TESIS – SS142501

PENGOPTIMALAN NAÏVE BAYES DAN REGRESI LOGISTIK MENGGUNAKAN ALGORITMA GENETIKA UNTUK DATA KLASIFIKASI

(Studi Kasus : Pembuangan Limbah Domestik di Surabaya Timur)

ABDURRAHMAN SALIM
NRP. 1315201015

DOSEN PEMBIMBING :
Irhamah, M.Si., Ph.D
Dr. Vita Ratnasari, S.Si., M.Si.

PROGRAM MAGISTER
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2017

(*halaman ini sengaja dikosongkan*)

THESIS – SS142501

OPTIMIZATION OF NAÏVE BAYES AND LOGISTIC REGRESSION USING GENETIC ALGORITHM FOR CLASSIFICATION DATA

(Case Study : Domestic Waste Disposal in East Surabaya)

ABDURRAHMAN SALIM
NRP. 1315201015

SUPERVISOR :
Irhamah, M.Si., Ph.D
Dr. Vita Ratnasari, S.Si., M.Si.

PROGRAM OF MAGISTER
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS AND NATURAL SCIENCES
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2017

(*halaman ini sengaja dikosongkan*)

PENGOPTIMALAN NAÏVE BAYES DAN REGRESI LOGISTIK MENGGUNAKAN ALGORITMA GENETIKA UNTUK DATA KLASIFIKASI
(Studi Kasus Pembuangan Limbah Domestik di Surabaya Timur)

Disusun untuk memenuhi syarat memperoleh gelar Magister Sains (M.Si)
di
Institut Teknologi Sepuluh Nopember

Oleh :
ABDURRAHMAN SALIM
NRP. 1315 2010 15

Tanggal Ujian : 21 Juni 2017
Periode Wisuda : September 2017

Disetujui Oleh :

1. Irhamah, M.Si., Ph. D (Pembimbing I)
   NIP. 19780406 200112 2 002
2. Dr. Vita Ratnasari, S.Si., M.Si (Pembimbing II)
   NIP. 19700910 199702 2 001
3. Prof. Drs. Nur Iriawan, MIkom, Ph.D (Penguji I)
   NIP. 19621015 198803 1 002
4. Dr. Wahyu Wibowo, S.Si, M.Si (Penguji II)
   NIP. 19740328 199802 1 001

Dekan
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Teknologi Sepuluh Nopember

Prof. Dr. Basuki Widodo, M.Sc
NIP. 19650605 198903 1 002

PENGOPTIMALAN NAÏVE BAYES DAN REGRESI LOGISTIK MENGGUNAKAN ALGORITMA GENETIKA UNTUK DATA KLASIFIKASI

(studi kasus pembuangan limbah domestik di Surabaya Timur)

Nama Mahasiswa : Abdurrahman Salim
NRP : 1315201015
Dosen Pembimbing : Irhamah, M.Si., Ph.D
Dr. Vita Ratnasari, S.Si., M.Si.

ABSTRAK

Klasifikasi pada data dalam jumlah banyak dan dengan fitur atau atribut yang beragam sering membuat hasil akurasi menjadi rendah. Untuk itu diperlukan metode yang dapat menangani pada data dengan jenis beragam tersebut. Metode yang dapat menangani masalah tersebut adalah metode *Naïve Bayes* dan Regresi Logistik. Metode *Naïve Bayes* merupakan salah satu metode data mining yang dapat mengatasi masalah data klasifikasi. Sedangkan regresi logistik merupakan salah satu metode klasifikasi, jika variabel respon tersebut bersifat biner dan terdapat banyak variabel prediktor berupa gabungan katagori dan kontinu. Metode Naive Bayes dan Regresi Logistik ini membutuhkan tahapan seleksi variabel Independen dalam meningkatkan keakurasian model dari Naive Bayes dan Regresi Logistik. Sehingga dibutuhkan metode yang bagus dalam memperbaiki kekurangan tersebut yaitu *Genetic Algorithm* (GA). Metode ini merupakan metode iteratif untuk mendapatkan global optimum. Hasil ketepatan klasifikasi dari Regresi Logistik Biner dan Naive Bayes pada kasus data septictank di wilayah Surabaya Timur dengan 11 variabel independen dan variabel dependennya berbentuk biner menghasilkan Naive Bayes lebih tinggi dibandingkan dengan Regresi Logistik dengan akurasi Naive Bayes sebesar 72.73 %, sedangkan Regresi Logistik Biner dengan akurasi sebesar 54.55 %. Namun ketika diseleksi dengan GA, hasil akurasi dari Naive Bayes dan Regresi Logistik Biner memiliki ketepatan klasifikasi yang sama yaitu 90.91 %.

Kata kunci: *Genetic Algorithm (GA)*, *Naïve Bayes*, Regresi Logistik, Klasifikasi.

(halaman ini sengaja dikosongkan)

OPTIMIZATION OF NAÏVE BAYES AND LOGISTIC REGRESSION USING GENETIC ALGORITHM FOR CLASSIFICATION DATA
(Case Study : Domestic Waste Disposal in East Surabaya)

Name : Abdurrahman Salim
NRP : 1315201015
Supervisor : Irhamah, M.Si., Ph.D
Dr. Vita Ratnasari, S.Si., M.Si.

ABSTRACT

Classification on large of data, and with a variety of features or attributes often makes the law accuracy. It required a method that has immunity in such diverse data types. The method can deal with the problem are Naïve Bayes method and Logistic Regression method. Naïve Bayes method is one of data mining that can be overcome the problem of data mining. While Logistic Regression is one of classification method, if response variable has binary characteristic and there are many predictor variable such as combination of category and continue. Method of Naive Bayes and Logistic Regression requires a stage selection independent variable in improving model accuration of Naive Bayes and Logistic Regression. So it takes a good method in fixing the deficiency is Genetic Algorithm (GA). This method is an iterative method to get global optimum. The results of the classification accuracy of Naive Bayes and Logistic Regression in the case of septictank data in East Surabaya with 11 independent variables and binary dependent variable is Naive Bayes higher than Logistic Regression with Naive Bayes accuracy of 72.73%, and Logiatic Regression accuracy of 54.55%. However when selected with GA, the accuracy of Naive Bayes and Binary Logistic Regression has the same classification accuracy of 90.91%.

Key Words: *Genetic Algorithm (GA)*, *Naïve Bayes*, Logistic Regression, Classification.

(halaman ini sengaja dikosongkan)

KATA PENGANTAR

Segala puji dan syukur bagi Allah SWT. yang telah memberikan limpahan rahmat, hidayah, dan kasih sayang-Nya, sehingga penulis dapat menyelesaikan Tesis dengan judul "***Pengoptimalan Naïve Bayes Dan Regresi Logistik Menggunakan Algoritma Genetika Untuk Data Klasifikasi***".

Dalam penyelesaian Tesis ini, penulis mendapatkan arahan, bimbingan, dan bantuan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis menghaturkan ucapan terima kasih yang tak terhingga kepada :

1. Kedua orang tua tercinta, Bapak H. M. Agus Salim dan Ibu Hj. Nur'aini atas kegigihan, perjuangan, air mata dalam doa, dan kasih sayangnya selama ini. Dan saudara-saudariku, M. Rijal Alfian dan Jihan Khairunnisa, atas selisih paham, canda tawa, doa serta dukungan dan kasih sayang tak terhingga selama penyelesaian Thesis ini.
2. Ibu Irhamah, M.Si., Ph.D. dan Ibu Dr. Vita Ratnasari, S.Si., M.Si. selaku dosen pembimbing, yang telah bersedia meluangkan waktu untuk memberikan bimbingan, saran, dan ilmu yang sangat bermanfaat dalam penyelesaian tesis ini.
3. Bapak Prof. Drs. Nur Iriawan, MIkom, Ph.D dan Bapak Dr. Wahyu Wibowo, S.Si, M.Si selaku dosen penguji yang telah memberikan banyak saran dan masukan agar tesis ini menjadi lebih baik.
4. Bapak Dr. Suhartono, M.Sc. selaku Ketua Jurusan Statistika ITS dan Bapak Dr. rer. pol. Heri Kuswanto, M.Si. selaku Kaprodi Pascasarjana Statistika FMIPA ITS.
5. Ibu Dr. Ismaini Zain, M.Si. selaku Dosen Wali atas motivasi, pelajaran berharga, bimbingan, arahan dan dukungan untuk melaksanakan Tesis ini.
6. Bapak /Ibu dosen pengajar di Jurusan Statistika ITS, terima kasih atas semua ilmu berharga yang telah diberikan.
7. Bapak/Ibu staf dan karyawan di Jurusan Statistika ITS, terima kasih atas segala bantuan selama masa perkuliahan penulis.

8. Penghuni Kos Keputih Gang Makam Blok A. No.3 (Rama, Jaya, Gozali, dll.)
9. Seseorang wanita (calon pendamping hidup penulis) yang selalu menyemangati, memberikan perhatian dan pengertian, dan banyak hal lainnya untuk mendukung penyelesaian tesis ini.
10. Semua teman-teman seperjuangan S2 Statistika ITS, terima kasih atas bantuan dan kebersamaan selama ini, khususnya Pencari Ilmu (Mbak Cinti, Mbak Tutus, Ifa, Titin, Rizfani, Asmita, Rani, dan Surya).
11. Serta, semua pihak yang telah membantu penulis, namun tidak dapat penulis sebutkan satu per satu.

Penulis mengharapkan kritik dan saran yang bersifat membangun demi kesempurnaan Tesis ini. Semoga bantuan dan bimbingan yang telah diberikan kepada penulis selama penyelesaian Tesis ini mendapat balasan setimpal dari Allah SWT dan dapat memberi manfaat bagi pembacanya. Amin.

Surabaya, September 2017

Penulis

DAFTAR ISI

(*halaman ini sengaja dikosongkan*)

DAFTAR TABEL

DAFTAR GAMBAR

(halaman ini sengaja dikosongkan)

DAFTAR LAMPIRAN

(halaman ini sengaja dikosongkan)

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Pengklasifikasian merupakan salah satu metode statistika untuk mengelompok atau mengklasifikasi suatu data yang disusun secara sistematis. Masalah klasifikasi sering dijumpai dalam kehidupan sehari-hari. Baik itu pengklasifikasian data pada bidang akademik, sosial, pemerintahan, maupun pada bidang lainnya. Masalah klasifikasi ini muncul ketika terdapat sejumlah ukuran yang terdiri dari satu atau beberapa kategori yang tidak dapat diidentifikasikan secara langsung tetapi harus menggunakan suatu ukuran.

Dalam statistika ada beberapa metode klasifikasi yang digunakan untuk melakukan klasifikasi data seperti: Analisis Diskriminan, Regresi Logistik, *Naïve Bayes*, dan lain-lain. Dalam penelitian ini yang akan dibahas ialah metode klasifikasi regresi logistik dan *Naïve Bayes*.

Regresi logistik adalah salah satu pendekatan model matematis yang digunakan untuk menganalisis hubungan antara satu atau beberapa variabel independen yang bersifat kontinu maupun biner dengan satu variabel dependen yang bersifat dikotomus (biner). Misalkan variabel dependen adalah *Y* dan variabel independen adalah *X*. Dalam hal ini regresi logistik tidak memodelkan secara langsung variabel dependen *Y* dengan variabel independen *X*, melainkan melalui transformasi variabel dependen ke variabel logit yang merupakan natural log dari odds rasio.

Selain regresi logistik, klasifikasi juga dapat dilakukan menggunakan metode *Naïve Bayes*. *Naïve Bayes* merupakan sebuah pengklasifikasi probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Dalam penggunaannya, *Naïve Bayes* merupakan metode yang hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifisian. *Naïve Bayes* sering bekerja jauh lebih baik dalam

kebanyakan situasi dunia nyata kompleks dari pada yang diharapkan (Patterkari dan Parveen, 2012).

Dalam penggunaan metode *Naïve Bayes* dan Regresi Logistik diperlukan tahapan penseleksian variabel untuk meningkatkan keakuratan model dalam menjelaskan data. Metode statistik yang biasanya digunakan untuk menseleksi variabel adalah metode *forward selection*, *stepwise* dan *backward elimination*. Namun metode seleksi variabel klasik tersebut memberikan kesimpulan bahwa pemodelan regresi akan memberikan informasi yang kurang lengkap karena adanya kasus multikolinieritas. Sehingga diperlukan metode lain dalam seleksi variabel salah satunya adalah *Genetic Algorithm* (GA).

Genetic Algorithm (GA) adalah salah satu metode metaheuristik yang digunakan dalam teknik optimasi. Menurut Haupt dan Haupt (2004) Algoritma Genetika adalah suatu teknik optimasi yang didasarkan pada prinsip genetik dan seleksi alam. Dalam Algoritma Genetika populasi terbentuk dari banyak individu yang berkembang sesuai aturan seleksi spesifik dengan memaksimalkan fitness. Menurut Sivanandam dan Deepa (2008) kelebihan yang dimiliki Algoritma Genetika dibanding metode-metode yang lain diantaranya yaitu sangat cocok digunakan untuk menyelesaikan masalah global optimum, mudah diubah atau fleksibel untuk diimplementasikan pada berbagai masalah dan ruang solusi lebih luas.

Menurut Guo et al (2010, 2990) Algoritma Genetika merupakan salah satu algoritma optimasi, yang diciptakan untuk meniru beberapa proses yang diamati dalam evolusi alam. Algoritma Genetika juga merupakan algoritma stokastik yang kuat berdasarkan prinsip-prinsip seleksi alam dan natural genetik yang cukup berhasil diterapkan dalam masalah machine learning dan optimasi.Algoritma Genetika juga merupakan metode yang berbasis perulangan-perulangan atau iterasi untuk mendapatkan nilai glibal optimum. Iterasi akan berhenti ketika kriteria yang digunakan untuk analisis telah sesuai dengan yang ditentukan. Kriteria yang digunakan untuk penelitian ini adalah nilai kesalahan klasifikasi pada setiap metode yang digunakan. Algoritma genetika digunakan untuk seleksi variabel. Kombinasi variabel independen yang menghasilkan kesalahan klasifikasi paling kecil yang akan terpilih sebagai solusi seleksi variabel menggunakan algoritma genetika.

Terdapat tiga langkah utama dalam algoritma Genetika yaitu seleksi orang tua, pindah silang, dan mutasi. Seleksi orang tua merupakan tahapan pemilihan beberapa kromosom untuk dijadikan orang tua generasi berikutnya, kromosom berisi kandidat (gen) solusi dari permasalahan yang akan diselesaikan. Pada penelitian ini gen berisi variabel independen dan estimasi parameter dari analisis Regresi Logistik dan *Naïve Bayes*. Pindah silang adalah proses pengombinasian gen pada 2 kromosom orang tua. Mutasi adalah pemilihan gen secara random untuk kromosom yang baru (Kusumawardani, 2015).

Kebaikan algoritma genetika untuk seleksi variabel telah dibuktikan oleh penelitian Back, Laitinen, Sere, dan Wezel (1996). Penelitian tersebut membandingkan metode seleksi variabel menggunakan *stepwise method* pada analisis diskriminan dan Regresi Logistik dengan algoritma genetika pada *neural network*, yang diaplikasikan pada tiga data, yakni data satu tahun, dua tahun, dan tiga tahun sebelum kegagalan. Penelitian tersebut memberikan hasil bahwa seleksi variabel dengan menggunakan algoritma genetika lebih bagus pada satu dan tiga tahun sebelum kegagalan daripada hanya stepwise yang bagus hanya di dua tahun sebelum kegagalan. Menurut Xu dan Zhang (2001) pada penelitiannya menyatakan bahwa seleksi variabel menggunakan pendekatan algoritma genetika lebih bagus dibandingkan metode seleksi variabel *backward elimination*, *forward selection*, dan *stepwise method* untuk regresi berganda.

Penelitian yang lainnya juga pernah dilakukan diantaranya penelitian Chiang dan Pell (2004) dengan menggunakan GA yang dikombinasikan dengan analisis diskriminan untuk mengidentifikasi variabel, dimana hasil yang diberikan bahwa GA dapat menyelesaikan masalah dalam mengidentifikasi variabel. Pada penelitian Wati (2016) dengan menerapkan GA dalam seleksi fitur pada *Naïve Bayes*, menyatakan bahwa penerapan GA pada *Naïve Bayes* dapat meningkatkan keakurasian klasifikasi.

Untuk melihat hasil klasifikasi tersebut, maka diperlukan kasus data yang memiliki sejumlah ukuran yang terdiri dari satu atau beberapa kategori yang tidak dapat diidentifikasikan secara langsung tetapi harus menggunakan suatu ukuran. Pada penelitian ini, kasus yang digunakan adalah faktor-faktor yang mempengaruhi rumah tangga membuang limbah domestik di Surabaya Timur pada

penelitian tugas akhir Kusumawati (2013) dengan menggunakan Regresi Logistik dan Naïve Bayes dengan optimasi *Genetic Algorithm* (GA). Pada kasus ini, mempunyai dua katagori di variabel responnya dan sebelas variabel prediktor serta merupakan data klasifikasi. Metode Regresi Logistik dan Naïve Bayes sangat tepat dalam mengklasifikasikan dan ditambah juga adanya algoritma GA dalam mengoptimalkan keakuratan dari model dengan menyeleksi variabel-variabel prediktornya. Penelitian ini akan membandingkan keakuratan klasfikasi dari masing-masing metode yaitu Regresi Logistik Biner dan *Naïve Bayes* dengan pengoptimuman GA. Metode yang terbaik adalah metode yang memiliki nilai kesalahan klasifikasi yang terkecil.

1.2 Rumusan Masalah

Rumusan masalah yang dapat diambil dari Penelitian ini adalah :

1. Bagaimanakah ketepatan klasifikasi yang didapatkan dengan menggunakan metode *Naïve Bayes* dan Regresi Logistik pada data kasus pembuangan limbah domestik daerah Surabaya Timur?
2. Bagaimanakah penerapan GA dalam penyeleksian variabel pada Regresi Logistik dan Naive Bayes?
3. Bagaimanakah Perbandingan diantara Naive Bayes, Regresi Logistik, dan Penggunaan GA pada Naive Bayes dan Regresi Logistik?

1.3 Tujuan Penelitian

Dari rumusan masalah di atas, didapatkan tujuan sebagai berikut :

1. Menentukan ketepatan klasifikasi dari *Naïve Bayes* dan Regresi Logistik pada data kasus pembuangan limbah domestik daerah Surabaya Timur.
2. Mengkaji dan menerapkan optimasi GA dalam penyeleksian variabel pada *Naïve Bayes* dan Regresi Logistik.
3. Membandingkan tingkat klasifikasi dari *Naïve Bayes*, Regresi Logistik, dan penggunaan optimasi GA pada *Naïve Bayes* dan Regresi Logistik berdasarkan akurasi, *error*, *sensitivity*, *specivicity* dan *G-Mean*.

1.4 Manfaat Penelitian

Manfaat yang dapat diambil dari penelitian ini adalah :

1. Dapat menambah dan mengembangkan wawasan dari metode data mining seperti Regresi Logistik dan *Naïve Bayes*.
2. Menambah keilmuan dari metode optimasi GA dalam penseleksian variabel secara global.
3. Dapat memberikan saran dan pembanding untuk pengembangan metode dari penelitian yang akan dilakukan selanjutnya.

1.5 Batasan Masalah

Batasan untuk penelitian ini adalah sebagai berikut:

1. Penelitian ini menggunakan metode Regresi Logistik dan metode Naïve Bayes dalam menentukan hasil klasifikasi.
2. Algoritma *Genetic Algorithm* (GA) yang digunakan dalam mengoptimasi metode Regresi Logistik dan Naïve Bayes.
3. Data yang diambil adalah data rumah tangga yang membuang limbah domestik (rumah tangga) atau *blackwater* (tinja) pada tangki septik dan rumah tangga yang membuang limbah langsung ke badan air (selokan, sungai, dan lain-lain) atau yang mencemari tanah di Surabaya Timur tahun 2013.

(*halaman ini sengaja dikosongkan*)

BAB 2
TINJAUAN PUSTAKA

2.1 Regresi Logistik

Regresi logistik merupakan salah satu metode klasifikasi yang sering digunakan. Regresi logistik biner digunakan saat variabel dependen merupakan vaiabel dikotomus. Regresi logistik multinomial digunakan pada saat variabel dependen adalah variabel katagorik dengan lebih dari dua katagori. Secara umum model regresi logistik adalah :

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \qquad (2.1)$$

Dimana $\pi(x)$ merupakan nilai probabilitas dari $0 \leq \pi(x) \leq 1$, yang berarti bahwa regresi logistik menggambarkan suatu probabilitas. Dengan mentransformasikan $\pi(x)$ pada persamaan di atas dengan transformasi logit $g(x)$, dimana :

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \qquad (2.2)$$

maka diperoleh bentuk logit :

$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ (Hosmer dan Lemeshow, 1989). (2.3)

Untuk memperoleh estimasi dari parameter regresi logistik dapat dilakukan dengan cara *Maximum Likelihood Estimation* (MLE) sebagai berikut :

Estimasi parameter dalam model logit menggunakan *Maximum Likelihood* dengan langkah-langkah sebagai berikut.

1. Fungsi likelihood dari Y

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{p}]^{\mathrm{y_i}} [\mathrm{q}]^{1-\mathrm{y_i}}$$

2. Fungsi ln-likelihood

$$\ln \mathrm{L}(\boldsymbol{\beta}) = \ln\left\{\prod_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{p}]^{\mathrm{y_i}} [\mathrm{q}]^{1-\mathrm{y_i}}\right\}$$

$$\ln \mathrm{L}(\boldsymbol{\beta}) = \ln\left\{\prod_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{p}]^{\mathrm{y_i}} [1-\mathrm{p}]^{1-\mathrm{y_i}}\right\}$$

$$\ln L(\boldsymbol{\beta}) = \ln\left\{\prod_{i=1}^{n}\left[\frac{\exp(g(x))}{1+\exp(g(x))}\right]^{y_i}\left[1-\left(\frac{\exp(g(x))}{1+\exp(g(x))}\right)\right]^{1-y_i}\right\}$$

$$\ln L(\boldsymbol{\beta}) = \sum_{s=0}^{q}\left\{\left[\sum_{i=1}^{n} y_i\, x_q\right]\beta_q - \sum_{i=1}^{n}\ln\left[1+\exp(\sum_{s=0}^{q}\beta_q\, x_q\right]\right\}$$

3. Fungsi ln-likelihood diturunkan terhadap β

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} y_i\, x_q - \sum_{i=1}^{n} x_q\left(\frac{\exp(g(x))}{1+\exp(g(x))}\right)$$

$$\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^{\mathbf{T}}} = \sum_{i=1}^{n} x_q {x_q}^{T}\, p(1-p)$$

4. Apabila menghasilkan bentuk yang tidak close form, maka estimasi parameter $\boldsymbol{\beta}$ diperoleh melalui prosedur iterasi dengan metode Newton Raphson. Metode Newton Raphson diperoleh dari pendekatan deret Taylor sebagai berikut.

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \left(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}\right)\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^{\mathbf{T}}} + \frac{1}{2!}\left(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}\right)\left(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}\right)^{T}\frac{\partial^3 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathbf{T}}\,\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^{\mathbf{T}}} + \cdots = 0$$

Vektor $\boldsymbol{\beta}$ merupakan nilai awal yang ditentukan. Apabila $\left|\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}\right|$ diasumsikan sangat kecil maka suku ketiga dan se-terusnya dapat diabaikan. Dengan demikian ekspansi deret Taylor dapat ditulis sebagai berikut.

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} + \left(\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}\right)\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^{\mathbf{T}}} = 0$$

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta} - \left(\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}^{\mathbf{T}}}\right)^{-1}\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

Secara umum iterasi ke-t metode Newton Raphson adalah sebagai berikut.

$$\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{(t-1)} - \left(\frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(t-1)}\,\partial \boldsymbol{\beta}^{\mathbf{T}(t-1)}}\right)^{-1}\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(t-1)}}$$

Proses iterasi akan berhenti pada saat $\left\|\boldsymbol{\beta}^{(t)}-\boldsymbol{\beta}^{(t-1)}\right\| \leq \varepsilon$.

Dalam melakukan evaluasi fungsi klasifikasi dilakukan dengan membagi data menjadi 2 bagian. Bagian pertama akan dipergunakan sebagai training set, yang diperlakukan untuk membentuk model klasifikasi regresi logistik. Berikutnya, bagian kedua akan dipergunakan sebagai validasi set, yang berfungsi sebagai cross-validasi fungsi klasifikasi regresi logistik. Dalam melakukan pengklasifikasian diharapkan untuk meminimalkan kesalahan klasifikasi atau meminimalkan rata-rata efek buruk dari kesalahan klasifikasi.

2.2 Uji Signifikasi

2.3.1 Uji Parameter secara keseluruhan

Dalam Hosmer, dkk. (1989) uji parameter secara keseluruhan atau uji rasio likelihood diperoleh dengan cara membandingkan fungsi log likelihood dari seluruh variabel bebas dengan fungsi log likelihood tanpa variabel bebas

a. Hipotesis

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

$H_1 :$ paling sedikit salah satu dari $\beta_m \neq 0$ dengan $m = 1, 2, \ldots, k$

b. Statistik uji Rasio Likelihood

$$X^2_{hit} = -2\log\left(\frac{likelihood \text{ tanpa variabel bebas}}{likelihood \text{ dengan variabel bebas}}\right)$$

$$= 2\log[likelihood \text{ tanpa variabel bebas}] - 2\log[likelihood \text{ dengan variabel bebas}]$$

c. Kriteria Uji

Tolak H_0 jika $X^2_{hit} > X^2_{(\alpha, p)}$

2.3.2 Uji Parameter secara individu

Dalam Hosmer, dkk. (1989) uji parameter secara individu diperoleh dengan cara mengkuadratkan rasio estimasi parameter dengan estimasi standar errornya. Uji ini mengunakan uji Wald yang berfungsi menguji signifikansi tiap parameter.

a. Hipotesis

$H_0 : \beta_m = 0$

$H_1 : \beta_m \neq 0$ dengan $m = 1, 2, \ldots, k$

b. Statistik uji Rasio Likelihood

$$W_m = \left[\frac{\hat{\beta}_m}{SE(\hat{\beta}_m)} \right]^2$$

c. Kriteria Uji

Tolak H_0 jika $W_m > X^2_{(\alpha,1)}$

2.3 *Naïve Bayes*

Naïve Bayes merupakan sebuah pengklasifikasi probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahan frekuensi dan kombnasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas (Patil dan Shereker, 2013). Definisi lain mengatakan *Naïve Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebeumnya (Bustami, 2013).

Naïve Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu (Ridwan, Suyono dan Sarosa, 2013). Keuntungan penggunaan Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifisian. *Naïve Bayes* sering bekerja jauh lebih baik dalam kebanykan situasi dunia nyata kompleks dari pada yang diharapkan (Patterkari dan Parveen, 2012).

2.3.1 Teorema Bayes

Persamaan dari teorema Bayes adalah :

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)} \qquad (2.4)$$

dimana :

X	:	Data dengan *class* yang belum dietahui
H	:	Hipotesis data merupakan suatu *class* spesifik
$P(H\|X)$	:	Probabilitas hipotesis H berdasar kondisi X (posterior probabilitas)
$P(H)$	:	Probabilitas hipotesis H (prior probabilitas)
$P(X\|H)$	:	Probabilitas hipotesis X berdasar kondisi pada hipotesis H
$P(X)$	:	Probabilitas hipotesis X

Untuk menjelaskan metode *Naïve Bayes,* perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, metode *Naïve Bayes* di atas disesuaikan sebagai berikut :

$$P(C \mid F1 \ldots Fn) = \frac{P(\mathrm{C})P(F1 \ldots Fn \mid C)}{P(F1 \ldots Fn)} \qquad (2.5)$$

Dimana variabel C mempresentasikan kelas, sementara variabel $F1 \cdots Fn$ mempresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi. Penjabaran lebih lanjut rumus *Bayes* tersebut dilakukan dengan menjabarkan $(C|F1, \cdots, Fn)$ menggunakan aturan perkalian sebagai berikut :

$$\begin{aligned}
&P(C \mid F_1, \ldots, F_n) = P(C)P(F_1, \ldots, F_n \mid C) \\
&= P(C)P(F_1 \mid C)P(F_2, \ldots, F_n \mid C, F_1) \\
&= P(C)P(F_1 \mid C)P(F_2 \mid C, F_1)P(F_3, \ldots, F_n \mid C, F_1, \mathrm{F}_2) \\
&= P(C)P(F_1 \mid C)P(F_2 \mid C, F_1)P(F_3 \mid C, F_1, \mathrm{F}_2)P(F_4, \ldots, F_n \mid C, F_1, \mathrm{F}_2, \mathrm{F}_3) \\
&= P(C)P(F_1 \mid C)P(F_2 \mid C, F_1)P(F_3 \mid C, F_1, \mathrm{F}_2) \ldots P(F_n, \ldots, F_{n-1} \mid C, F_1, \mathrm{F}_2, \ldots, \mathrm{F}_{n-1}) \quad (2.6)
\end{aligned}$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Disinilah digunakan asumsi independensi yang sangat tinggi (*naïve*), bahwa masing-masing petunjuk $(F_1, F_2, \cdots, F_n)$ saling bebas (*independen*) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut :

$$P(F_i \mid F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i), \quad \text{untuk } i \neq j, \qquad (2.7)$$

sehingga,

$$P(F_i \mid C, F_j) = P(F_i \mid C) \tag{2.8}$$

Persamaan di atas merupakan model dari teorema *Naïve Bayes* yang selanjutnya akan digunakan dalam proses klasifikasi. Untuk klasifikasi dengan data kontinu digunakan rumus *Densitas Gauss* :

$$P(X_i = x_i \mid Y = y) = \frac{1}{\sqrt{2\pi\sigma_y{}^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma^2{}_y}} \tag{2.9}$$

dimana :

P : Peluang

X_i : Atribut ke – i

x_i : nilai atribut ke – i

Y : Kelas yang dicari

y : Sub kelas Y yang dicari

μ_y : *mean*, menyatakan rata-rata dari seluruh atribut

σ_y : Standar deviasi, menyatakan varian dari seluruh atribut

2.3.2 Naïve Bayes Untuk Klasifikasi

Klasifikasi adalah proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya. Selain itu, klasifikasi Naïve Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikaskan ke dalam basis data dengan jumlah yang besar (Han dan Kamber, 2006).

Salah satu penerapan theorem Bayes dalam klasifikasi adalah *Naïve Bayes*.

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n \mid v_j) P(v_j) \tag{2.10}$$

Naïve Bayes didasarkan pada asumsi penyederhanaan bahwa nilai aribut secara conditional saling bebas jika diberikan nilai output. Atau dapat diberikan $P(a_1, a_2, \ldots, a_n \mid v_j) = \prod_i P(a_i \mid v_j)$, kemudian dimasukkan pada persamaan di atas, maka akan didapat pendekatan yang dipakai dalam *Naïve Bayes* klasifikasi.

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j) \quad (2.11)$$

dimana v_{NB} adalah nilai output hasil klasifikasi Naïve *Bayes.*

2.4 Algoritma Genetika

Algoritma Genetika merupakan suatu teknik optimasi yang didasarkan pada prinsip genetik dan seleksi alam. Dalam Algoritma Genetika populasi terbentuk dari banyak individu yang berkembang sesuai aturan seleksi spesifik dengan memaksimalkan fitness (Haupt dan Haupt, 2004). Algoritma ini juga digunakan untuk mendapatkan nilai global optimum dengan cara melakukan perulangan atau iterasi pada konsep evolusi darwin.

Menurut Trevino dan Falciani (2006) terdapat 7 tahapan untuk menjalankan algoritma genetika, yaitu :

1. Membentuk populasi awal terdiri dari beberapa kromosom yang didalamnya memuat gen kromosom pada algoritma genetika yang digunakan untuk menunjukkan kandidat sekelompok gen yang dapat digunakan sebagai solusi permaslahan. Gen pada algoritma genetika berisi variabel yang ingin dioptimumkan, pada penelitian ini gen berisi variabel indepnden.
2. Masing-masing kromosom dalam populasi dievaluasi kemampuannya dengan menggunakan fungsi fitness. Pada penelitian ini fungsi fitnesnya berupa kesalahan klasifikasi.
3. Ketika sebuah kromosom memiliki nilai fitness lebih optimum daripada nilai inisialnya, maka kromosom dihentikan, namun apabila tidak maka tahapan analisis dilanjutkan ke tahap 4. Nilai fitnes terkecil yang dipilih sebagai solusi permasalahan dari penelitian ini, karena fungsi fitnes yang digunakan adalah tingkat kesalahan klasifikasi.
4. Memilih kromosom dengan nilai fitnes yang optimum yang dijadikan orang tua.
5. Mengkombinasikan informasi genetika yang ada dalam replikasi orang tua melalui pindah silang. Dua induk secara random dipilih dan digunakan untuk membentuk dua kromosom baru.

6. Melakukan mutasi untuk memperkenalkan unsur gen baru pada kromosom secara acak.
7. Tahapan diulangi dari tahapan 2 sampai kromosom yang memberikan nilai fitnes paling optimum atau sudah mencapai konvergen.

2.4.1 Pengkodean

Pengkodean adalah proses menggambarkan bentuk gen dalam kromosom. Pengkodean dapat berupa *bits, number, trees, arrays, list*, dan lain-lain. Pengkodean yang dilakukan pada penelitian ini adalah pengkodean *bits* (biner) untuk seleksi variabel.

2.4.2 *Fitness*

Fungsi *fitness* digunakan untuk mengukur tingkat kebaikan atau kesesuaian suatu solusi yang dicari. Fungsi fitness bisa berhubungan dengan langsung fungsi tujuan, atau bisa juga sedikit modifikasi terhadap fungsi tujuan. Sejumlah solusi yang dibangkitkan dalam populasi akan dievaluasi menggunakan fungsi *fitness*. Fungsi *fitness* ($F(x)$) yang digunakan adalah:

$$F(x) = \frac{1}{1 + f(x)} \tag{2.14}$$

dimana $f(x)$ adalah fungsi tujuan dari masalah yang dapat terseleseikan.

Untuk kasus minimasi, jika didapatkan $f(x)$ yang kecil maka nilai *fitness*-nya besar. Sebaliknya, untuk kasus maksimasi, fungsi *fitness*-nya bias menggunakan nilai $f(x)$ sendiri, jadi $\mathrm{F}(x) = f(x)$ (Santosa dan Willy, 2011).

2.4.3 Seleksi Orang Tua

Desiani dan Arhami (2006) mengemukakan bahwa seleksi orang tua bertujuan untuk memberikan kesempatan reproduksi bagi anggota populasi yang memiliki nilai fitness tinggi. Pemilih-an dua buah kromosom dalam suatu populasi sebagai orang tua yang akan dipindahsilangkan biasanya secara proporsional sesuai dengan nilai fitness masing-masing. Metode umum yang dipakai adalah Roulette Wheel (Roda Roulette). Pada metode ini, masing-masing kromosom menempati potongan lingkaran pada Roda Roulette secara proporsional sesuai dengan nilai

fitnessnya (Suyanto, 2005). Sebuah kromosom yang nantinya akan terpilih adalah apabila bilangan random yang dibangkitkan berada dalam nilai interval kumulatifnya. Nilai kumulatif ini didapatkan dari membagi nilai fitness dari tiap kromosom dengan total nilai fitness keseluruhan.

2.4.4 Pindah Silang

Pindah silang merupakan metode pengkombinasian 2 orang tua kromosom untuk membentuk generasi baru. Orang tua diperoleh dari tahapan seleksi orang tua menggunakan RWS. Kromosom dengan nilai fitnes yang tinggi akan dijadikan sebagai orang tua. Terdapat beberapa metode untuk pindah silang yaitu, single point, two points, arithmetic, intermediate, dan scattered. Penelitian yang dilakukan oleh Bocko, Nohajova, & Harcarik (2011) terkait perbandingan hasil beberapa metode pindah silang dengan berbagai macam proses seleksi orang tua memberikan hasil bahwa untuk seluruh proses seleksi orang tua, metode pindah silang scattered memberikan perfoma paling bagus. Oleh karena itu, pada penelitian ini digunakan metode scattered untuk proses pindah silang. Metode ini menghasilkan kromosom baru yang berisi randoman nilai 0 dan 1, nilai 1 menunjukkan gen yang berasal dari orang tua pertama dan nilai 0 menunjukkan gen yang berasal dari orang tua kedua (Kusumawardani, 2015).

2.4.5 Mutasi

Metode yang digunakan untuk mutasi pada penelitian ini adalah metode uniform. Tujuan mutasi adalah untuk mendapatkan keberagaman gen. Terdapat 2 tahapan dalam metode ini, pertama adalah memilih sebagian gen dari kromosom yang akan dimutasi dengan peluang sebesar 0.1, kedua mengganti gen yang terpilih dengan bilangan random yang dibangkitkan dari batas bawah dan batas atas nilai gen dalam kromosom.

2.4.6 Elitisme

Suatu individu yang memiliki nilai fitness tertinggi tidak akan selalu terpilih karena proses seleksi dilakukan secara random. Oleh karena itu perlu

dilakukan elitisme, yaitu suatu prosedur pengopian individu agar individu yang bernilai fitness tertinggi tidak hilang selama proses evolusi (Suyanto, 2005).

2.5 Klasifikasi

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumah kelas yang tersedia. Dalam klasifikasi ada 2 pekerjaan utama yang dilakukan, yaitu : pertama, pembangunan model sebagai prototype untuk disimpan sebagai memori dan kedua, penggunaan model tersebut untuk melakukan pengenalan/ klasifikasi/ prediksi pada suatu objek data tersebut dalam model yang mudah disimpan.

Sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua set data dengan benar, tetapi tidak dapat dipungkiri bahwa kinerja suatu sistem tidak bias 100% benar sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya, pengukuran kinerja dilakukan dengan matriks kofusi.

2.6 Evaluasi Performansi Metode Klasifikasi

Data aktual dan data hasil prediksi dari model klasifikasi disajikan dengan menggunakan Tabulasi silang (*Confusion matrix*), yang mengandung informasi tentang kelas data aktual direpresentasikan pada baris matriks dan kelas data hasil prediksi pada kolom (Han dan Kamber, 2006). Dimana untuk Confusion Matrix dapat dilihat pada Tabel 2.1.

Tabel 2.1 *Confusion Matrix*

Aktual	Prediksi	
	Positif	Negatif
Positif	*TP*	*FN*
Negatif	*FP*	*TN*

Keterangan :

1. *True Positive* (TP) menunjukkan bahwa kelas yang dihasilkan prediksi klasifikasi adalah positif dan kelas sebenarnya adalah positif.
2. *True Negatif* (TN) menunjukkan bahwa kelas yang dihasilkan dari prediksi klasifikasi adalah negatif dan kelas sebenarnya adalah negative.

3. *False Positif* (FP) menunjukkan bahwa kelas yang dihasilkan dari prediksi klasifikasi adalah negative dan kelas sebenarnya adalah positif.
4. *False Negatif* (FN) menunjukkan bahwa kelas yang dihasilkan dari prediksi klasifikasi adalah positif dan kelas sebenarnya adalah negatif.

Ketepatan klasifikasi dapat dilihat dari akurasi klasifikasi. Akurasi klasifikasi menunjukkan performansi model klasifikasi secara keseluruhan, dimana semakin tinggi akurasi klasifikasi hal ini berarti semakin baik performansi model klasifikasi.

$$Akurasi\ Total = \frac{Jumlah\ prediksi\ benar}{jumlah\ total\ prediksi} \times 100\%$$
$$Akurasi\ Total = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \quad (2.16)$$

Kemudian dapat juga menghitung APER atau yang disebut laju error merupaka ukuran evaluasi yang digunakan untuk melihat peluang kesalahan klasifikasi yang dihasilkan oleh suatu fungsi klasifikasi. Semakin kecil nilai APER maka hasil pengklasifikasian semakin baik (Prasetyo, 2012).

Formulasi untuk menhitung APER yaitu :

$$\text{APER} = \frac{FP + FN}{TP + FP + TN + FN} \times 100\% \quad (2.17)$$

Untuk mendapatkan klasifikasi yang optimal dan lebih spesifik maka dapat diuji *Sensitivity* dan *Specificity*. *Sensitifity* adalah tingkat positif benar atau ukuran performansi untuk mengukur kelas yang positif (minor) sedangkan *Specifity* adalah tingkat negatif benar atau ukuran performansi untuk mengukur kelas yang negatif (mayor). Rumus *Sensitivity* dan *Specificity* adalah sebagai berikut :

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100\% \quad (2.18)$$

$$Specificity = \frac{TN}{(TN + FP)} \times 100\% \quad (2.19)$$

Selain itu, evaluasi performansi model klasifikasi dapat dilakukan dengan menggunakan *G-mean*. *G-mean* merupakan rata-rata *geometric Sensitivity* dan *Specificity*. Apabila semua kelas positif tidak dapat diprediksi maka *G-mean* akan

bernilai nol, sehingga diharapkan suatu algoritm klasifikasi mencapai nilai *G-mean* yang tinggi (Kubat, Matwin dan Holte, 1997).

$$G - Mean = \sqrt{Sensitivity \times Specificity} \quad (2.20)$$

BAB 3
METODE PENELITIAN

3.1 Sumber Data

Sumber data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari penelitian Kusumawati (2013). Data yang diambil adalah rumah tangga yang membuang limbah domestik (rumah tangga) atau *blackwater* (tinja) pada tangki septik dan rumah tangga yang membuang limbah langsung ke badan air (selokan, sungai, dan lain-lain) atau yang mencemari tanah. Jumlah data yang digunakan adalah 102 observasi.

3.2 Variabel Penelitian

Variabel respon yang digunakan seperti pada Tabel 3.1.

Tabel 3.1 Variabel Penelitian (Variabel Respon)

Variabel Respon (Y)	
Y = Pembuangan Air Limbah Domestik *blackwater* (tinja) yang dibuang oleh setiap rumah tangga	0 = mempunyai *septictank*
	1 = tidak mempunyai *septictank*

Variabel prediktor yang digunakan untuk mengetahui faktor-faktor rumah tangga yang membuang limbah domestik yang berada di Surabaya Timur adalah seperti pada Tabel 3.2.

Tabel 3.2 Variabel Penelitian (Variabel Prediktor)

Variabel Independen (X)	
Variabel	Keterangan
X1 = Jumlah Keluarga	0 : ≤ 4 1 : > 4
X2 = Pendidikan Kepala Keluarga	0 : Tidak Sekolah dan SD 1 : SMP / Sederajat 2 : SMA / Sederajat 3 : Diploma dan Sarjana
X3 = Pekerjaan Kepala Keluarga	0 : PNS / BUMN 1 : Karyawan Swasta 2 : Wiraswasta 3 : Pertukangan

Tabel 3.2 Lanjutan

Variabel Independen (X)	
Variabel	Keterangan
X4 = Pendapatan KK	0 : < UMR 1 : > UMR
X5 = Pengeluaran KK	0 : < UMR 1 : > UMR
X6 = Status Rumah	0 : Status Hak Milik 1 : Kontrak / Kos 2 : Milik Orangtua 3 : Rumah Dinas
X7 = Lama Waktu Pengurasan	0 : belum pernah nguras 1 : < 4 2 : > 4
X8 = Jenis Kloset	0 : duduk 1 : Jongkok
X9 = Jenis Penyakit	0 : Diare 1 : Paratipus 2 : Demam Berdarah 3 : Lainnya
X10 = Air Mandi Cuci	0 : Air Tanah 1 : PDAM
X11 = Prilaku Sanitasi	0 : Baik 1 : Kurang Baik

3.3 Langkah-langkah Penelitian

Untuk mencapai penelitian ini akan dilakukan dengan tahapan-tahapan sebagai berikut :

A. Pengumpulan Data

Pada pengumpulan data ini merupakan langkah awal pada suatu penelitian. Data yang digunakan pada penelitian ini adalah data sekunder dari penelitian Yuriko (2013).

B. Eksperimen dan Pengolahan data

Pada langkah ini akan membahas proses pengolahan data untuk metode regresi logistik dan naïve bayes serta penggunaan optimasi algoritma Genetik. Untuk setiap metodenya akan dibahas sebagai berikut :

I. Untuk pengolahan data menggunakan metode Regresi Logistik dan Naive Bayes :

a. Regresi Logistik :
    1. Membagi data menjadi dua, yaitu data training 90 % dan data testing 10 %.
    2. Melakukan uji independensi dengan menggunakan data training.
    3. Membentuk model Regresi Logistik menggunakan data training.
    4. Menguji signifikansi parameter secara individu dan secara keseluruhan
    5. Melakukan validasi keakuratan prediksi dari model dengan data testing.
    6. Menghitung nilai Akurasi, APER, *Sensitivity*, *Specificity*, dan *G-Mean* dari model regresi logistik yang terbentuk
b. Naive Bayes :
    1. Membagi data menjadi data training dan data testing
    2. Menghitung probabilitas awal (prior probability) $Y(P(Y))$.
    3. Menghitung Nilai Probabilitas independen kelas *Y* dari semua fitur dalam vektor *X* ($\prod_{i=1}^{k} P(X_i|Y)$).
    4. Menghitung nilai posterior probability untuk masing-masing klasifikasi $(P(Y|X))$.
    5. Menghitung prediksi yang didapat dari nilai maksmum dalam perhitungan posterior probability.
    6. Menghitung nilai Akurasi, APER, *Sensitivity*, *Specificity*, dan *G-Mean* dari model yang terbentuk dari Naive Bayes.

II. Langkah selanjutnya adalah pengerjaan Algoritma Genetika untuk Regresi Logistik dan Naive Bayes, sebagai berikut :
    1. Membentuk populasi awal yang terdiri dari beberapa kromosom.
    2. Mengevaluasi nilai setiap masing-masing kromosom dengan nilai fitness. Dimana nilai fitness yang digunakan adalah kesalahan klasifikasi model. Untuk fungsi evaluasi fitnessnya digunakan tergantung pada setiap metode yang akan diseleksi seperti Regresi Logistik menggunakan persamaan regresi logistik dalam fungsi evaluasi fitness dan Untuk Naive Bayes menggunakan fungsi fitness dari Naive Bayes.

3. Melakukan proses seleksi sebanyak N kromosom dari sejumlah P induk yang berasal dari populasi dengan seleksi roulette wheel. Kromosom dengan nilai fitness yang tinggi memiliki peluang yang lebih besar untuk terseleksi dan memilih pasangan induk secara acak untuk bereproduksi.
4. Melakukan proses pindah silang.
5. Melakukan proses mutasi.
6. Pergantian populasi yang lama dengan populasi generasi yang baru dengan cara memilih kromosom terbaik dari induk dan anak baru yang memiliki nilai fitness tertinggi setelah terjadinya seleksi, pindah silang dan mutasi.
7. Melihat apakah solusi yang didapatkan sudah memenuhi kriteria atau belum. Apabila solusi yang didapatkan belum mencapai kriteria maka kembali ke-2. Kriteria yang dimaksud disini adalah ketika nilai fitness terbaik sudah konvergen dari hasil generasi sebelumnya dan selanjutnya

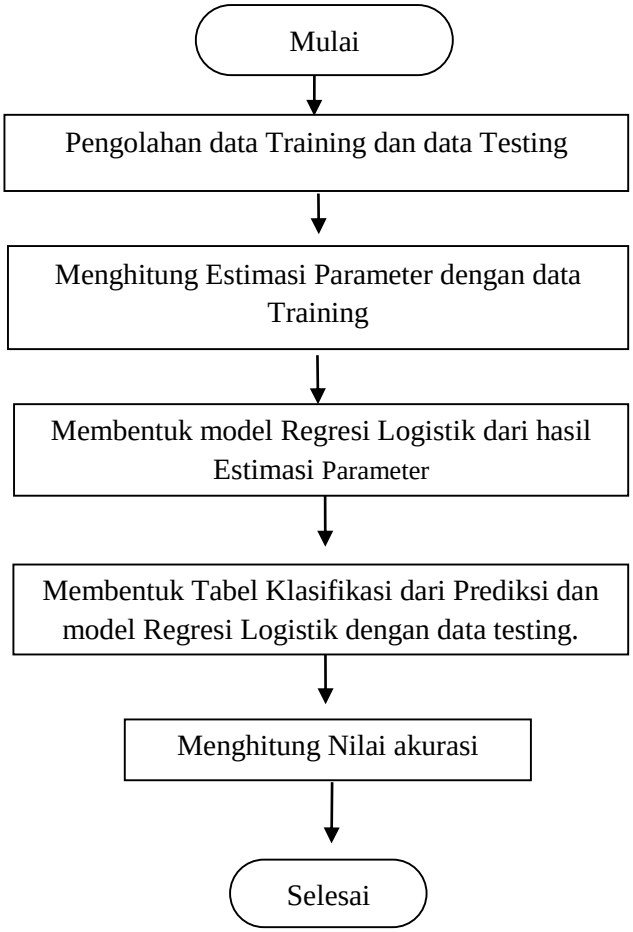
III. Pada langkah ini akan dilihat perbandingan pengklasifikasian terbaik menggunakan Akurasi, APER, *Sensitivity*, *Specificity*, dan *G-Mean*.

C. Pembahasan dan kesimpulan Penelitian

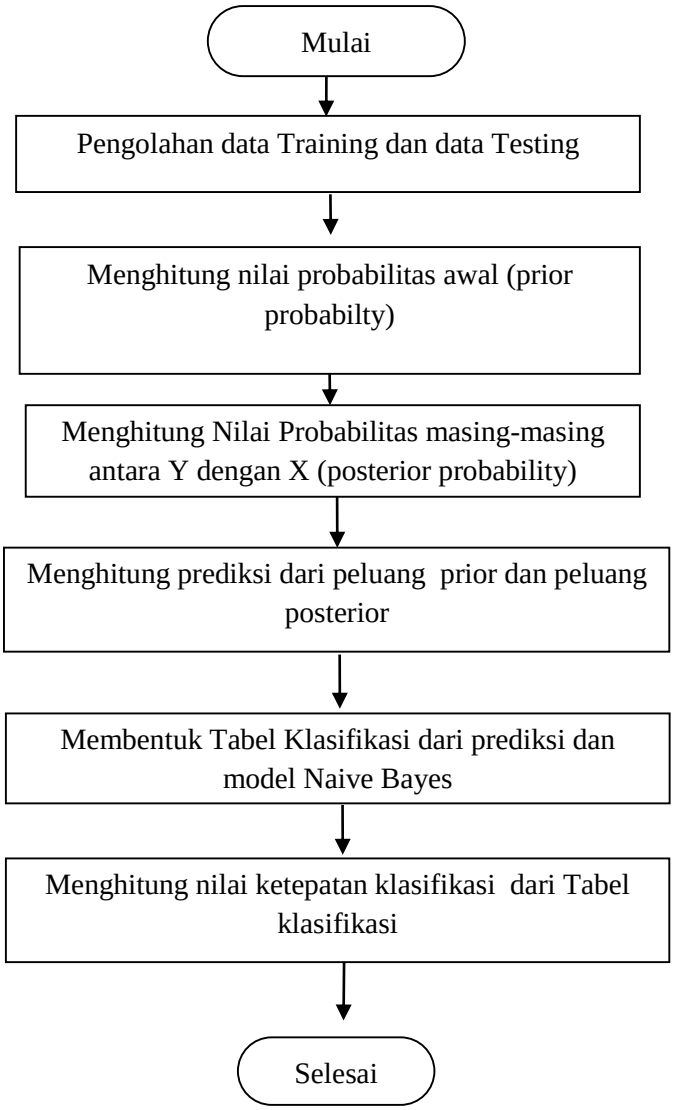
Pada langkah ini akan dibahas hasil penelitian yang telah dilakukan dan kemudian disimpulkan.

Langkah-langkah di atas dapat digambarkan menggunakan diagram alir, sebagai berikut :

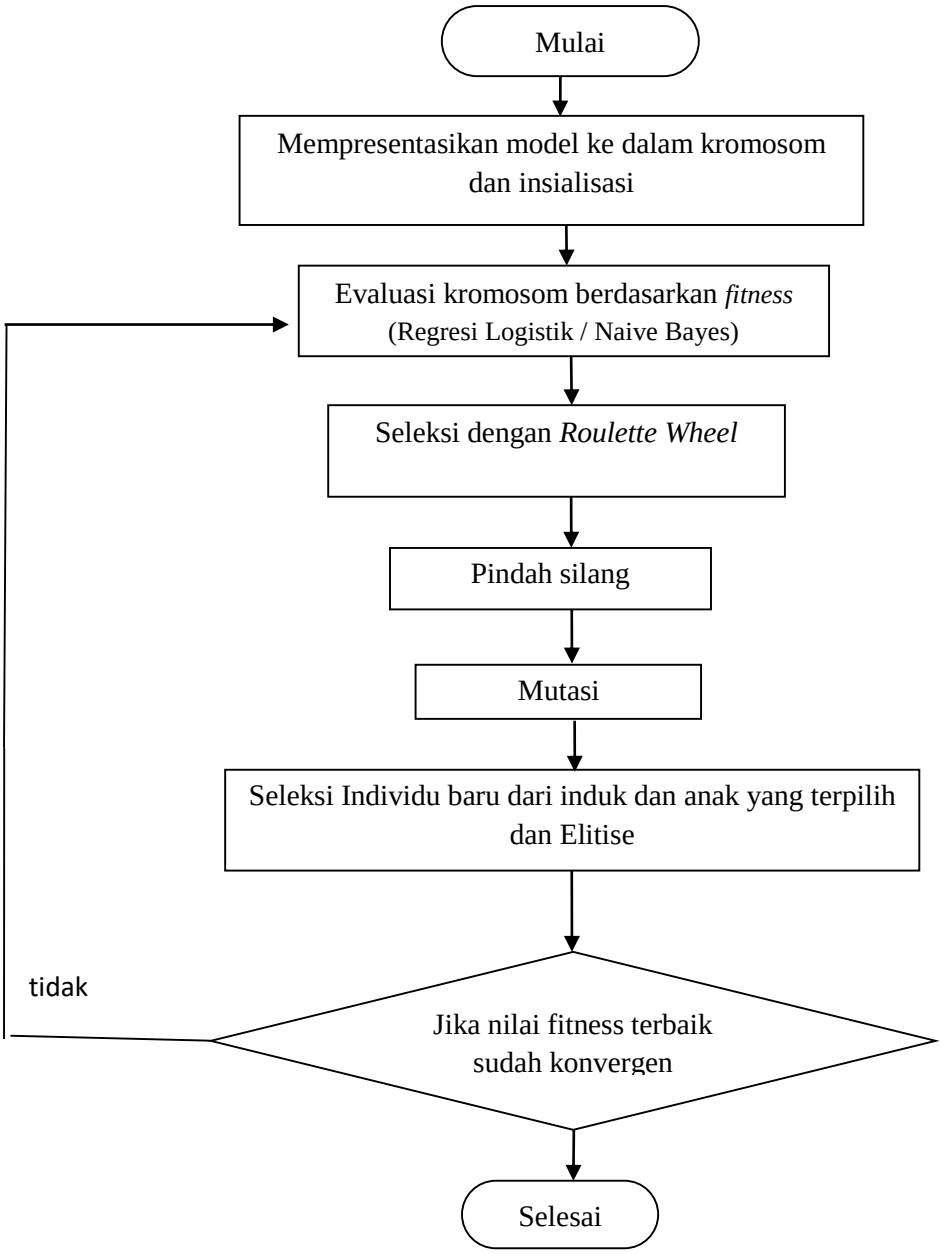
1. Diagram alir Regresi Logistik pada Gambar 1.
2. Diagram alir Naive Bayes pada Gambar 2.
3. Diagram alir Algoritma Genetika pada Gambar 3.



Gambar 3.1 Diagram alir Regresi Logistik



Gambar 3.2 Diagram alir Naive Bayes



Gambr 3.3 Diagram alir Algoritma Genetika

(*halaman ini sengaja dikosongkan*)

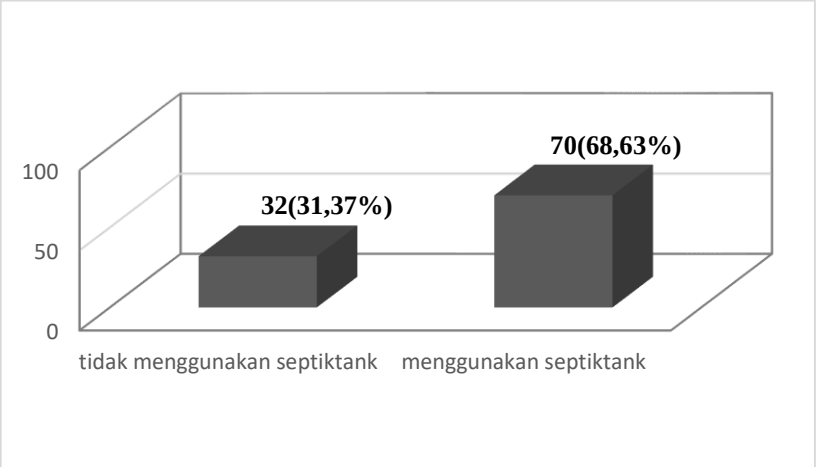
BAB 4
HASIL DAN PEMBAHASAN

Pada Penelitian ini bertujuan untuk mengetahui ketepatan klasifikasi dari Regresi Logistik, *Naïve Bayes*, dan penggunan Algoritma Genetika dalam meningkatkan ketepatan klasifikasi dari Regresi Logistik dan *Naïve Bayes* pada data klasifikasi. Dan penelitian ini juga melihat perbandingan klasifikasi dari Regresi Logistik dan *Naïve Bayes* dalam menggunakan Algoritma Genetika sebagai optimasi. Perbandingan yang digunakan pada penelitian ini dilihat dari akurasi, *misclassification*, *specitivity*, *sensitivity*, dan *G-Mean*.

Data pada penelitian ini, terdapat 102 pengamatan yang digunakan dalam penelitian ini. Kemudian data akan dibagi menjadi data Training dan data Testing. Peluang data Training dan data Testing pada penelitian ini adalah 90% : 10%, untuk data *Training* dengan peluang 90% dari total observasi dan untuk data *Testing* dengan peluang 10% dari total observasi. Kemudian dilakukan pengolahan masing-masing metode yakni Regresi Logistik, *Naïve Bayes*, dan Algoritma Genetika.

4.1 Analisis Karakteristik Data

Analisis deskriptif digunakan untuk memperoleh gambaran secara umum tentang karateristik pembuangan limbah domestik *blackwater* (*feces*). Adapun jumlah rumah tangga yang sudah menggunakan septictank yang sesuai dengan standar SNI dan belum menggunakan septictank adalah sebagai berikut.

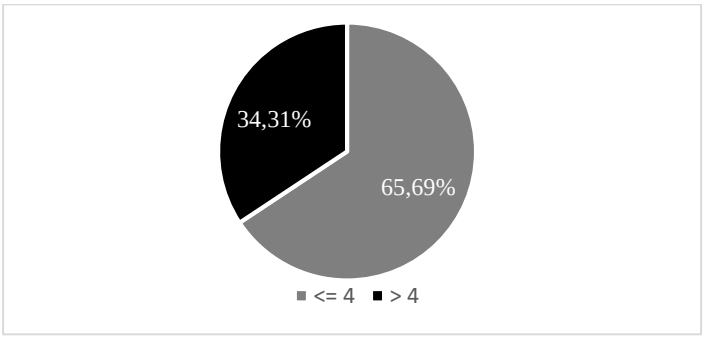


Gambar 4.1 Jumlah Pembuangan Limbah Domestik Rumah Tangga

Gambar 4.1 menunjukkan bahwa rumah tangga yang membuang limbah domestik *blackwater* (*feses*) dari tujuh Kelurahan meliputi Kelurahan Keputih, Gebang Putih, Nginden Jangkungan, Semolowaru, Mulyorejo, Kalisari, dan Kalijudan. Sebanyak 32 rumah tangga atau sekitar 31,37% rumah tangga sudah membuang limbah ke septictank yang sesuai dengan standar SNI sedangkan sebanyak 70 rumah tangga atau sekitar 68,63% belum membuang limbah *blackwater* (*feses*) ke septictank yang memenuhi standar SNI. Rata-rata rumah tangga masih membuang ke jamban cemplung, leher angsa dan jamban plengsengan yang landasannya tanah.

4.1.1 Karateristik Jumlah Anggota Keluarga

Jumlah anggota keluarga dikaitkan dengan faktor yang mempengaruhi pembuangan limbah domestik *blackwater* (*feses*). Keluarga yang memenuhi aturan BKKBN memiliki jumlah anak sebanyak dua orang, dengan total anggota keluarga sebanyak empat orang. Gambar 4.2 menunjukkan banyaknya anggota keluarga yang berada di tujuh Kelurahan.



Gambar 4.2 Jumlah Anggota Keluarga

Pada Gambar 4.2 menunjukkan bahwa sebagian besar di tujuh Kelurahan yang diambil sebagai sampel lebih banyak yang mempunyai anggota kurang dari sama dengan 4 yaitu sebanyak 65,69%. Anggota keluarga yang lebih dari empat anggota keluarga sebanyak 34,31%. Semakin banyaknyaknya anggota keluarga maka akan semakin sering juga tangki septik akan dikuras, hal ini mengakibatkan semakin meningkatnya pengeluaran rumah tangga. Anggota keluarga yang mempunyai jumlah di atas empat kemungkinan dipengaruhi ketika anaknya menikah sehingga menantu juga ikut tinggal dalam satu rumah. Hal yang lain juga adanya saudara atau nenek yang menjadi tanggungan keluarga.

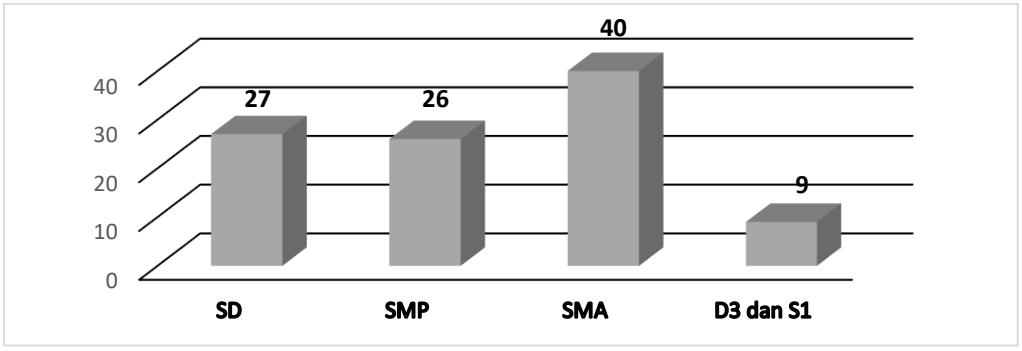
Tabel 4.1 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Jumlah Anggota Keluarga

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Jumlah Anggota Keluarga	≤ 4	Count	23	44	67
		Persentase Count	22,55%	43,14%	65,69%
		Expected Count	21	46	67
	> 4	Count	9	26	35
		Persentase Count	8,82%	25,49%	34,31%
		Expected Count	11	24	35
Total		Count	32	70	102
		Persentase Count	31,37%	68,63%	100%

Berdasarkan Tabel 4.1 tabulasi silang antara kepemilikan pembuangan limbah domestik dengan jumlah anggota keluarga. Diketahui bahwa jumlah anggota keluarga kurang dari empat lebih banyak yang membuang limbah domestik *blackwater* (feses) ke *septictank* sebanyak 23 rumah tangga dibanding yang mempunyai anggota keluarga lebih dari empat, hanya 9 kepala rumah tangga yang membuang limbah domestik *black-water* (feses) ke *sepictank* tetapi anggota keluarga yang dibawah sama dengan empat masih banyak yang masih membuang limbah tidak pada *septictank* yang memenuhi standar SNI dibandingkan yang lebih dari empat anggota keluarga.

4.1.2 Karateristik Pendidikan Kepala Keluarga

Pendidikan suatu keluarga sangat penting dalam peningkatan taraf hidup suatu keluarga, semakin tinggi pendidikan yang telah ditempuh semakin tinggi pula mendapatkan pekerjaan dan gaji yang layak. Pada Gambar 4.3 adalah pendidikan kepala keluarga yang telah diambil sebagai sampel.



Gambar 4.3 Pendidikan Kepala Keluarga

Berdasarkan Gambar 4.3 sebagian besar kepala rumah tangga di beberapa wilayah Surabaya Timur telah menempuh pendidikan hingga tahap SMA atau sederajat sebanyak 40 kepala keluarga. Kemudian disusul oleh SD atau sederajat sebanyak 27 kepala keluarga, kemudian SMP atau sederajat sebanyak 26 kepala keluarga. Dan yang paling sedikit Diploma dan Sarjana sebanyak 9 kepala keluarga.

Tabel 4.2 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Pendidikan Kepala Keluarga

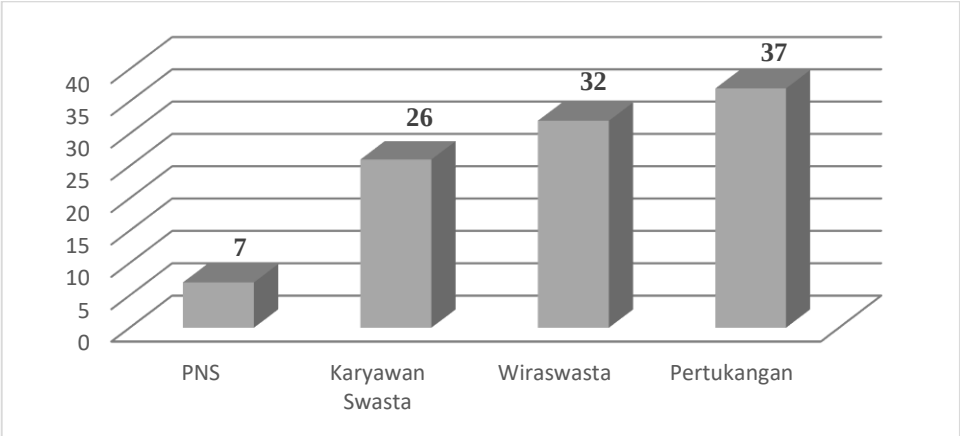
			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Pendidikan	SD	Count	6	21	27
		Persentase Count	5,90%	20,60%	26.50%
		Expected Count	8,5	18,5	27
	SMP	Count	8	18	26
		Persentase Count	7,80%	17,60%	25.50%
		Expected Count	8,2	17,8	26
	SMA	Count	12	28	40
		Persentase Count	11,80%	27,50%	39.20%
		Expected Count	12.5	27.5	40
	Diploma / Sarjana	Count	6	3	9
		Persentase Count	5.90%	2.90%	8.80%
		Expected Count	2.8	6.2	9
Total		Count	32	70	102
		Persentase Count	31.40%	68.60%	100%

Berdasarkan Tabel 4.2 terlihat bahwa presentase terbesar yang tidak mempunyai septitank sesuai dengan standar SNI pada saat kepala rumah tangga menempuh pendidikan terakhir SMA, dan SD. Hal ini mungkin karena tidak mengetahui bahwa setiap rumah tangga harus mempunyai pembuangan limbah yang sesuai dengan standar SNI. Diketahui yang paling sedikit tidak membuang limbah domestik *blackwater* (feses) tidak ke *septictank* adalah pada pendidikan lebih dari diploma termasuk Diploma dan Sarjana.

4.1.3 Karateristik Pekerjaan Kepala Keluarga

Pekerjaan suatu keluarga sangat penting perannya dalam peningkatan taraf hidup keluarga, karena dalam pekerjaan yang mempunyai gaji yang layak mempu memberikan pendapatan yang sesuai untuk keluarganya. Gambar 4.4 merupakan

pekerjaan kepala keluarga di tujuh Kelurahan yang telah diambil sebagai sampel adalah sebagai berikut.



Gambar 4.4 Pekerjaan Kepala Keluarga

Berdasarkan Gambar 4.4 sebagian besar kepala rumah tangga di tujuh Kelurahan mempunyai pekerjaan sebagai wira-swasta, yang terbagi dari pedagang hingga yang mempunyai bengkel sendiri sebanyak 32 kepala keluarga. Kemudian kepala keluarga yang sebagai karyawan swasta sebanyak 26 kepala rumah tangga. Pada variabel petani pertukangan, buruh dan jasa akan digabung menjadi satu karena mempunyai karakteristik pekerjaan yang hampir sama.

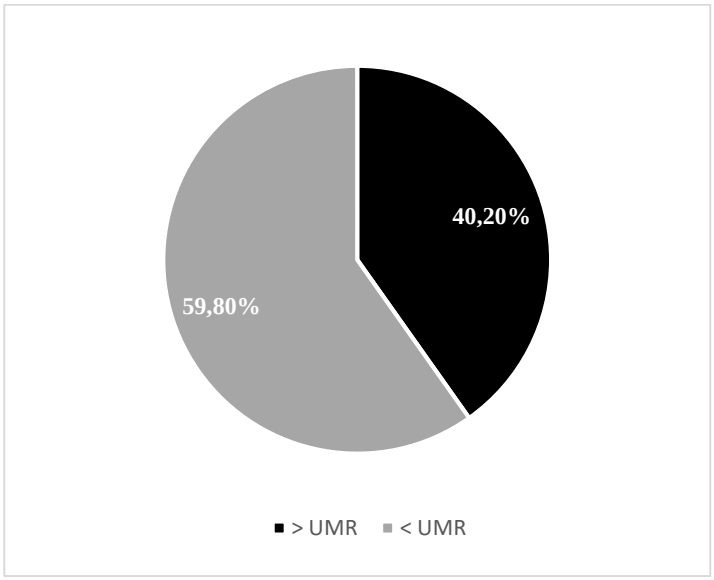
Tabel 4.3 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Pekerjaan Kepala Keluarga

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Pekerjaan Kepala Keluarga	PNS	Count	6	1	7
		Persentase Count	5,90%	1,00%	6.90%
		Expected Count	2,2	4,8	7
	Karyawan Swasta	Count	10	16	26
		Persentase Count	9,80%	15,70%	25,50%
		Expected Count	8,2	17,8	26
	Wiraswasta	Count	8	24	32
		Persentase Count	7,80%	23,50%	31,30%
		Expected Count	10	22	32
	Pertukangan	Count	8	29	37
		Persentase Count	7,80%	28,40%	36,20%
		Expected Count	11,6	25,4	37
Total		Count	32	70	102
		Persentase Count	31,40%	68,60%	100%

Berdasarkan Tabel 4.3 diketahui bahwa prosentase terbesar yang tidak mempunyai *septictank* sesuai dengan SNI yaitu pertukangan yang meliputi buruh, petani atau nelayan, tukang, dan jasa seperti calo sebesar 28,4% atau 29 rumah tangga. Hal ini terjadi mungkin karena tidak ada biaya untuk membangun *septic-tank* yang sesuai dengan standar SNI karena gaji pertukangan cenderung dibawah UMR atau dibawah Rp 1.750.000. Pada pekerjaan PNS yang paling sedikit tidak menggunakan *septictank* sesuai dengan SNI yaitu hanya 1%.

4.1.4 Karateristik Pendapatan Kepala Keluarga

Pendapatan kepala keluarga adalah menjadi tanggung jawab pemenuh kebutuhan untuk rumah tangga. Pendapatan kepala rumah tangga akan dibagi menjadi dua yaitu di atas UMR (Upah Minimal Regional) atau dibawah UMR. UMR kota Surabaya adalah sebesar 1.750.000 maka akan dijelaskan pendpatan kepala keluarga di tujuh Kelurahan pada Gambar 4.5.



Gambar 4.5 Pendapatan Kepala Keluarga

Berdasarkan Gambar 4.5 terlihat bahwa sudah 60% pen-dapatan kepala keluarga di tujuh Kelurahan sudah berada di atas UMR (Upah Minimal Regional) atau Rp1.750.000 dan 40% masih di bawah UMR. Hal ini sudah dapat dikatakan baik karena lebih banyak kepala keluarga yang sudah mempunyai gaji di atas UMR. Berikut ini tabulasi silang antara pendapatan kepala rumah tangga dengan kepemilikan pembuangan limbah.

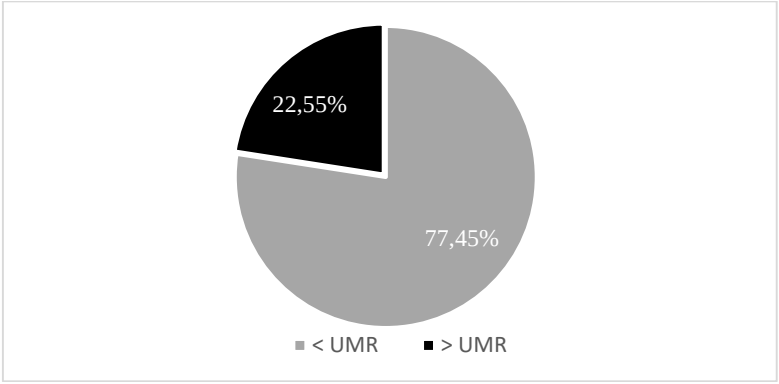
Tabel 4.4 Tabulasi Silang Kepemilikan Pembuangan LimbahDomestik dengan Pendapatan Kepala Keluarga

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Pendapatan	<UMR	Count	7	34	41
		Persentase Count	6,86%	43,10%	40,20%
		Expected Count	12.9	28.1	41
	>UMR	Count	25	36	61
		Persentase Count	8,80%	25,50%	59,80%
		Expected Count	19.1	41.9	61
Total		Count	32	70	102
		Persentase Count	31,37%	68,62%	100%

Berdasarkan Tabel 4.4 prosentase kepala keluarga yang mempunyai pendapatan di bawah UMR atau Rp 1.750.000 sangat sedikit sekali yang mempunyai *septictank* sesuai dengan standar SNI sebesar 6,89% atau 7 rumah tangga. Hal ini mungkin disebabkan tidak adanya biaya untuk membangun *septictank* yang sesuai dengan standar SNI, karena pendapatannya untuk memenuhi kebutuhan keluarga yang mungkin juga terbatas karena mempunyai pendapatan yang kurang.

4.1.5 Karateristik Pengeluaran Rumah Tangga

Pengeluaran merupakan jumlah rupiah yang dibelanjakan rumah tangga setiap bulannya untuk kebutuhan sehari-hari. Pe-ngeluaran kepala rumah tangga akan dibagi menjadi dua yaitu di atas UMR (Upah Minimal Regional) atau dibawah UMR. UMR kota Surabaya adalah sebesar 1.750.000. Presentase pengeluaran rumah tangga akan dijelaskan pada Gambar 4.6.



Gambar 4.6 Pengeluaran Rumah Tangga

Berdasarkan Gambar 4.6 sebagian besar pengeluaran rumah tangga di tujuh Kelurahan kurang dari Rp 1.750.000 dengan pro-sentase 77,45%. Pengeluaran rumah tangga yang diatas Rp 1.750.000 mempunyai prosentase sebesar 22,55%. Sedikitnya peng-eluaran setiap rumah tangga karena rata-rata jumlah anggota keluarga di tujuh Kelurahan kurang dari empat anggota, hal ini terjadi karena untuk menghemat pengeluaran agar dapat me-nabung membuat usaha atau sebagainya. Tabulasi silang antara pengeluaran rumah tangga dan kepemilikan pembuangan limbah dapat dilihat pada Tabel 4.5 sebagai berikut.

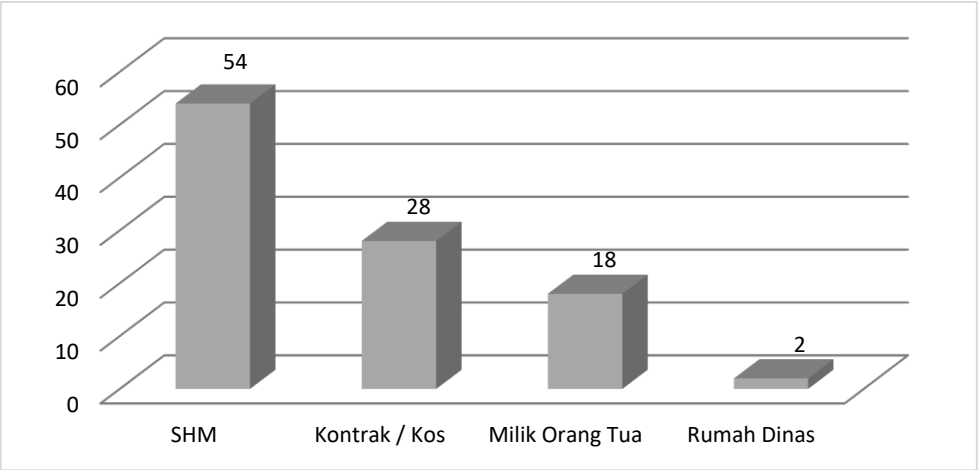
Tabel 4.5 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Pengeluaran Rumah Tangga

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Pengeluaran	< UMR	Count	19	60	79
		Persentase Count	18,60%	58,80%	77,45%
		Expected Count	24,8	54,2	79
	> UMR	Count	13	10	23
		Persentase Count	12,70%	9,80%	22,55%
		Expected Count	7,2	15,8	23
Total		Count	32	70	102
		Persentase Count	31,40%	68,60%	100%

Pada Tabel 4.5 diketahui pengeluaran rumah tangga kurang dari Rp 1.750.000, prosentase yang membuang limbah *blackwater* (feses) tidak ke *septictank* sebesar 58,8%. Terlihat bahwa yang mempunyai pengeluaran di bawah Rp 1.750.000 cenderung tidak mempunyai *septictank* sesuai dengan standar SNI, hal ini mungkin disebabkan tidak adanya biaya untuk membangun *septictank* yang sesuai dengan standar SNI. Pada rumah tangga yang mempunyai pengeluaran di atas Rp 1.750.000 yang sudah mempunyai *septictank* sesuai standar SNI cenderung lebih banyak daripada yang tidak ke *septictank.*

4.1.6 Karateristik Status Kepemilikan Rumah

Hal penting sebelum memutuskan untuk membeli properti, dan tanah adalah status kepemilikan rumah tersebut. Tempat tinggal adalah hal yang sangat penting perannya bagi anggota keluarga. Berikut ini akan dijelaskan pada Gambar 4.7 mengenai status kepemilikan rumah yang berada di tujuh Kelurahan.



Gambar 4.7 Status Kepemilikan Rumah

Berdasarkan Gambar 4.7 terlihat bahwa status kepemilikan rumah di tujuh Kelurahan yang paling banyak adalah SHM (Status Hak Milik) sebesar 54 rumah tangga, kontrak dan kos sebesar 28 rumah tangga, milik orangtua sebesar 18 rumah tangga, dan rumah dinas sebesar 2 rumah tangga. Tabulasi silang antara pengeluaran rumah tangga dan status kepemilikan rumah dapat dilihat pada Tabel 4.6 sebagai berikut.

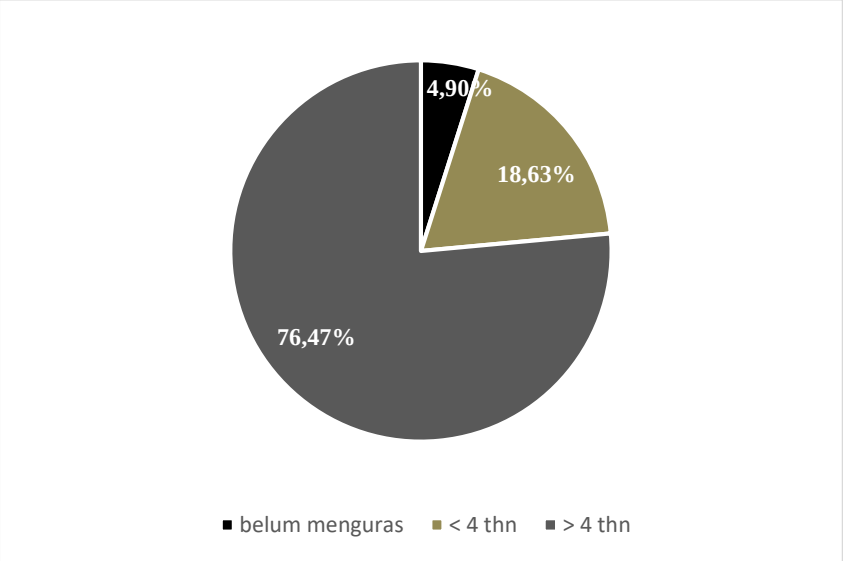
Tabel 4.6 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Status Kepemilikan Rumah

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Pekerjaan Kepala Keluarga	SHM	Count	17	37	54
		Persentase Count	24,32%	36,27%	52,94%
		Expected Count	16.9	37,1	54
	Kontrak / Kos	Count	7	21	28
		Persentase Count	6,86%	20,59%	27,45%
		Expected Count	8.8	19.2	28
	Milik Orang Tua	Count	6	12	18
		Persentase Count	5,88%	11,77%	17,65%
		Expected Count	5.6	12.4	18
	Rumah Dinas	Count	2	0	2
		Persentase Count	1,96%	0%	1,96%
		Expected Count	0.6	1,4	2
Total		Count	32	70	102
		Persentase Count	31,40%	68,60%	100%

Berdasarkan Tabel 4.6 dapat diketahui bahwa yang paling banyak tidak mempunyai *septictank* untuk membuang limbah domestik *blackwater* (feses) adalah rumah yang mempunyai status hak milik sebesar 36,27% atau 37 rumah tangga dan status kepemilikan rumah dinas tidak ada yang tidak membuang limbah domestik *blackwater* (feses) tidak ke *septictank*.

4.1.7 Karateristik Pengurasan Limbah Domestik

Pengurasan tangki septik merupakan aspek yang penting dari sanitasi lingkungan, lamanya pengurasan untuk tangki septik menurut ketentuan SNI (Standart Nasional Indonesia) selama tiga tahun mampu bertahan tanpa pengurasan, sehingga yang menjadi acuan lamanya pengurasan pada tangki septik adalah selama empat tahun. Waktu pengurasan di tujuh Kelurahan dapat dijelas-kan pada Gambar 4.8.



Gambar 4.8 Lama Waktu Pengurasan Tangki Septik

Pada gambar 4.8 menunjukkan bahwa rata-rata lama pengurasan di tujuh Kelurahan mempunyai prosentase sebesar 76,47%, sehingga masih banyak yang menguras tangki septik lebih dari empat tahun. Prosentase yang mengguras sesuai ketentuan SNI atau yang kurang dari sama dengan empat tahun sebesar 18,63% dan yang belum pernah dikuras sebesar 4,90% karena anggota keluarga menempati lokasi belum mencapai satu tahun. Tabulasi silang antara lama waktu pengurasan dan kepemilikan pembuangan limbah dapat dilihat pada Tabel 4.7 sebagai berikut.

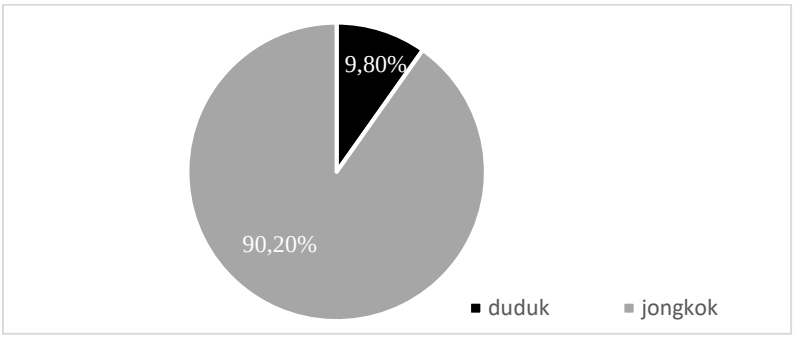
Tabel 4.7 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Lama Waktu Pengurasan

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Lama Waktu Pengurasan	belum pernah nguras	Count	0	5	5
		Persentase Count	0,00%	4,90%	4,90%
		Expected Count	1,6	3.4	5
	< 4 thn	Count	8	11	19
		Persentase Count	7,80%	10,80%	18,63%
		Expected Count	6	13	19
	>4 thn	Count	24	54	78
		Persentase Count	23,50%	52,90%	76,47%
		Expected Count	24,5	53,5	78
Total		Count	32	70	102
		Persentase Count	31,40%	68,60%	100%

Berdasarkan Tabel 4.7 diketahui yang menguras limbah domestik *blackwater* (feses) diatas empat tahun yang paling banyak dilakukan oleh rumah tangga yang tidak membuang limbah *blackwater* (feses) ke *septictank* sesuai dengan standar SNI, mempunyai prosentase 52,9% atau 54 rumah tangga. Hal ini disebabkan karena rumah tangga tersebut tidak mengetahui bila tangki septik harus dikuras paling lama empat tahun.

4.1.8 Karateristik Jenis Kloset

Pembuangan feses dalam MDGs sebagai sanitasi meliputi jenis kloset yang digunakan. Berikut ini adalah jenis kloset yang digunakan anggota keluarga di tujuh Kelurahan pada Gambar 4.9.



Gambar 4.9 Jenis Kloset

Berdasarkan Gambar 4.9 diketahui rumah tangga di tujuh Kelurahan rata-rata menggunakan kloset jongkok dengan prosentase sebesar 90,20% dan yang

menggunakan kloset duduk hanya sebesar 9,80%. Tabulasi silang antara jenis kloset dan kepemilikan pembuangan limbah dapat dilihat pada tabel 4.10 sebagai berikut.

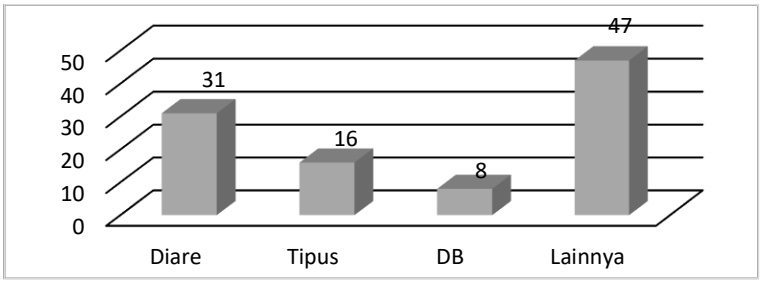
Tabel 4.8 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Jenis Kloset

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Jenis Kloset	Duduk	Count	4	6	10
		Persentase Count	3,90%	5,90%	9,80%
		Expected Count	3,1	6,9	10
	Jongkok	Count	28	64	92
		Persentase Count	27,50%	62,70%	90,20%
		Expected Count	28,9	63.1	92
Total		Count	32	70	102
		Persentase Count	31,40%	68,60%	100%

Berdasarkan Tabel 4.8 dapat diketahui rata-rata yang menggunakan kloset jongkok cenderung tidak membuang limbah domestik *blackwater* (feses) ke *septictank* yang sesuai dengan standar SNI dengan prosentase sebesar 62,7% atau 64 rumah tangga.

4.1.9 Karateristik Penyakit Anggota Keluarga

Sanitasi yang baik berhubungan dengan penyakit yang dialami oleh anggota keluarga. Semakin baik sanitasi maka se-makin sedikit kemungkinan anggota keluarga terkena penyakit. Penyakit yang dialami anggota keluarga di tujuh Kelurahan adalah pada Gambar 4.10 sebagai berikut.



Gambar 4.10 Penyakit yang Pernah dialami Anggota Keluarga

Gambar 4.10 menunjukkan penyakit yang pernah dialami oleh setiap anggota keluarga, penyakit yang berhubungan dengan sanitasi lingkungan adalah diare, tipus dan demam berdarah sedangkan yang lainnya meliputi, TBC, batuk dan

demam. Dari tujuh Kelurahan yang menderita penyakit diare (sakit perut) yang paling tinggi diantara tipus dan demam berdarah yaitu sebanyak 31 rumah tangga kemudian disusul oleh tipus sebanyak 16. Selain penyakit sanitasi lingkungan, penyakit lainnya juga banyak me-nyerang anggota keluarga. Hal ini disebabkan anggota keluarga kurang peduli terhadap sanitasi lingkungan sehingga banyak anggota keluarga yang terserang penyakit, hal lain juga yang mendukung anggota keluarga terserang penyakit seperti asupan gizi dan lain-lain. Tabulasi silang antara penyakit yang pernah dialami dan kepemilikan pembuangan limbah dapat dilihat pada Tabel 4.9 sebagai berikut.

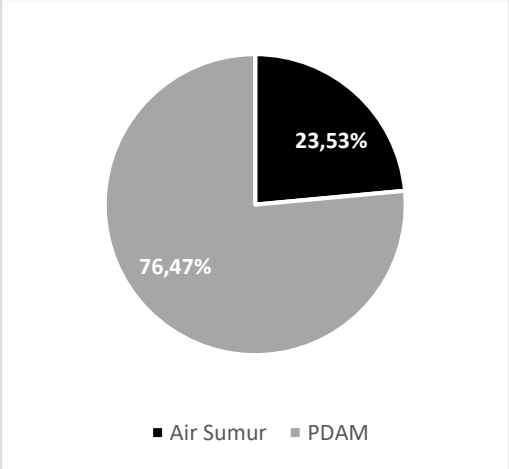
Tabel 4.9 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Penyakit yang Pernah dialami

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Pekerjaan Kepala Keluarga	Diare	Count	10	21	31
		Persentase Count	9,80%	31,40%	30,40%
		Expected Count	9,7	21,3	31
	Tipus	Count	1	15	16
		Persentase Count	1,00%	14,70%	15,70%
		Expected Count	5	11	16
	Demam Berdarah	Count	1	7	8
		Persentase Count	1,00%	6,90%	7,90%
		Expected Count	2,5	5.5	8
	Lain-lain	Count	20	27	47
		Persentase Count	19,60%	26,50%	26,10%
		Expected Count	14,7	32.3	47
Total		Count	32	70	102
		Persentase Count	31.40%	68,60%	100%

Berdasarkan Tabel 4.9 diketahui bahwa yang paling banyak terserang penyakit yang berhubungan dengan sanitasi adalah diare sebanyak 31,4% atau 21 rumah tangga pada rumah tangga yang tidak mempunyai *septictank* sesuai dengan standar SNI. Sebanyak 14,7% terserang tipus, 6,9% terserang demam berdarah, dan 26,5% terserang penyakit yang lain meliputi flu, demam, dan syaraf. Hal ini sangat jelas, bila yang mempunyai sanitasi baik maka akan kecil kemungkinan terserang penyakit dibandingkan dengan yang tidak mempunyai sanitasi yang baik. Maka pem-buangan limbah domestik *blackwater* (feses) sangat berpengaruh dengan kesehatan.

4.1.10 Karateristik Air Mandi dan Mencuci

Air merupakan salah satu dari sanitasi lingkungan, begitu juga dengan air yang digunakan anggota keluarga untuk mandi dan mencuci. Air yang digunakan anggota keluarga di tujuh Kelurahan dapat dijelaskan pada Gambar 4.11 sebagai berikut.



Gambar 4.11 Air Cuci dan Mandi yang digunakan Rumah Tangga

Gambar 4.11 menunjukkan air yang digunakan setiap anggota keluarga di tujuh Kelurahan untuk mencuci dan mandi rata-rata rumah tangga di tujuh Kelurahan sudah menggunakan air PDAM dengan prosentase sebesar 76,47% dan masih ada anggota keluarga yang masih menggunakan air tanah atau air sumur dengan prosentase sebesar 23,53%. Tabulasi silang antara air mencuci dan mandi dengan kepemilikan pembuangan limbah dapat dilihat pada Tabel 4.10 sebagai berikut.

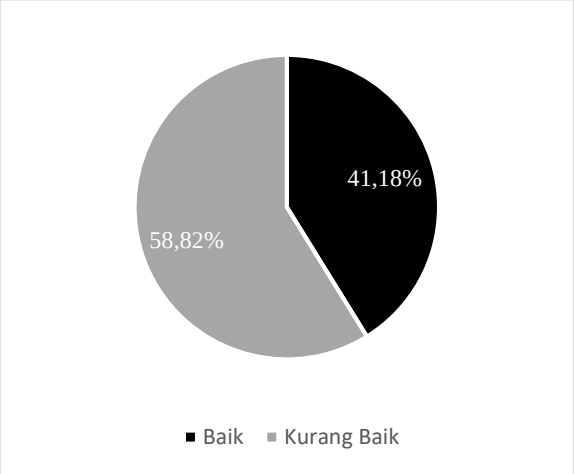
Tabel 4.10 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Air Mandi dan Mencuci

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Air Mandi Cuci	Air Tanah / Sumur	Count	9	15	24
		Persentase Count	8,80%	14,70%	23,53%
		Expected Count	7.5	16.5	24
	PDAM	Count	23	55	78
		Persentase Count	22,50%	53,90%	76,47%
		Expected Count	24.5	53,5	78
Total		Count	32	70	102
		Persentase Count	31,40%	68,60%	100%

Pada Tabel 4.10 diketahui prosentase yang masih menggunakan air tanah yaitu pada rumah tangga yang membuang limbah *blackwater* (feses) tidak ke *septictank* sebesar 14,7% lebih besar daripada yang membuang limbah *blackwater* (feses) ke *septictank* sesuai dengan standar SNI yaitu sebesar 8,8%. Begitu juga dengan yang menggunakan air PDAM sebesar 53,9% atau 55 rumah tangga masih membuang limbah domestik *blackwater* (feses) ke *septictank* tidak sesuai dengan standar SNI.

4.1.11 Karateristik Perilaku Sanitasi

Perilaku sanitasi hal dasar dari masing-masing pribadi setiap anggota keluarga, apakah di keluarga tersebut sudah mem-punyai perilaku yang baik atau kurang baik. Penilaian perilaku sanitasi dinilai dari beberapa aspek, yaitu apakah menggunakan air bersih untuk sehari hari, pencahayaan di kamar mandi, ventilasi di kamar mandi dan lain-lain. Perilaku sanitasi akan dijelaskan pada Gambar 4.12 sebagai berikut.



Gambar 4.12 Perilaku Sanitasi

Gambar 4.12 menunjukkan bahwa dari tujuh Kelurahan yang diambil sebagai sampel sebanyak 58,82% rumah tangga mem-punyai perilaku sanitasi yang kurang baik dan 41,18% yang mempunyai perilaku sanitasi yang baik. Tabulasi silang antara perilaku sanitasi dengan kepemilikan pembuangan limbah dapat dilihat pada tabel 4.11 sebagai berikut.

Tabel 4.11 Tabulasi Silang Kepemilikan Pembuangan Limbah Domestik dengan Perilaku Sanitasi

			Pembuangan Limbah		Total
			Menggunakan Septiktank	Tanpa Septiktank	
Prilaku Sanitasi	Baik	Count	20	22	42
		Persentase Count	19,60%	21,60%	41,18%
		Expected Count	13.2	28.8	42
	Kurang Baik	Count	12	48	60
		Persentase Count	11,80%	47,00%	58,82%
		Expected Count	18.8	41.2	60
Total		Count	32	70	102
		Persentase Count	31,40%	68,60%	100%

Berdasarkan Tabel 4.11 terlihat prosentase rumah tangga yang mempunyai perilaku sanitasi yang kurang baik cenderung tidak mempunyai *septictank* yang sesuai dengan standar SNI sebesar 47,00% atau 48 rumah tangga dan yang mempunyai *septictank* sesuai dengan standar SNI sebesar 11,80% atau 12 rumah tangga. Hal ini terjadi karena masyarakat kurang sadar pentingnya sanitasi yang baik sehingga masih banyak yang berperilaku sanitasi tidak baik dan tidak mengetahui mengenai pem-buangan limbah yang sesuai dengan standar SNI.

4.2 Korelasi Antar Variabel Independen

Hubungan antar variabel dapat diketahui berdasarkan nilai korelasi. Metode mencari korelasi yang digunakan adalah Spearsman's rho. Tujuan untuk melihat korelasi ini adalah untuk melihat apakah antar variabel independennya mengalami multikolinieritas atau adanya hubungan yang signifikan antar variabel independennya. Sehingga digunakan nilai korelasi Spearsman's rho diantara variabel Independen dapat dilihat pada Tabel 4.12.

Tabel 4.12 Korelasi antar Variabel independen

		X2	X3	X4	X5	. . .	X9	X10	X11
X1	korelasi	0,055	-0,077	**0,213**	**0,252**		**-0,212**	0,060	-0,067
	p-value	0,581	0,441	**0,031**	**0,010**		**0,032**	0,548	0,506
X2	korelasi	1	**-0,505**	**0,257**	**0,272**		-0,097	0,038	**-0,300**
	p-value		**0,000**	**0,009**	**0,006**		0,332	0,705	**0,002**
X3	korelasi		1	-0,191	-0,177		-0,001	-0,047	0,150
	p-value			0,055	0,075		0,996	0,639	0,132

Lanjutan Tabel 4.12

		X2	X3	X4	X5	.	.	.	X9	X10	X11
X4	korelasi			1	**0,442**				0,011	-0,125	**-0,442**
	p-value				**0,000**				0,916	0,211	**0,000**
X5	korelasi				1				**-0,338**	-0,033	**-0,407**
	p-value								**0,001**	0,745	**0,000**
X6	korelasi								-0,125	-0,038	**-0,206**
	p-value								0,211	0,704	**0,038**
X7	korelasi								**0,254**	-0,151	0,049
	p-value								**0,010**	0,129	0,625
X8	korelasi								0,067	0,050	0,126
	p-value								0,502	0,616	0,207
X9	korelasi								1	-0,048	0,191
	p-value									0,629	0,054
X10	korelasi									1	-0,088
	p-value										0,377

Hasil Korelasi pada Tabel 4.12 terlihat bahwa angka yang bercetak tebal tersebut memiliki p-value < 0,05, sehingga terdapat beberapa variabel independen yang berkorelasi secara signifikan. Sehingga ini juga membuktikan bahwa adanya multikolinieritas antar variabel indenpenden.

4.3 Regresi Logistik Biner

Regresi Logistik merupakan salah satu metode dalam mengatasi data klasifikasi. Dalam metode ini penentuan estimasi model yang digunakan adalah untuk mengetahui faktor mana yang mempengaruhi secara signifikan untuk model terbaik sehingga dapat dilihat dari akurasinya. Maka dilakukan pembagian data Training 90% dan data Testing sebesar 10 % dari 102 observasi. Pada Tabel 4.13 akan dilihat estimasi secara keseluruhan dengan menggunakan data Training untuk melihat model yang siginifikan sebelum penentuan klasifikasinya, sebagai berikut:

Tabel 4.13 Estimasi Parameter Regresi Logistik

Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	11,566	2095,101	0,006	0,996
X1 (1)	**4,536**	**1,508**	**3,007**	**0,003**
X2 (1)	0,749	0,986	0,760	0,448
X2 (2)	2,511	1,322	1,900	0,057
X2 (3)	3,165	1,627	1,944	0,052
X3 (1)	3,354	2,065	1,624	0,104

Lanjutan Tabel 4.13

Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
X3 (2)	3,493	1,887	1,852	0,064
X3 (3)	**4,938**	**2,130**	**2,319**	**0,020**
X4 (1)	-0,902	0,922	-0,977	0,328
X5 (1)	**-3,835**	**1,456**	**-2,635**	**0,008**
X6 (1)	**2,029**	**1,027**	**1,977**	**0,048**
X6 (2)	2.218	1,359	1,632	0,103
X6 (3)	-15,050	4612,203	-0,003	0,997
X7 (1)	-20,268	2095,099	-0,010	0,992
X7 (2)	-18,438	2095,099	-0,009	0,993
X8 (1)	1,424	1,664	0,856	0,392
X9 (1)	1,947	1,681	1,158	0,247
X9 (2)	1,375	2,122	0,648	0,517
X9 (3)	-1,824	1,043	-1,749	0,080
X10 (1)	0,797	0,907	0,879	0,379
X11 (1)	**2,106**	**0,955**	**2,205**	**0,027**

Dengan pengujian alpha 5 % atau 0,05, maka kriteria yang diperoleh untuk $p - value < 0{,}05$ adalah variabel yang signifikan yang akan dibentuk menjadi model logitnya. Dari Tabel 4.13 terlihat bahwa variabel yang signifikan adalah X_1, X_3 , X_5, X_6, dan X_{11}. Jika dibuat model dari keseluruhan variabelnya, maka model regresi Logistiknya menjadi :

$$\pi(x) = \frac{\exp\left(4{,}536X_1(1) + 0{,}749X_2(1) + \cdots + 0{,}797X_{10}(1) + 2{,}106X_{11}(1)\right)}{1 + \exp\left(4{,}536X_1(1) + 0{,}749X_2(1) + \cdots + 0{,}797X_{10}(1) + 2{,}106X_{11}(1)\right)}$$

Kemudian menghitung ketepatan klasifikasi dari model Regresi Logistik Biner dengan data Testing 10 %. Sebelumnya perhitungan prediksi yang digunakan adalah peluang dengan nilai *Cut off* sebesar 0,3 akan bernilai $P(Y = 0)$ dan nilai *Cut off* sebesar 0,7 akan bernilai $P(Y = 1)$. Berdasarkan hasil prediksi dan aktual pada Lampiran 3. Diberikan Tabel 4.14 berikut ini.

Tabel 4.14 Tabel Klasifikasi dari model Regresi Logistik Biner

Real	**Prediction**	
	0	**1**
0	2	2
1	3	4

Dari Tabel 4.14 di atas didapatkan ketepatan klasifikasi dari model Regresi Logistik Biner pada Tabel 4.15 menggunakan persamaan 2.16, 2.17, 2.18, 2.19, dan 2.120.

Tabel 4.15 Ketepatan Klasifikasi model Regresi Logistik Biner

	Accuracy	APER	Specificity	Sensitivity	G-Mean
Binary Logistic Regression	54,55%	44,45%	0,57	0,50	0,54

Pada Tabel 4.15, terlihat bahwa dari keakuratan model didapatkan akurasi sebesar 54,55%, dengan kesalahannya 44,45%. Kemudian juga terdapat Specificity nya atau *true positive* sebesar 0,57 dan sensitivity atau *true negative* sebesar 0,50. Untuk melihat apakah datanya mengalami *unbalance*, dapat dilihat dari nilai G-Mean sebesar 0,54, berarti nilai keakuratan dari data *balance*-nya sebesar 0,54 atau 53,57%.

4.3.1 Regresi Logistik Biner Menggunakan Seleksi Variabel dengan Metode Backward

Sebelumnya sudah dilihat estimasi parameter yang telah dilakukan pada Tabel 4.13, dan terdapat tiga variabel yang secara signifikan mempengaruhi model yaitu X_1, X_3, X_5, X_6 dan X_{11}. Kemudian akan dilakukan model Regresi Logistik dalam menseleksi variabel menggunakan metode *Backward*. Menggunakan data Training sebesar 90 % dari 102 observasi, maka akan dilihat bagaimana tahapan penseleksian yang dilakukan menggunakan metode *backward* pada Tabel 4.16 :

Tabel 4.16 Tahapan Seleksi Variabel metode Backward

Tahap	Seleksi	Model Akhir Persamaan	AIC
0	Null	Y = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11	98,87
1	X10	Y = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X11	97,65
2	X10, X8	Y = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X9 + X11	96,49
3	X10, X8, X2	Y = X1 + X3 + X4 + X5 + X6 + X7 + X9 + X11	94,98
4	X10, X8, X2, X6	Y = X1+ X3 + X4 + X5 + X7 + X9 + X11	93,14
5	X10, X8, X2, X6, X4	Y = X1 + X3 + X5 + X7 + X9 + X11	91,85
6	X10, X8, X2, X6, X4 , X3	Y = X3 + X5 + X7 + X9 + X11	91,63

Dari Tabel 4.16 didapatkan hasil seleksi metode *backward* dengan mengeluarkan X_2, X_3, X_4, X_6, X_8, dan X_{10} dengan menghasilkan nilai AIC paling kecil yaitu 103,15. Dimana Tahap-tahap penseleksian metode *backward* berada di Lampiran. Maka variabel yang terpilih dari hasil seleksi *backward* adalah X_1, X_5,

X_7, X_9, dan X_{11}. Tahap selanjutnya adalah melakukan uji Parsial untuk mengetahui variabel mana yang mempengaruhi secara signifikan pada hasil seleksi metode *backward*. Kemudian pada Tabel 4.17 pengujian secara parsial pada penyeleksian metode *backward*.

Tabel 4.17 Uji Parsial seleksi Backward

Tahap	Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
6	(Intercept)	18,4067	1337,949	0,014	0,989
	X1 (1)	**2,3141**	**0,909**	**2,545**	**0,011**
	X5 (1)	**-2,7810**	**0,970**	**-2,866**	**0,004**
	X7 (1)	-19,1703	1337,949	-0,014	0,989
	X7 (2)	-17,9067	1337,949	-0,013	0,989
	X9 (1)	**2,9672**	**1,381**	**2,149**	**0,032**
	X9 (2)	1,1130	1,764	0,631	0,528
	X9 (3)	-1,3839	0,801	-1,727	0,084
	X11 (1)	**1,6884**	**0,678**	**2,489**	**0,013**

*dengan alpha = 5 % = 0,05

Terlihat pada Tabel 4.17 bahwa terdapat pada tahap keenam dalam penseleksian terakhir metode *backward*. Sehingga modelnya sebagai berikut:

$$\pi(x) = \frac{\exp(18,407 + 2,314X_1(1) - 2,781X_5(1) - 19,170X_7(1)\cdots)}{1+\exp(18,407 + 2,314X_1(1) - 2,781X_5(1) - 19,170X_7(1)\cdots)}$$

$$= \frac{\exp(\cdots - 17,907X_7(2) + 2,967X_9(1) + 1,113X_9(2) - 1,384X_9(3) + 1,688X_{11}(1))}{1+\exp(\cdots - 17,907X_7(2) + 2,967X_9(1) + 1,113X_9(2) - 1,384X_9(3) + 1,688X_{11}(1))}$$

Kemudian menghitung ketepatan klasifikasi dari model Regresi Logistik Biner dengan data Testing 10 %. Sebelumnya perhitungan prediksi yang digunakan adalah peluang dengan nilai *Cut off* sebesar 0,3 akan bernilai $P(Y = 0)$ dan nilai *Cut off* sebesar 0,7 akan bernilai $P(Y = 1)$. Berdasarkan hasil prediksi dan aktual pada Lampiran 3. Diberikan Tabel 4.18 berikut ini.

Tabel 4.18 Tabel Kontingensi Regresi Logistik dengan seleksi *backward*

Real	**Prediction**	
	0	**1**
0	2	1
1	2	6

Dari Tabel 4.18, dapat dihitung ketepatan klasifikasi dari Regresi Logistik dengan seleksi *backward* menggunakan persamaan 2.16, 2.17, 2.18, 2.19, dan 2.120.. Berikut ditampilkan hasil ketepatan klasifikasi pada Tabel 4.19.

Table 4.19 Ketepatan Klasifikasi Regresi Logistik Seleksi *Backward*

	Accuracy	APER	Specificity	Sensitivity	G-Mean
Regresi Logistik-seleksi backward	72,73%	27,27%	0,86	0,50	0,68

Pada Tabel 4.19, terlihat bahwa dari keakurat model didapatkan akurasi sebesar 72,73 %, dengan kesalahannya 27,27 %. Kemudian juga terdapat Specificity nya atau *true positive* sebesar 0,86 dan sensitivity atau *true negative* sebesar 0,50. Untuk melihat apakah datanya mengalami *unbalance*, dapat dilihat dari nilai G-Mean sebesar 0,68, karena nilai G-mean > 0,5, berarti datanya cukup balance pada tipe responsnya.

4.4 Regresi Logistik Biner Menggunakan Seleksi Variabel dengan Algoritma Genetika

Tahapan pertama dari metode algoritma genetika adalah membentuk kromosom awal. Kromosom untuk seleksi variabel berisikan gen yang berupa variabel independen, tipe kromosom yang sesuai untuk seleksi variabel adalah tipe biner. Variabel independen yang akan masuk dalam fungsi Regresi Logistik dikode 1, sedangkan yang tidak masuk dikode dengan 0. Panjang kromosom sesuai dengan jumlah variabel independen, dimana variabel independennya adalah data septictank di wilayah Surabaya Timur dengan panjang kromosomnya adalah 11 variabel independen. Pada Gambar 4.13 merupakan ilustrasi dari kromosom awal yang dibentuk sebagai berikut :

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	1	1	1	1	1	1	1	1	1	1

Gambar 4.13 Ilustrasi Kromosom awal

Pada penelitian ini telah ditentukan ukuran populasi sebanyak 100. Jadi ada sebanyak 100 kromosom yang akan dibangkitkan. Tabel 4.20 berikut ini menyajikkan ilustrasi 100 kromosom yang dibangkitkan untuk populasi awal.

Tabel 4.20 Ilustrasi Populasi Awal pada 100 Kromosom untuk Regresi Logistik

Kromosom	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	1	1	1	1	1	1	1	1	1	1	1
2	0	0	1	0	1	0	1	0	1	1	1
.											
.											
.											
100	0	0	1	1	1	0	1	1	1	1	1

Pada Tahap kedua adalah mengevaluasi masing-masing kromosom yang telah dibentuk menggunakan fungsi fitnes yang telah ditentukan. Fungsi *fitness* yang digunakan untuk menyelesaikan permasalahan seleksi variabel pada Regresi Logistik adalah kesalahan klasifikasi pembentukan peluang Regresi Logistik. Hasil evaluasi masing-masing kromosom pada populasi awal dapat dilihat pada Tabel 4.21.

Tabel 4.21 Ilustrasi Nilai *Fitness* pada masing-masing Kromosom

Kromosom	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Nilai Fitnes
1	1	1	1	1	1	1	1	1	1	1	1	0,500
2	0	0	1	0	1	0	1	0	1	1	1	0,134
.												.
.												.
.												.
100	0	0	1	1	1	0	1	1	1	1	1	0,402

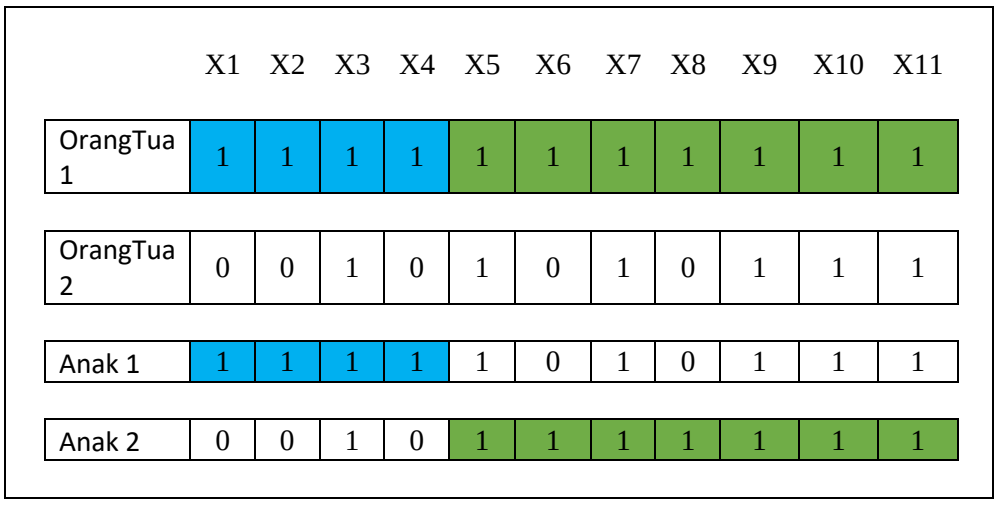
Kemudian pada tahapan ketiga adalah memilih kromosom dari populasi awal untuk dijadikan orang tua pada generasi berikutnya menggunakan metode RWS. Tahapan metode RWS secara umum adalah menentukan proporsi nilai fitnes, menentukan nilai komulatifnya, membangkitkan bilangan random yang beridistribusi uniform antar 0-1, kemudian membandingkan nilai komulatif dengan bilangan random tersebut. Apabila nilai komulatif kromosom ke-*i* lebih besar dari bilangan randomnya, maka kromosom ke-*i* tersebut terpilih menjadi orang tua untuk generasi berikutnya. Berikut adalah ilustrasi proses RWS pada Tabel 4.22.

Tabel 4.22 Ilustrasi Proses RWS

Kromosom	Nilai *Fitness*	Proporsi Nilai *fitness*	*fitness* Komulatif	Bilangan Random
1	0,500	0,022	0,022	0.250
2	0,134	0,006	0,028	0.844
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
100	0,402	0,018	1,000	0.963

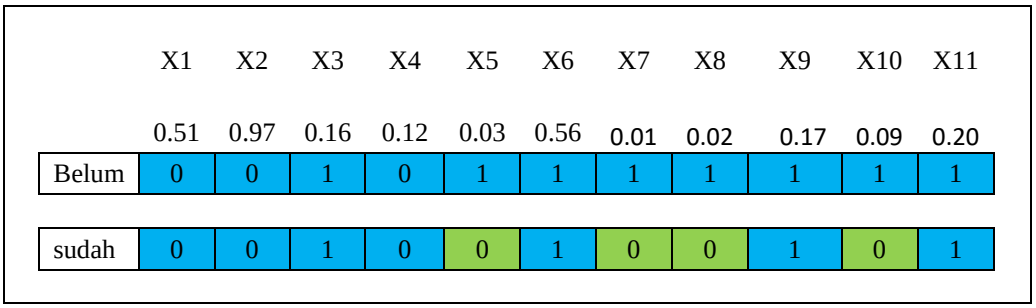
Berdasarkan Tabel 4.22 dapat diketahui nilai bilangan *random* pertama adalah 0,250 dan untuk nilai *fitness* kumulatif pertama sebesar 0,022 sehingga kromosom pertama berubah menjadi individu baru. Dan begitu seterusnya untuk kromosom selanjutnya.

Proses Selanjutnya Tahap keempat adalah melakukan pindah silang, yang tujuannya untuk mengasilkan kromosom baru dari perpaduan 2 kromosom orang tua. Pindah silang dalam penelitian ini menggunakan pindah silang satu titik atau *crossover* sederhana. Penetuan titik ini melibatkan probabilitas pindah silang (Ps) = 0,8. Proses penentuan titik potong dilanjutkan dengan membangkitkan bilangan *random* antara 0 sampai 1 kemudian membadingkan nilai bilangan *random* dengan nilai probabilitas pindah silang (Ps) = 0,8, jika bilangan *random* lebih kecil dari probabilitas pindah silang maka orangtua yang terpilih dari individu satu dan individu dua akan dikawinkan dan keturunanya akan menjadi individu baru. Sehingga misalkan titik potong ditentukan secara *random* serta ditemukan pada X_4, maka mulai X_5 akan ditukarsilangkan antara variabel pada kromosom bapak dan ibu. Berikut ditunjukkan Ilustrasi proses pindah silang pada Gambar 4.14.



Gambar 4.14 Ilustrasi Proses Pindah Silang

Kemudian pada tahapan kelima adalah mutasi, metode yang digunakan adalah metode uniform. Tujuan mutasi adalah untuk mendapatkan keberagaman gen atau variabel pada kasus ini. Tahapan mutasi adalah dengan membangkitkan bilangan random pada setiap variabel dan membandingkan dengan peluang mutasi sebesar 0.1. Apabila nilai bangkitan pada suatu variabel kurang dari peluang mutasi, maka dilakukan proses mutasi pada gen tersebut. Proses mutasi yaitu dengan mengubah kode 1 menjadi 0 atau 0 menjadi 1 pada variabel yang dilakukan proses mutasi. Pada Gambar 4.15 akan ditunjukkan ilustrasi tahapan dalam melakukan mutasi.



Gambar 4.15 Ilustrasi tahapan Mutasi

Pada Gambar 4.15 terlihat bahwa X_5, X_7, X_8, dan X_{10} mengalami mutasi karena probabilitis bilangan randomnya lebih keci dari probabilitis mutasi (*Pm*) = 0.1.

Tahapan keenam adalah melakukan elitisme, yaitu menggandakan kromosom yang memiliki nilai fitnes paling kecil. Jumlah kromosom yang digandakan adalah 2. Berdasarkan perhitungan nilai fitnes tiap kromosom pada Tabel 4.23 diketahui bahwa nilai fitnes paling kecil adalah 0,134.

Tabel 4.23 Kromosom dan Nilai Fitness pada seleksi Algoritma Genetika pada Regresi Logistik

Kromosom	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Nilai *Fitness*
1	0	1	0	0	0	0	1	1	1	1	1	0.134
2	0	0	1	0	1	0	1	0	1	1	1	0,134
.						.						.
.						.						.
.						.						.
100	0	0	1	1	1	0	1	1	1	1	1	0,134

Dari Tabel 4.23 diberikan 100 kromosom dengan nilai *fitness*-nya dan hasil seleksi Algoritma Genetika di atas didaptakan 5 variabel yang terpilih yaitu X_2, X_7, X_8, X_9, X_{10} dan X_{11} yang terdapat pada Kromosom pertama dengan nilai kesalahan 0,134.

Kemudian menghitung ketepatan klasifikasi dari model Regresi Logistik dengan data Testing 10 %. Sebelumnya perhitungan prediksi yang digunakan adalah peluang dengan nilai *Cut off* sebesar 0,3 akan bernilai $P(Y = 0)$ dan nilai *Cut off* sebesar 0,7 akan bernilai $P(Y = 1)$. Berdasarkan hasil prediksi dan aktual pada Lampiran 3. Diberikan Tabel 4.24 berikut ini.

Tabel 4.24 Klasifikasi Algoritma Genetika-Regresi Logistik

Real	**Prediction**	
	0	**1**
0	3	0
1	1	7

Dari Tabel 4.24, kita dapat mengetahui ketepatan klasifikasinya untuk melihat keakuratan dari variabel yang terpilih atau faktor yang terpilih itu cukup mewakili dalam menganalisa faktor yang mempengaruhi pembuangan limbah rumah tangga yang memiliki septiktank atau tidak. Berikut Tabel 4.25 akan merangkum hasil ketepatan klasifikasinya.

Tabel 4.25 Ketepatan Klasifikasi GA - Reglog

	Accuracy	APER	Specificity	Sensitivity	G-Mean
Algoritma Genetika-Regresi Logistik	90,91%	9,09%	1	0,75	0,875

Pada Tabel 4.25, terlihat bahwa dari keakurat model didapatkan akurasi sebesar 90,91%, dengan kesalahannya 9,09 %. Kemudian juga terdapat Specificity nya atau *true positive* sebesar 1 dan sensitivity atau *true negative* sebesar 0,75. Untuk melihat apakah datanya mengalami *unbalance*, dapat dilihat dari nilai G-Mean sebesar 0,875, berarti datanya cukup balance pada tipe responsenya.

4.5 *Naïve Bayes*

Untuk membuat model *Naïve Bayes* langkah pertama adalah melihat prior dari data, dan data yang akan digunakan adalah data Training dengan 90 % dari data 102 observasi yaitu 91 data. Dimana formula untuk menghitung peluang prior adalah :

$$P(0) = \frac{N(0)}{N} = \frac{28}{91} = 0,3077$$

$$P(1) = \frac{N(1)}{N} = \frac{63}{91} = 0,6923$$

Dan dapat dijadikan pada Tabel 4.26 untuk peluang prior.

Tabel 4.26 *Prior Probability* Naive Bayes

Y	
0	1
0,3077	0,6923

Dan kemudian menghitung peluang masing-masing atau peluang marjinal untuk setiap variabel X terhadap variabel Y, pada tahapan ini ada 2 tipe data yang dapat diketahui dengan formula yang berbeda yaitu data kategori dan data kontinu. Namun pada penelitian ini untuk semua variabel independen (bebas) berbentuk kategori, sehingga formula yang digunakan adalah persamaan 2.8. Kemudian Hasil dari peluang untuk masing-masing variabel dirangkum pada Tabel 4.27 sebagai berikut :

Tabel 4.27 Peluang Marjinal masing-masing Variabel X terhadap variabel Y

Y	X1			
	0	1		
0	0.7500	0.2500		
1	0.6190	0.3810		
Y	X2			
	0	1	2	3
0	0.1786	0.2857	0.3571	0.1786
1	0.2857	0.2381	0.4286	0.0476
Y	X3			
	0	1	2	3
0	0.2143	0.3214	0.2500	0.2143
1	0.0159	0.2540	0.3333	0.3968
Y	X4			
	0	1		
0	0.2143	0.7857		
1	0.5079	0.4921		
Y	X5			
	0	1		
0	0.6071	0.3929		
1	0.8730	0.1270		
Y	X6			
	0	1	2	3
0	0.5000	0.2143	0.2143	0.0714
1	0.4921	0.3175	0.1905	0.0000
Y	X7			
	0	1		
0	0.0000	0.2857		
1	0.0794	0.1429		

Lanjutan Tabel 4.27

Y	X8			
	0	1		
0	0.1071	0.8929		
1	0.0635	0.9365		
Y	X9			
	0	1	2	3
0	0.3571	0.0357	0.0357	0.5714
1	0.3016	0.2063	0.0794	0.4127
Y	X10			
	0	1		
0	0.3214	0.6786		
1	0.1905	0.8095		
Y	X11			
	0	1		
0	0.6071	0.3929		
1	0.3175	0.6825		

Setelah menghitung peluang prior dan peluang marjinal, kemudian didapatkan hasil prediksi yang dibentuk dari model peluang *Naïve Bayes* yang digunakan untuk membandingkan dengan peluang data testing 10 % dari 102 pengamatan yaitu 11 observasi, dan ditulis pada Tabel 4.28, sebagai berikut :

Tabel 4.28 Prediksi Y ($\hat{Y}$)

Y	$\hat{Y}$
1	1
1	0
0	1
1	1
0	0
0	1
0	0
1	1
1	1
1	1
1	1

Dari hasil prediksi pada Tabel 4.28 kemudian dibandingkan data yang sesungguhnya dan dibentuk tabel kontingensi *Naïve Bayes* pada Tabel 4.29, sebagai berikut :

Tabel 4.29 Tabel Kontingensi *Naïve Bayes*

Real	**Prediction**	
	0	**1**
0	2	1
1	2	6

Dari Tabel 4.29, akan dihitung ketepatan klasifikasi menggunakan persamaan 2.16, 2.17, 2.18, 2.19 dan persamaan 2.20, dirangkum pada Tabel 4.30, sebagai berikut :

Table 4.30 Ketepatan Klasifikasi *Naïve Bayes*

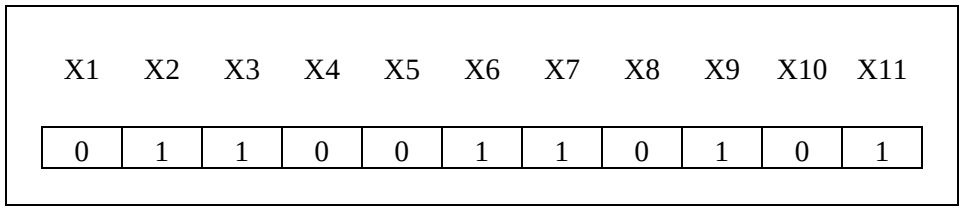
	Accuracy	APER	Specificity	Sensitivity	G-Mean
Naïve Bayes	72.73%	27.27%	0.86	0.50	0.68

Pada Tabel 4.30, terlihat bahwa dari keakurat model didapatkan akurasi sebesar 72.73 %, dengan kesalahannya 27.27 %. Kemudian juga terdapat Specificity nya atau *true positive* sebesar 0.86 dan sensitivity atau *true negative* sebesar 0.50. Untuk melihat apakah datanya mengalami *unbalance*, dapat dilihat dari nilai G-Mean sebesar 0.68, karena nilai G-mean > 0.5, berarti datanya cukup balance pada tipe responsnya.

4.6 Algoritma Genetika - *Naïve Bayes*

Pada seleksi Algoritma Genetika dengan *Naïve Bayes*, dimana langkah-langkah atau tahapan dalam melakukan seleksi ini hampir sama dengan sebelumnya pada tahapan seleksi Algoritma Genetika dengan Regresi Logistik. Namun pada seleksi ini fungsi *fitness*-nya menggunakan peluang *Naive Bayes*. Sehingga nilai *fitness*-nya berbeda dengan nilai *fitness* pada Algoritma Genetika dengan Regresi Logistik.

Tahapan pertama dari metode algoritma genetika adalah membentuk kromosom awal. Kromosom untuk seleksi variabel berisikan gen yang berupa variabel independen, dimana tipe kromosom yang sesuai untuk seleksi variabel adalah tipe biner. Variabel independen yang akan masuk dalam fungsi *Naïve Bayes* dikode 1, sedangkan yang tidak masuk dikode dengan 0. Panjang kromosom sesuai dengan jumlah variabel independen, dimana variabel independennya adalah data septictank di wilayah Surabaya Timur dengan panjang kromosomnya adalah 11 variabel independen. Pada Gambar 4.16 merupakan ilustrasi dari kromosom awal yang dibentuk sebagai berikut :



Gambar 4.16 Ilustrasi Kromosom awal

Pada penelitian ini telah ditentukan ukuran populasi sebanyak 100. Jadi ada sebanyak 100 kromosom yang akan dibangkitkan. Tabel 4.31 berikut ini menyajikkan ilustrasi 100 kromosom yang dibangkitkan untuk populasi awal.

Tabel 4.31 Ilustrasi Populasi Awal pada 100 Kromosom untuk Naive Bayes

Kromosom	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	0	1	1	0	0	1	1	0	1	0	1
2	1	0	0	1	0	1	1	0	1	0	1
.											
.											
.											
99	0	1	0	0	0	1	1	0	1	0	1
100	0	0	0	0	0	0	1	0	1	0	1

Pada Tahap kedua adalah mengevaluasi masing-masing kromosom yang telah dibentuk menggunakan fungsi fitnes yang telah ditentukan. Fungsi *fitness* yang digunakan untuk menyelesaikan permasalahan seleksi variabel pada Regresi Logistik adalah kesalahan klasifikasi pembentukan peluang Regresi Logistik. Hasil evaluasi masing-masing kromosom pada populasi awal dapat dilihat pada Tabel 4.32.

Tabel 4.32 Ilustrasi Nilai *Fitness* pada masing-masing Kromosom

Kromosom	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Nilai Fitnes
1	0	1	1	0	0	1	1	0	1	0	1	0,293
2	1	0	0	1	0	1	1	0	1	0	1	0,345
.												.
.												.
.												.
99	0	1	0	0	0	1	1	0	1	0	1	0,293
100	0	0	0	0	0	0	1	0	1	0	1	0,134

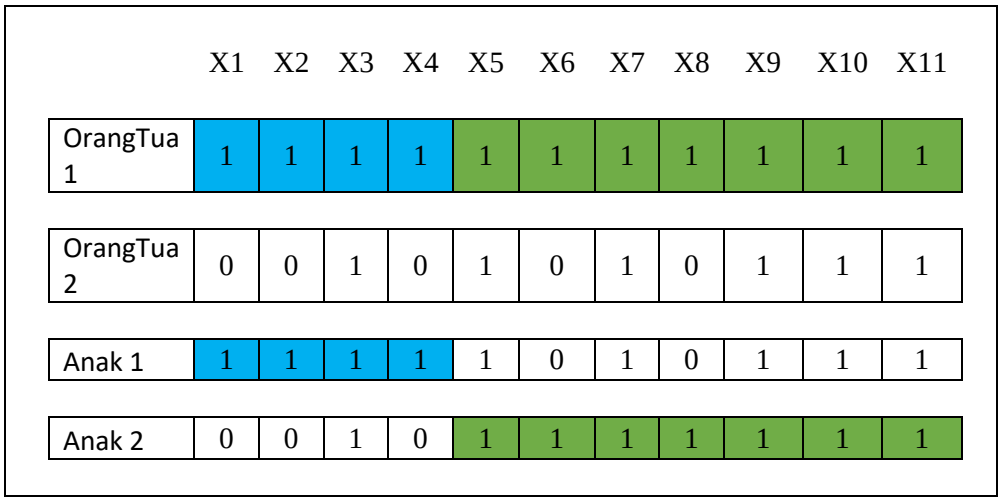
Kemudian pada tahapan ketiga adalah memilih kromosom dari populasi awal untuk dijadikan orang tua pada generasi berikutnya menggunakan metode RWS. Tahapan metode RWS secara umum adalah menentukan proporsi nilai fitnes, menentukan nilai komulatifnya, membangkitkan bilangan random yang beridistribusi uniform antar 0-1, kemudian membandingkan nilai komulatif dengan bilangan random tersebut. Apabila nilai komulatif kromosom ke-*i* lebih besar dari bilangan randomnya, maka kromosom ke-*i* tersebut terpilih menjadi orang tua untuk generasi berikutnya. Berikut adalah ilustrasi proses RWS pada Tabel 4.33.

Tabel 4.33 Ilustrasi Proses RWS

Kromosom	Nilai *Fitness*	Proporsi Nilai *fitness*	*fitness* Komulatif	Bilangan Random
1	0.293	0.012	0.012	0.625
2	0.345	0.015	0.027	0.720
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
99	0.293	0.012	0.994	0.373
100	0.134	0.006	1.000	0.280

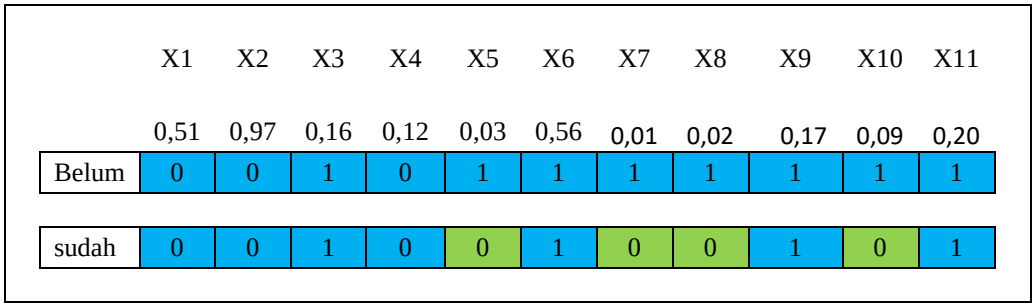
Berdasarkan Tabel 4.33 dapat diketahui nilai bilangan *random* pertama adalah 0,625 dan untuk nilai *fitness* kumulatif pertama sebesar 0,012 sehingga kromosom pertama berubah menjadi individu baru. Dan begitu seterusnya untuk kromosom selanjutnya.

Proses Selanjutnya Tahap keempat adalah melakukan pindah silang, yang tujuannya untuk mengasilkan kromosom baru dari perpaduan 2 kromosom orang tua. Pindah silang dalam penelitian ini menggunakan pindah silang satu titik atau *crossover* sederhana. Penetuan titik ini melibatkan probabilitas pindah silang (Ps) = 0,8. Proses penentuan titik potong dilanjutkan dengan membangkitkan bilangan *random* antara 0 sampai 1 kemudian membadingkan nilai bilangan *random* dengan nilai probabilitas pindah silang (Ps) = 0,8, jika bilangan *random* lebih kecil dari probabilitas pindah silang maka orangtua yang terpilih dari individu satu dan individu dua akan dikawinkan dan keturunanya akan menjadi individu baru. Sehingga misalkan titik potong ditentukan secara *random* serta ditemukan pada X_4, maka mulai X_5 akan ditukarsilangkan antara variabel pada kromosom bapak dan ibu. Berikut ditunjukkan Ilustrasi proses pindah silang pada Gambar 4.17.



Gambar 4.17 Ilustrasi Proses Pindah Silang

Kemudian pada tahapan kelima adalah mutasi, metode yang digunakan adalah metode uniform. Tujuan mutasi adalah untuk mendapatkan keberagaman gen atau variabel pada kasus ini. Tahapan mutasi adalah dengan membangkitkan bilangan random pada setiap variabel dan membandingkan dengan peluang mutasi sebesar 0.1. Apabila nilai bangkitan pada suatu variabel kurang dari peluang mutasi, maka dilakukan proses mutasi pada gen tersebut. Proses mutasi yaitu dengan mengubah kode 1 menjadi 0 atau 0 menjadi 1 pada variabel yang dilakukan proses mutasi. Pada Gambar 4.18 akan ditunjukkan ilustrasi tahapan dalam melakukan mutasi.



Gambar 4.18 Ilustrasi tahapan Mutasi

Pada Gambar 4.18 terlihat bahwa X_5, X_7, X_8, dan X_{10} mengalami mutasi karena probabilitis bilangan randomnya lebih keci dari probabilitis mutasi (*Pm*) = 0,1.

Tahapan keenam adalah melakukan elitisme, yaitu menggandakan kromosom yang memiliki nilai fitnes paling kecil. Jumlah kromosom yang digandakan adalah 2. Berdasarkan perhitungan nilai fitnes tiap kromosom pada Tabel 4.34 diketahui bahwa nilai fitnes paling kecil adalah 0,134.

Tabel 4.34 Kromosom dan Nilai Fitness pada seleksi Algoritma Genetika pada *Naïve Bayes*

Kromosom	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Nilai *Fitness*
1	0	0	0	0	0	1	1	0	1	0	1	0.134
2	0	0	0	0	0	1	1	0	1	0	1	0.134
.						.						.
.						.						.
.						.						.
98	0	0	0	0	0	1	1	0	1	1	1	1.000
99	0	0	0	0	0	1	1	0	1	0	1	0.134
100	0	0	0	0	0	0	1	0	1	0	1	0.134

Dari Tabel 4.34 diberikan 100 kromosom dengan nilai *fitness*-nya dan hasil seleksi Algoritma Genetika di atas didaptakan 5 variabel yang terpilih yaitu X_2, X_7, X_8, X_9, X_{10} dan X_{11} yang terdapat pada Kromosom pertama dengan nilai kesalahan 0,134. Kemudian akan dibentuk tabel klasifikasinya pada Tabel 4.35, sebagai berikut :

Tabel 4.35 Klasifikasi Algoritma Genetika- *Naïve Bayes*

Real	**Prediction**	
	0	**1**
0	3	0
1	1	7

Dari Tabel 4.35, kita dapat mengetahui ketepatan klasifikasinya untuk melihat keakuratan dari variabel yang terpilih atau faktor yang terpilih itu cukup mewakili dalam menganalisa faktor yang mempengaruhi pembuangan limbah rumah tangga yang memiliki septiktank atau tidak. Berikut Tabel 4.36 akan merangkum hasil ketepatan klasifikasinya.

Tabel 4.36 Ketepatan Klasifikasi GA – *Naïve Bayes*

	Accuracy	APER	Specificity	Sensitivity	G-Mean
Algoritma Genetika-Naive Bayes	90,91%	9,09%	1	0,75	0,875

Pada Tabel 4.36, terlihat bahwa dari keakurat model didapatkan akurasi sebesar 90,91%, dengan kesalahannya 9,09 %. Kemudian juga terdapat Specificity nya atau *true positive* sebesar 1 dan sensitivity atau *true negative* sebesar 0,75. Untuk melihat apakah datanya mengalami *unbalance*, dapat dilihat dari nilai G-Mean sebesar 0,875, berarti datanya cukup balance pada tipe responsenya.

4.7 Regresi Logistik – *Naïve Bayes*

Dari hasil Tabel 4.17 didapatkan variabel yang terpilih dari model Regresi Logistik pada seleksi backward yaitu X_1, X_5, X_7, X_9, dan X_{11}. Kemudian melihat prior untuk peluang yang dihasilkan dari peluang Y, sebagai berikut :

$$P(0) = \frac{N(0)}{N} = \frac{28}{91} = 0,3077$$

$$P(1) = \frac{N(1)}{N} = \frac{63}{91} = 0,6923$$

Ternyata hasilnya sama saja dengan peluang prior pada perhitungan untuk Tabel 4.25, karena untuk data yang sama dengan variabel independen yang terpilih X_1, X_5, X_7, X_9, dan X_{11} tidak berpengaruh terhadap peluang priornya.

Setelah itu melihat peluang marjinal untuk setiap variabel independen yang terpilih dengan menggunakan persamaan pada 2.8, didapatkan peluangnya pada Tabel 4.37, sebagai berikut :

Tabel 4.37 Kumpulan Peluang Marjinal masing-masing Variabel X yang terpilih terhadap variabel Y

Y	X1	
	0	1
0	0,7500	0,2500
1	0,6190	0,3810

Y	X5	
	0	1
0	0,6071	0,3929
1	0,8730	0,1270

Y	X6			
	0	1	2	3
0	0,5000	0,2143	0,2143	0,0714
1	0,4921	0,3175	0,1905	0,0000

Y	X7	
	0	1
0	0,0000	0,2857
1	0,0794	0,1429

Y	X9			
	0	1	2	3
0	0,3571	0,0357	0,0357	0,5714
1	0,3016	0,2063	0,0794	0,4127

Lanjutan Tabel 4.37

Y	X11	
	0	1
0	0,6071	0,3929
1	0,3175	0,6825

Dan pada Tabel 4.37, terlihat bahwa peluang marjinal X_1, X_5, X_7, X_9, dan X_{11} juga sama dengan peluang marjinal pada Tabel 4.26, karena untuk data yang sama, meskipun pada variabel yang digunakan tidak semua atau sebagian, tetapi memiliki peluang marjinal yang sama pada setiap variabelnya. Sehingga dapat dihitung untuk nilai prediksi dengan variabel independen yang terpilih X_1, X_5, X_7, X_9, dan X_{11}, dapat dilihat pada Tabel 4.38, sebagai berikut :

Tabel 4.38 Prediksi Y ($\hat{Y}$) dari hasil seleksi Regresi Logistik

Y	Y*
1	1
1	1
0	0
1	1
0	0
0	1
0	0
1	1
1	1
1	1
1	1

Kemudian dihitung keakuratan dari variabel yang terpilih dari Regresi Logistik di modelkan dengan *Naïve Bayes*, pada Tabel 4.39 dapat dilihat tabel klasifikasinya, sebagai berikut :

Tabel 4.39 Klasifikasi Regresi Logistik – *Naïve Bayes*

Real	**Prediction**	
	0	**1**
0	2	2
1	2	5

Dari Tabel 4.39 bahwa klasifikasi antara Regresi Logistik yang digabungkan dengan Naive Bayes terlihat cukup bagus, dimana regresi Logistik merupakan sebagai seleksi variabel kemudian dimodelkan dengan Naive Bayes untuk melihat keakuratan dari model penggabungan tersebut, untuk lebih jelasnya didapatkan

Tabel 4.40 untuk melihat besarnya ketepatan klasifikasi dari Regresi Logistik dengan *Naïve Bayes*, sebagai berikut :

Tabel 4.40 Ketepatan Klasifikasi Regresi Logistik-Naive Bayes

	Accuracy	APER	Specificity	Sensitivity	G-Mean
Regresi Logistik-*Naïve Bayes*	63,64%	36,36%	0,71	0,50	0,61

Pada Tabel 4.40, terlihat bahwa dari keakurat model didapatkan akurasi sebesar 63,64 %, dengan kesalahannya 36,36 %. Kemudian juga terdapat Specificity nya atau *true positive* sebesar 0,71 dan sensitivity atau *true negative* sebesar 0,50. Untuk melihat apakah datanya mengalami *unbalance*, dapat dilihat dari nilai G-Mean sebesar 0,61, berarti datanya cukup balance pada tipe responsenya.

4.8 Perbandingan Ketepatan Klasifikasi

Dari hasil yang didapatkan, maka perbandingan dari keseluruhan metode yang digunakan, diantaranya Regresi Logistik, *Naïve Bayes*, Algoritma Genetika – Regresi Logistik, Algoritma Genetika – *Naïve Bayes*, dan Regresi Logistik – *Naïve Bayes* yang dirangkum pada Tabel 4.41, sebagai berikut :

Tabel 4.41 Perbandingan Ketepatan Klasifikasi

Metode	Variabel Terpilih	Ukuran Ketepatan Klasifikasi				
		Accuracy	APER	Specificity	Sensitivity	G-Mean
Regresi Logistik	X1, X2, X3, ..., X11	54,55%	44,45%	0,57	0,50	0,54
Naïve Bayes	X1, X2, X3, ..., X11	72,73%	27,27%	0,86	0,50	0,68
Algoritma Genetika - Regresi Logistik	**X2, X7, X8, X9, X10, X11**	**90,91%**	**9,09%**	**1**	**0,75**	**0,875**
Algoritma Genetika - Naive Bayes	**X6, X7, X9, X11**	**90,91%**	**9,09%**	**1**	**0,75**	**0,875**
Regresi Logistik-seleksi *backward*	X1, X5, X7, X9, X11	72,73%	27,27%	0,86	0,50	0,68
Regresi Logistik - Naive Bayes	X1, X5, X7, X9, X11	63,64%	36,36%	0,71	0,50	0,61

(*halaman ini sengaja dikosongkan*)

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan analisis dari hasil dan pembahasan, maka akan ditarik kesimpulan, sebagai berikut :

1. Keakuratan klasifikasi antara Regresi Logistik dan Naïve Bayes untuk semua variabel independen yang digunakan adalah Naïve Bayes lebih baik daripada Regresi Logistik pada kasus ini. Dimana Naïve Bayes memiliki akurasi sebesar 72,73 %, Specificity sebesar 0,86, Sensitivity sebesar 0,50, dan G-Mean sebesar 0,68. Sedangkan Regresi Logistik memiliki akurasi sebesar 54,55 %, Specificity sebesar 0,57, Sensitivity sebesar 0,50, dan G-Mean sebesar 0,68.
2. Pada Kajian Algoritma Genetika ini bertujuan untuk melihat seleksi variabel yang digunakan pada setiap Regresi Logistik dan Naïve Bayes Dan hasil seleksi Algoritma Genetika untuk Regresi Logistik adalah X_2, X_7, X_8, X_9, X_{10}, dan X_11 dengan nilai kesalahannya adalah 0,134 dan untuk Naïve Bayes variabel yang terpilih adalah X_6, X_7, X_9, dan X_{11} dengan nilai kesalahannya yaitu 0,134.
3. Hasil perbandingan ketepatan klasifikasi antara Algoritma Genetika-Regresi Logistik dan Algoritma Genetika – Naïve Bayes dapat dilihat bahwa ukuran ketepatan klasifikasi antara Algoritma Genetika – Regresi Logistik dengan Algoritma genetika – Naïve Bayes adalah sama. Namun hasil seleksi variabel untuk Regresi Logistik lebih banyak yakni 6 variabel sedangkan Naïve Bayes lebih sedikit yakni 4 variabel.

5.2 Saran

Berdasarkan kesimpulan diatas, maka didapatkan saran sebagai berikut :

1. Dapat membandingkan metode klasifikasi lainnya, seperti SVM, Regesi *three*, KNN, *decision three*, dan lain sebagainya.
2. Diharapkan untuk menggunakan metode optimasi yang lain seperti PSO, ACO, dll. Kemudian dapat membandingkan beberapa optimasi tersebut.

3. Diharapkan menggunakan data yang lebih banyak, untuk lebih mengetahui tingkat keakuratan yang lebih baik dari metode klasifikasi yang dibandingkan.

DAFTAR PUSTAKA

Back B., Laitinen T., Sere K., dan Wezel M. V, "Choosing Bankruptcy Predictors Using Discriminan Analysis, Logit Analysis, and Genetic Algorithms", Turku Centre for Computer Science, (1996)

Bustami., (2013), "Penerapan Algoritma Naïve Bayes Untuk Mengklasifikasi Data Nasabah Asuransi", *TECHSI : Jurnal PenelitianTeknik Informatika*, Vol. 3, No.2, Hal. 127-146.

Chiang, L. H., dan Pell, R. J. (2004). Genetic Algorithms Combined with Discriminant Analysis for Key Variable Identification. Journal of Process Control, 14, 143-155.

Desiani, A., dan Arhami, M., 2006. Konsep Kecerdasan Buatan. Yogyakarta: Andi Offset

Guo, P., Wang, X. dan Han, Y., 2010. The Enhanced Genetic Algorithms for the Optimization Design. , (Bmei), pp.2990–2994.

Han J. dan Kamber M., (2006), *Data mining : Concepts and Techniques*, Second Edition, Morgan Kaufmann, California.

Haupt, S. E. dan Haupt, R. L. 2004. Practical Genetic Algorithms. New Jersey: A John Wiley & Sons Inc.

Hosmer, D. W. dan Lemeshow, S., (1989), *Applied Logistic Regression*, John Wiley and Sons Inc, Canada.

Jadaan, O. A., Rajamani, L., dan Rao, C. R. (2005-2008). Improved selection Operator for GA. *Journal of Theoretical and Applied Information Technology (JATIT)*, 269-277.

Johnson, D. W., dan Dean W. W., (2007), *Applied Multivariat Statistical Analysis*, Sixth Edition, Prentice Hall Iternational Inc, New Jersey.

Kubat, M., Matwin, S., dan Holte, R., (1997), Addressing The Curse of Inbalanced Traing Set: One Side Selection, 14 th International Conference on Machine Learning Nashville, TN, USA, pp. 179-186.

Kusumawardani, K., (2015), *Seleksi Variabel dan Optimasi Parameter Menggunakan Hybrid Analisis Diskriminan dan Algoritma Genetika untuk Klasifikasi*, Surabaya: Institut Teknologi Sepuluh Nopember.

Kusumawati Y., (2013), *Pemodelan Faktor-Faktor Yang Mempengaruhi Rumah Tangga Membuang Limbah Domestik Menggunakan Regresi Logistik Dan Algoritma Genetika*, Surabaya : Institut Teknologi Sepuluh Nopember.

Patil, T. R., dan Shereker, M. S., (2013), "Performance Analysis of Naïve Bayes and j48 Classification Algorithm for Data Classification", *Internasional Journal of Computer Science and Applications*, Vol. 6, No.2, Hal 256-261.

Patterkari, S. A., dan Parveen, A., (2012), "Prediction System for Heart Disease Using Naïve Bayes", *Internastional Journal of Advance Computer and Mathematical Science*, Vol. 3, No. 3, Hal 290-294.

Prasetyo, E., (2006), *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*, ANDI Yogyakarta, Yogyakarta.

Ridwan, M., Suyono, H., dan Sarosa, M., (2013), "Penerapan Data Mining untk Evaluasi Kinerja Akademik Mahsiswa Menggunakan Algoritma Naïve Bayes Classifier", *Jurnal EECCIS*, Vol 1, No. 7, Hal. 59-64.

Park, T. S., Lee, J. H., dan Choi, B., (2009) "Optimization for Artificial Neural Networ with adaptive inertial weight of Particle Swarm Optimization," *8th IEEE International Conference on Cognitive Informatics*, Hal. 481-485.

Santosa, B. dan Willy, P., (2011), *Metode Metaheuristik: Konsep dan Implementasi*. Guna Widya. Surabaya.

Sivanandam, S.N., dan Deepa, S.N. 2008. Introduction to Genetic Algorithms. Berlin Heidelberg New York: Springer.

Sumathi, S dan Surekha, P., (2010), Computational Intelligence Paradigms Theory and Applications Using Matlab, *Taylor and Francis Group*, ISBN 978-1-4398-0902-0.

Suyanto. (2005). Algoritma Genetika dalam Matlab. Yogyakarta: Andi Offset.

Trevino,V. dan Falciani,F., (2006), GALGO: an R package for multivariate variable selection using genetic algorithms. Bioinformatics, 22, 1154–1156.

Wati, R., (2016), Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan Menggunakan Naive Bayes, *Jurnal Evolusi*, Volume 4, Nomor 1.

Xu, L., dan Zhang, W. J. (2001). Comparison of Different methods for Variable Selection. Analytica Chimica Acta 446, 477-483.

LAMPIRAN

Lampiran 1. Data coding dari penelitian Kusumawati (2013)

No.	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	1	1	2	3	0	0	1	0	0	0	1	1
2	1	0	1	3	0	0	0	2	1	0	1	1
3	1	1	0	3	0	0	0	1	1	0	1	1
4	1	1	0	2	1	1	0	1	1	0	1	1
. . .						. . .						. . .
99	1	0	0	1	0	0	2	2	1	1	1	0
100	1	0	2	1	0	0	2	2	1	0	1	1
101	1	0	1	3	0	0	2	2	1	2	1	1
102	1	0	2	1	0	0	2	2	1	0	1	1

(Kusumawati, 2013)

Keterangan :
Y = Kepemilikan septictank
X1 = Jumlah keluarga
X2 = Pendidikan kepala keluarga
X3 = Pekerjaan kepala keluarga
X4 = Pendapatan kepala keluarga
X5 = Pengeluaran kepala keluarga
X6 = Status kepemilikan rumah
X7 = Waktu pengurasan tangki septik
X8 = Jenis kloset
X9 = Penyakit yang pernah dialami anggota keluarga
X10 = Air yang digunakan untuk mandi dan cuci
X11 = Perilaku sanitasi

Lampiran 2. Data yang Digunakan (data yang sudah acak satu kali)

No.	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	1	1	2	3	0	0	1	0	0	0	1	1
2	1	0	1	3	0	0	0	2	1	0	1	1
3	1	1	0	3	0	0	0	1	1	0	1	1
. . .						. . .						. . .
99	1	0	1	3	1	0	0	2	0	1	0	1
100	1	1	1	3	1	0	0	1	0	1	1	0
101	1	0	1	3	0	0	0	2	1	2	1	1
102	1	0	0	3	0	0	0	2	1	2	0	1

Keterangan :
Y = Kepemilikan septictank
X1 = Jumlah keluarga
X2 = Pendidikan kepala keluarga
X3 = Pekerjaan kepala keluarga
X4 = Pendapatan kepala keluarga
X5 = Pengeluaran kepala keluarga
X6 = Status kepemilikan rumah
X7 = Waktu pengurasan tangki septik
X8 = Jenis kloset
X9 = Penyakit yang pernah dialami anggota keluarga
X10 = Air yang digunakan untuk mandi dan cuci
X11 = Perilaku sanitasi

Lampiran 3. Hasil Prediksi dan Aktual

Reglog	
Aktual	Prediksi
1	0.2534
1	0.0328
0	0.9807
1	0.3671
0	0.0407
0	0.9853
0	0.0104
1	0.8773
1	0.9666
1	0.9892
1	0.9513

Reglog-seleksi backward	
Aktual	Prediksi
1	0.6125
1	0.0927
0	0.8069
1	0.6909
0	0.2057
0	0.6909
0	0.0250
1	0.9943
1	0.9892
1	0.9645
1	0.9645

Reglog-GA	
Aktual	Prediksi
1	0.7707
1	0.6621
0	0.4294
1	0.6881
0	0.2783
0	0.8319
0	0.3087
1	0.9476
1	0.6169
1	0.9076
1	0.8795

Lampiran 4. Hasil Output tahapan *Backward*

Start: 0 AIC = 98.870

Y = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11

Variabel	Df	Deviance	AIC
X10	1	57.650	97.650
X8	1	57.650	97.650
X4	1	57.838	97.838
X2	3	62.607	98.607
X6	3	62.793	98.793
<none>		56.868	98.868
X3	3	65.130	101.130
X11	1	62.365	102.365
X7	2	64.827	102.827
X9	3	67.539	103.539
X5	1	66.673	106.673
X1	1	72.801	112.801

Step: 1 AIC = 97.650

Y = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X11

Variabel	Df	Deviance	AIC
X8	1	58.487	96.487
X2	3	62.901	96.901
X4	1	59.145	97.145
X6	3	63.287	97.287
<none>		57.650	97.650
X3	3	66.158	100.158
X11	1	62.676	100.676
X7	2	66.103	102.103
X9	3	68.975	102.975
X5	1	67.561	105.561
X1	1	73.642	111.642

Step: 2 AIC = 96.490

Y = X1 + X2 + X3 + X4 + X5 + X6 + X7 + X9 + X11

variabel	Df	Deviance	AIC
X2	3	62.976	94.976
X6	3	63.290	95.290
X4	1	60.077	96.077

<none>		58.487	96.487
X11	1	63.024	99.024
X3	3	67.604	99.604
X7	2	66.356	100.356
X9	3	71.557	103.557
X5	1	67.833	103.833
X1	1	74.108	110.108

Step : 3 AIC = 94.98

Y = X1 + X3 + X4 + X5 + X6 + X7 + X9 + X11

Variabel	Df	Deviance	AIC
X6	3	67.143	93.143
X4	1	63.799	93.799
<none>		62.976	94.976
X3	3	69.683	95.683
X11	1	65.894	95.894
X7	2	70.132	98.132
X5	1	69.546	99.546
X9	3	74.231	100.231
X1	1	75.713	105.713

Step : 4 AIC = 93.14

Y = X1+ X3 + X4 + X5 + X7 + X9 + X11

Variabel	Df	Deviance	AIC
X4	1	67.850	91.850
X3	3	72.756	92.756
<none>		67.143	93.143
X11	1	70.120	94.120
X7	2	74.387	96.387
X5	1	73.893	97.893
X9	3	81.577	101.577
X1	1	78.025	102.025

Step : 5 AIC = 91.85

Y = X1 + X3 + X5 + X7 + X9 + X11

Variabel	Df	Deviance	AIC
X3	3	73.630	91.630
<none>		67.850	91.850
X11	1	73.329	95.329

X7	2	75.519	95.519
X5	1	76.264	98.264
X1	1	78.031	100.031
X9	3	82.975	100.975

Step: 6 AIC=91.63

Y = X3 + X5 + X7 + X9 + X11

Variabel	Df	Deviance	AIC
<none>		73.630	91.630
X11	1	80.476	96.476
X7	2	83.644	97.644
X1	1	82.793	98.793
X5	1	84.346	100.346
X9	3	90.063	102.063

Lampiran 5. Estimasi Parameter Seleksi *Bakcward*

Tahap	Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
1	(Intercept)	12.489	2104.057	0.006	0.995
	X1 (1)	4.462	1.473	3.029	0.002
	X2 (1)	0.738	0.974	0.758	0.449
	X2 (2)	2.393	1.304	1.836	0.066
	X2 (3)	2.958	1.614	1.833	0.067
	X3 (1)	3.633	2.038	1.782	0.075
	X3 (2)	3.571	1.883	1.897	0.058
	X3 (3)	4.944	2.082	2.374	0.018
	X4 (1)	-1.089	0.902	-1.207	0.227
	X5 (1)	-3.784	1.415	-2.674	0.007
	X6 (1)	1.926	1.004	1.918	0.055
	X6 (2)	2.251	1.343	1.676	0.094
	X6 (3)	-14.638	4612.203	-0.003	0.997
	X7 (1)	-20.396	2104.056	-0.010	0.992
	X7 (2)	-18.674	2104.056	-0.009	0.993
	X8 (1)	1.432	1.612	0.888	0.374
	X9 (1)	2.172	1.645	1.320	0.187
	X9 (2)	0.958	1.969	0.486	0.627
	X9 (3)	-1.803	1.034	-1.744	0.081
	X11 (1)	1.996	0.941	2.120	0.034

Tahap	Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
2	(Intercept)	13.763	2015.161	0.007	0.995
	X1 (1)	4.301	1.424	3.020	0.003
	X2 (1)	0.679	0.957	0.709	0.478
	X2 (2)	2.030	1.190	1.705	0.088
	X2 (3)	2.679	1.563	1.713	0.087
	X3 (1)	4.041	2.045	1.976	0.048
	X3 (2)	3.792	1.930	1.964	0.050
	X3 (3)	5.160	2.120	2.434	0.015
	X4 (1)	-1.126	0.903	-1.246	0.213
	X5 (1)	-3.585	1.372	-2.613	0.009
	X6 (1)	1.748	0.955	1.831	0.067
	X6 (2)	1.574	1.061	1.483	0.138
	X6 (3)	-14.502	4612.203	-0.003	0.997
	X7 (1)	-20.096	2015.160	-0.010	0.992
	X7 (2)	-18.742	2015.160	-0.009	0.993

	X9 (1)	2.641	1.606	1.644	0.100
	X9 (2)	1.710	1.817	0.941	0.347
	X9 (3)	-1.603	0.998	-1.606	0.108
	X11 (1)	1.849	0.913	2.024	0.043

Tahap	Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
3	(Intercept)	14.992	2111.835	0.007	0.994
	X1 (1)	3.263	1.125	2.902	0.004
	X3 (1)	3.854	1.949	1.978	0.048
	X3 (2)	3.384	1.847	1.833	0.067
	X3 (3)	3.891	1.872	2.078	0.038
	X4 (1)	-0.749	0.828	-0.905	0.366
	X5 (1)	-2.527	1.081	-2.339	0.019
	X6 (1)	1.590	0.900	1.766	0.077
	X6 (2)	1.222	0.955	1.280	0.201
	X6 (3)	-14.029	4612.203	-0.003	0.998
	X7 (1)	-19.516	2111.834	-0.009	0.993
	X7 (2)	-18.214	2111.834	-0.009	0.993
	X9 (1)	2.243	1.409	1.592	0.111
	X9 (2)	1.282	1.767	0.725	0.468
	X9 (3)	-1.387	0.907	-1.530	0.126
	X11 (1)	1.394	0.847	1.646	0.100

Tahap	Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
4	(Intercept)	16.758	2140.333	0.008	0.994
	X1 (1)	2.804	1.030	2.721	0.007
	X3 (1)	3.048	1.804	1.689	0.091
	X3 (2)	2.826	1.740	1.624	0.104
	X3 (3)	3.473	1.813	1.915	0.056
	X4 (1)	-0.673	0.801	-0.841	0.401
	X5 (1)	-2.478	1.045	-2.372	0.018
	X7 (1)	-19.719	2140.332	-0.009	0.993
	X7 (2)	-18.704	2140.332	-0.009	0.993
	X9 (1)	2.546	1.397	1.822	0.068
	X9 (2)	1.653	1.749	0.945	0.345
	X9 (3)	-1.449	0.882	-1.643	0.100
	X11 (1)	1.319	0.785	1.681	0.093

Tahap	Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
5	(Intercept)	16.369	2165.180	0.008	0.994
	X1 (1)	2.641	1.004	2.632	0.009
	X3 (1)	2.920	1.733	1.685	0.092
	X3 (2)	2.821	1.680	1.680	0.093
	X3 (3)	3.430	1.746	1.964	0.050
	X5 (1)	-2.657	1.019	-2.608	0.009
	X7 (1)	-19.709	2165.179	-0.009	0.993
	X7 (2)	-18.676	2165.179	-0.009	0.993
	X9 (1)	2.420	1.382	1.751	0.080
	X9 (2)	1.593	1.743	0.914	0.361
	X9 (3)	-1.566	0.866	-1.808	0.071
	X11 (1)	1.604	0.717	2.237	0.025

Tahap	Coefficient	Estimate	Std. Error	z-value	Pr(>|z|)
6	(Intercept)	18.4067	1337.949	0.014	0.989
	X1 (1)	2.3141	0.909	2.545	0.011
	X5 (1)	-2.7810	0.970	-2.866	0.004
	X7 (1)	-19.1703	1337.949	-0.014	0.989
	X7 (2)	-17.9067	1337.949	-0.013	0.989
	X9 (1)	2.9672	1.381	2.149	0.032
	X9 (2)	1.1130	1.764	0.631	0.528
	X9 (3)	-1.3839	0.801	-1.727	0.084
	X11 (1)	1.6884	0.678	2.489	0.013

Lampiran 6. Hasil seleksi Algoritma Genetika – Regresi Logistik

Kromosom												Nilai Fitness
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	
1	0	1	0	0	0	0	1	1	1	1	1	0.134
2	0	0	1	0	1	0	1	0	1	1	1	0.134
3	0	0	1	0	1	0	1	0	1	1	1	0.134
4	0	0	0	0	0	1	1	1	1	1	1	0.134
5	0	1	1	0	1	0	0	0	1	1	1	0.134
6	0	1	0	0	1	0	0	1	1	1	1	0.293
7	0	0	1	0	1	0	1	1	1	1	1	0.134
8	0	0	1	1	1	0	0	1	1	1	1	0.134
9	0	0	1	0	1	0	1	0	1	1	1	0.134
.						.						.
.						.						.
.						.						.
92	1	1	1	0	1	1	1	0	1	1	1	0.345
93	0	1	1	0	0	0	0	1	1	1	0	1.000
94	0	1	0	1	0	0	0	1	1	1	1	0.293
95	1	1	1	0	1	0	1	1	1	1	1	0.402
96	0	0	1	0	1	1	1	0	1	0	1	0.198
97	0	0	1	0	1	0	1	0	1	0	1	0.134
98	0	0	1	1	1	0	1	0	0	1	1	0.537
99	0	0	1	0	1	0	1	1	1	1	1	0.134
100	0	0	1	1	1	0	1	1	1	1	1	0.134

Lampiran 7. Hasil Seleksi Algoritma Genetika – Naive Bayes

Kromosom												Nilai Fitness
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	
1	0	0	0	0	0	1	1	0	1	0	1	0.134
2	0	0	0	0	0	1	1	0	1	0	1	0.134
3	0	0	0	0	0	1	1	0	1	0	1	0.134
4	0	0	0	0	0	1	1	0	1	0	1	0.134
5	0	0	0	0	0	1	1	0	1	0	1	0.134
6	0	0	0	0	0	1	1	0	1	0	1	0.134
7	0	0	0	0	0	1	1	0	1	0	1	0.134
8	0	0	0	0	0	1	1	0	1	0	1	0.134
9	0	0	0	0	0	1	1	0	1	0	1	0.134
.						.						.
.						.						.
.						.						.
92	0	0	0	0	0	1	1	0	1	0	1	0.134
93	0	0	0	1	0	1	1	0	1	0	1	0.345
94	0	0	1	0	0	0	1	0	1	1	1	0.500
95	1	0	1	0	0	0	1	0	1	0	1	0.293
96	0	1	0	0	0	1	1	0	1	0	1	0.293
97	0	0	0	0	1	1	1	0	1	0	1	0.402
98	0	0	0	0	0	1	1	0	1	1	1	1.000
99	0	0	0	0	0	1	1	0	1	0	1	0.134
100	0	0	0	0	0	0	1	0	1	0	1	0.134

Lampiran 8. Syntax R Model Regresi Logistik Biner

```
dataku=read.csv('data1.csv', header=T, sep=';')
dataku
dataku$X1<-factor(dataku$X1)
dataku$X2<-factor(dataku$X2)
dataku$X3<-factor(dataku$X3)
dataku$X4<-factor(dataku$X4)
dataku$X5<-factor(dataku$X5)
dataku$X6<-factor(dataku$X6)
dataku$X8<-factor(dataku$X8)
dataku$X7<-factor(dataku$X7)
dataku$X9<-factor(dataku$X9)
dataku$X10<-factor(dataku$X10)
dataku$X11<-factor(dataku$X11)
dataTrain<-dataku[1:91,]
dataTest<-dataku[92:102,]
model<-glm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11,
           data=dataTrain, family=binomial('logit'))
model
summary(model)
backward <- step(model, direction = 'backward')
summary(backward)
backward1<-glm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X11, data=dataTrain,
           family=binomial('logit'))
summary(backward1)
backward2<-glm(Y~X1+X2+X3+X4+X5+X6+X7+X9+X11, data=dataTrain,
           family=binomial('logit'))
summary(backward2)
backward3<-glm(Y~X1+X3+X4+X5+X6+X7+X9+X11, data=dataTrain,
           family=binomial('logit'))
summary(backward3)
backward4<-glm(Y~X1+X3+X4+X5+X7+X9+X11, data=dataTrain,
           family=binomial('logit'))
summary(backward4)
backward5<-glm(Y~X1+X3+X5+X7+X9+X11, data=dataTrain,
           family=binomial('logit'))
summary(backward5)
backward6<-glm(Y~X1+X5+X7+X9+X11, data=dataTrain,
           family=binomial('logit'))
summary(backward6)
predicted <- predict(model, dataTest, type="response")
predicted
library(InformationValue)
library(caret)
optCutOff <- optimalCutoff(dataTest$Y, predicted)[1]
optCutOff
```

```
p.survive=round(predicted)
tabel=confusionMatrix(p.survive1, dataTest$Y)
tabel
predicted1 <- predict(backward, dataTest, type="response")
predicted1
library(InformationValue)
optCutOff1 <- optimalCutoff(dataTest$Y, predicted1)[1]
optCutOff1
p.survive1=round(predicted1)
tabel1=confusionMatrix(p.survive1, dataTest$Y)
tabel1
```

Lampiran 9. Syntax R Model Naive Bayes

```
library(e1071)
library(caret)
dataku=read.csv('data1.csv', header=T, sep=';')
dataku
dataku$Y<-factor(dataku$Y)
dataku$X1<-factor(dataku$X1)
dataku$X2<-factor(dataku$X2)
dataku$X3<-factor(dataku$X3)
dataku$X4<-factor(dataku$X4)
dataku$X5<-factor(dataku$X5)
dataku$X6<-factor(dataku$X6)
dataku$X7<-factor(dataku$X7)
dataku$X8<-factor(dataku$X8)
dataku$X9<-factor(dataku$X9)
dataku$X10<-factor(dataku$X10)
dataku$X11<-factor(dataku$X11)
dataTrain<-dataku[1:91,]
dataTest<-dataku[92:102,]
nb<-naiveBayes(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11,
               data=dataTrain)
nb
predicted2<-predict(nb, dataTest, type=c("class"))
predicted2
confusionMatrix(table(predicted2, dataTest$Y))
```

Lampiran 10. Code Matlab Algoritma Genetika

```
%================reglog====================
ga
options = gaoptimset ('PopulationSize', 100, 'PopulationType',
'bitstring','FitnessScalingFcn',@fitscalingrank,'InitialPopulation
',
[],'EliteCount',[2],'MutationFcn',{@mutationuniform,0.1},'Crossove
rFraction',[0.8],'CrossoverFcn',@crossoversinglepoint,
'SelectionFcn', @selectionroulette)
[a,fval,exitflag,output,population,scores]=ga(@seleksi_reglog,11,o
ptions)
```

```
%Fungsi Regresi Logistik Biner
function [salah]=reglog_biner(ytrain,xtrain,ytest,xtest)
[b dev stats]=glmfit(xtrain,ytrain,'binomial');
n1=length(xtest(:,1));
p1=length(xtest(1,:));

for i=1:n1
    a1=0;
    for j=2:(p1+1)
        a1=a1+(xtest(i,(j-1))*b(j));
    end
    pi1(i)=a1+b(1);
end
for i=1:n1
    if pi1(i)>50
        pi1(i)=50;
    end
    pii1(i)=exp(pi1(i))/(1+exp(pi1(i)));
end
for i=1:n1
    if pii1(i)>0.5
        ypred(i)=1;
    else
        ypred(i)=0;
    end
end
ypred=ypred';

mat1=confusionmat(ytest,ypred);
sensitifity1=mat1(1,1)./(mat1(1,1)+mat1(1,2));
specifity1=mat1(2,2)./(mat1(2,1)+mat1(2,2));
akurasi_test=sum(diag(mat1))./sum(sum(mat1));
GMean_test=sqrt(sensitifity1*specifity1);
salah=1-GMean_test;
end
```

```
%optimasi paremeter
inisial=[];
n=length(inisial)
option = gaoptimset ('PopulationSize', 100, 'InitialPopulation',
inisial ,'MutationFcn',{@mutationuniform,0.1},'CrossoverFraction',
```

```
[0.8], 'CrossoverFcn',@crossoversinglepoint,'EliteCount', 20,
'SelectionFcn', @selectionroulette)
[a,fval,exitflag,output,population,scores]=ga(@optimasi_parameter,
n,[],[],[],[],[],[],[],option)
```

%Seleksi Regresi Logistik

```
function [salah]= seleksi_variabel(a)
load('data90_10.mat')
ytraining=datatraining(:,1);
xtraining=datatraining(:,2:12);
ytesting=datatesting(:,1);
xtesting=datatesting(:,2:12);
%Mengganti nilai biner dengan nilai matriks
if a==[0 0 0 0 0 0 0 0 0 0 0];
    salah=10;
else
    tr=xtraining;
    ts=xtesting;
    p=length(a);
    j=1;
    for i=1:p
        if a(i)==1
        train(:,j)=tr(:,i);
        test(:,j)=ts(:,i);
        j=j+1;
        end
    end
    salah=reglog_biner(ytraining,train,ytesting,test);

end
salah;
```

%Fungsi Optimasi Parameter Regresi Logistik

```
function [salah]=optimasi_parameter(b)
load('data90_10_bacward.mat')
ytesting=datatesting(:,1);
xtesting=datatesting(:,2:5);
n1=length(xtesting(:,1));
p1=length(xtesting(1,:));

for i=1:n1
    a1=0;
    for j=2:(p1+1)
        a1=a1+(xtesting(i,(j-1))*b(j));
    end
    pi1(i)=a1+b(1);
end
for i=1:n1
    if pi1(i)>50
        pi1(i)=50;
    end
    pii1(i)=exp(pi1(i))/(1+exp(pi1(i)));
end
for i=1:n1
```

```
    if pii1(i)>0.5
        ypred(i)=1;
    else
        ypred(i)=0;
    end
end
ypred=ypred';

mat1=confusionmat(ytesting,ypred);
sensitifity1=mat1(1,1)./(mat1(1,1)+mat1(1,2));
specifity1=mat1(2,2)./(mat1(2,1)+mat1(2,2));
akurasi_test=sum(diag(mat1))./sum(sum(mat1));
GMean_test=sqrt(sensitifity1*specifity1);
salah=1-GMean_test;
end
```

%============naivebayes==================

%seleksi variabel

```
options = gaoptimset ('PopulationSize', 100, 'PopulationType',
'bitstring','FitnessScalingFcn',@fitscalingrank,'InitialPopulation
',
[],'EliteCount',[2],'MutationFcn',{@mutationuniform,0.1},'Crossove
rFraction',[0.8],'CrossoverFcn',@crossoversinglepoint,
'SelectionFcn', @selectionroulette)
[a,fval,exitflag,output,population,scores]=ga(@seleksi_bayes,11,op
tions)
```

%fungsi Naive Bayes

```
function [salah]=naive_bayes(ytrain,xtrain,ytest,xtest)
NB=fitNaiveBayes(xtrain,ytrain);
pred=predict(NB,xtest);
mat1=confusionmat(ytest,pred);
sensitifity1=mat1(1,1)./(mat1(1,1)+mat1(1,2));
specifity1=mat1(2,2)./(mat1(2,1)+mat1(2,2));
akurasi_test=sum(diag(mat1))./sum(sum(mat1));
GMean_test=sqrt(sensitifity1*specifity1);
salah=1-GMean_test;
end
```

```
%seleksi Naive Bayes
function [salah]= seleksi_variabel(a)
load('data90_10.mat')
ytraining=datatraining(:,1);
xtraining=datatraining(:,2:12);
ytesting=datatesting(:,1);
xtesting=datatesting(:,2:12);
%Mengganti nilai biner dengan nilai matriks
if a==[0 0 0 0 0 0 0 0 0 0 0];
    salah=10;
else
    tr=xtraining;
    ts=xtesting;
    p=length(a);
    j=1;
    for i=1:p
```

```
        if a(i)==1
        train(:,j)=tr(:,i);
        test(:,j)=ts(:,i);
        j=j+1;
        end
    end
    salah=naive_bayes(ytraining,train,ytesting,test);

end
salah;
```

BIOGRAFI PENULIS

Penulis lahir di Kota Mataram, Provinsi Nusa Tenggara Barat pada tanggal 03 Desember 1992 dengan nama lengkap Abdurrahman Salim, sebagai anak kedua dari tiga bersaudara dari pasangan H. M. Agus Salim dan Hj. Nur'aini. Penulis menempuh pendidikan formal di TK Dharma Wanita Dahlia Mataram (1997-1998), SD Negeri 36 Mataram (1998-2004), SMP Negeri 4 Mataram (2004-2007) dan SMA Negeri 3 Mataram (2007-2010). Penulis kemudian melanjutkan jenjang S1 di Prodi Matematika FMIPA Universitas Mataram (2010-2015). Penulis melanjutkan studi ke jenjang S2 di Program Pascasarjana Statistika FMIPA Institut Tekonologi Sepuluh Nopember Surabaya (2015-2017).

Saran, kritik, dan pertanyaan seputar tesis ini dapat disampaikan ke alamat email abdurrahmansalim18@gmail.com.

(*halaman ini sengaja dikosongkan*)