



Available online at : <http://bit.ly/InfoTekJar>

## InfoTekJar : Jurnal Nasional Informatika dan Teknologi Jaringan

ISSN (Print) 2540-7597 | ISSN (Online) 2540-7600



# Data Mining untuk Nasabah Bank Telemarketing Menggunakan kombinasi Algoritma Naïve Bayes Dan Algoritma Genetik

Ahmad Asifuddin Aqham, Kristoko Dwi Hartomo

Universitas Kristen Satya Wacana, Salatiga, 50711, Indonesia

### KEYWORDS

telemarketing, naïve bayes, algoritma genetika

### CORRESPONDENCE

Phone: +62 813 2977 2630

E-mail: asif@stekom.ac.id

### A B S T R A C T

Strategi yang digunakan untuk telemarketing dengan melakukan media promosi, strategi tersebut merupakan sebuah cara pemasaran yang digunakan pihak bank, dalam penawaran produk kepada nasabah, bank, salah satu produk yang akan di tawarkan adalah deposito berjangka, pihak bank kesulitan dalam mengetahui kendala yang dialami nasabah dalam pengambilan keputusan untuk melakukan deposito terhadap bank, sehingga nantinya akan memiliki efek krisis keuangan di bank. Telemarketing bank harus memiliki target terhadap nasabah, dimana nasabah yang memiliki potensi untuk bergabung salah satu produk bank, yaitu deposito dengan melihat data nasabah yang ada.

Dengan permasalahan yang ada akan diatasi dengan teknik datamining yang akan digunakan untuk penelitian ini adalah algoritma *Naïve Bayes* dan algoritma genetika yang bertujuan untuk memprediksi keinginan nasabah Telemarketing yang bersumber dari data *public UCI Repository* sehingga bank menawarkan sebuah produknya terhadap nasabah tepat sesuai target. Pengujian *Naïve Bayes* dengan hasil eksperimen akurasi 86,71% sedangkan pengujian cross validation menggunakan algoritma Genetika menghasilkan akurasi tinggi 90,27%, *Root* membuktikan prediksi data time series metode *Naïve Bayes* dan Genetika menghasilkan akurasi sebesar 90,27%, Sehingga dapat di simpulkan bahwa menggunakan algoritma *Naïve Bayes* dan Genetika dapat mengoptimalkan dalam memprediksi keputusan klien Telemarketing tepat dalam penawaran deposito.

*The strategy used for telemarketing by conducting promotional media, this strategy is a marketing method used by banks, in offering products to customers, banks, one of the products that will be offered is time deposits, the bank has difficulty in knowing the obstacles experienced by customers in making a decision to make deposits against the bank, so that later it will have the effect of a financial crisis at the bank. Telemarketing banks must have targets for customers, where customers have the potential to join one of the bank's products, namely deposits by looking at existing customer data.*

*With the existing problems will be overcome by the datamining technique that will be used for this research is the Naïve Bayes algorithm and genetic algorithm which aims to predict the Telemarketing customers' sources sourced from public UCI Repository data so that the bank offers a product to the customer right at the target. Naïve Bayes test with experimental results of 86.71% accuracy while cross validation testing using Genetic algorithm produces high accuracy 90.27%, Root proves the prediction of time series data Naïve Bayes method and Genetics produces an accuracy of 90.27%, so it can be concluded that using the Naïve Bayes algorithm and Genetics can optimize in predicting Telemarketing client decisions right in the deposit offer.*

## PENDAHULUAN

Bank memiliki data besar yang disimpan sebagai *database* dan diolah menghasilkan sebuah informasi yang berkaitan tentang nasabah, data tersebut dapat digunakan untuk menjaga hubungan bank dengan nasabah yang jelas agar dapat menentukan secara individual target penawaran produk bank, agar menjadi pelanggan dengan cara: menghubungi langsung

dalam kontak pribadi, surat, telepon seluler, *e-mail*, maupun media lain dalam melakukan penawaran produk baru maupun pelayanan produk, media promosi ini disebut dengan promosi langsung, mengingat promosi secara langsung banyak digunakan oleh Bank dan perusahaan asuransi dalam melakukan interaksi dengan pelanggan. [1]

Dahulu bank melakukan pemasaran bertemu secara langsung dengan tujuan mempertahankan dan pencarian nasabah baru, tetapi beda dengan sekarang di saat teknologi pertumbuhan

semakin besar, bank dapat memanfaatkan teknologi, salah satunya pemanfaatan telepon, media elektronik yang dapat melakukan komunikasi jarak jauh, dengan demikian bank dapat melakukan promosi kepada konsumen bank secara efektif daripada melakukan penawaran secara langsung maupun lewat brosur dan beretemu langsung.

Bank harus memiliki cara bagaimana mengetahui kendala nasabah telemarketing yang ingin melakukan deposito pada bank, agar bank terhindar dari ancaman krisis keuangan, dengan demikian Bank harus berupaya meningkatkan modal dengan cara menawarkan produk, salah satunya deposito jangka panjang terhadap nasabah.

Telemarketing bank harus memiliki dan membuat sebuah rencana untuk target nasabah mana yang berpotensi, agar dapat ditawarkan dan melakukan deposito dengan menganalisis data klien bank yang ada. [2],

Penulis ingin menguji keberhasilan telemarketing dengan cara melakukan prediksi keputusan nasabah dengan menerapkan *machine learning* sehingga mendapatkan sebuah informasi dari pola yang ada dan mengekstraksi informasi, salah satu algoritma klasifikasi yang digunakan yaitu *Naïve Bayes* dengan metode algoritma sebagai yang menghasilkan klasifikasi nasabah bank telemarketing, data yang digunakan merupakan data primer dan sekunder yang didapat dari data public UCI WEKA yang terdiri dari 45211 data dan 17 atribut, dari hasil penelitian diharapkan memberikan referensi mengenai akurasi dari klasifikasi nasabah bank telemarketing yang menggunakan algoritma *naïve bayes* dan penggunaan algoritma *naïve bayes* yang disertai pemilihan variabel menggunakan algoritma genetika sebagai *feature selection*.

## STATE OF THE ART

Metode *Naïve Bayes* juga digunakan pada penelitian Dwi Cahaya, P, B. (2016) dengan menerapkan metode *Naïve Bayes* untuk mengetahui tingkat kesuburan, dikarenakan terdapat sebuah kasus yang di temui oleh peneliti dalam dua *decade* tingkat kesuburan (Fertilitas) mengalami penurunan, yang menyebabkan kesuburan menurun salah satunya karena faktor lingkungan dan gaya hidup, misalkan alkohol, rokok yang dapat mempengaruhi kualitas sperma, sehingga peneliti menggunakan *Naïve Bayes* untuk memprediksi faktor yang sangat berpengaruh didalam penentuan kesuburan, di dalam pengklasifikasi masih terdapat kekurangan sehingga peneliti berusaha untuk meningkatkan dengan cara mengoptimalkan kinerja metode *Naïve Bayes* dengan menambahkan Algoritma Genetika yang awal mula *result* 97,66% menjadi 99,33% [3]

Risa Wati (20016) membahas tentang bagaimana mendapatkan sebuah klasifikasi yang tepat untuk menentukan opini positif dan negatif dari hasil review kualitas layanan maskapai, dengan menggunakan algoritma *Naïve Bayes* tetapi dirasa masih memiliki kelemahan untuk pemilihan fitur sehingga penulis menggunakan algoritma genetika yang merupakan salah satu algoritma optimasi dengan nilai awal 60,00% menjadi 89,50%. [4]

Penelitian yang dilakukan oleh Oman Somantri dan Mohammad Khambali (2017) dengan judul "Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan *Naïve Bayes* dan Algoritma Genetika" dapat disimpulkan bahwa penelitian tersebut penulis membahas tentang bagaimana menentukan cerpen yang sesuai dengan usia dan keinginan pembaca, sehingga peneliti mengusulkan untuk membuat sistem pengambil keputusan agar dapat mengklasifikasi sesuai jenis dan kategori

cerpen, pada penelitian ini penulis menggunakan metode algoritma *Naïve Bayes* dikarenakan metode tersebut sesuai untuk mengklasifikasi jenis kategori cerpen, penulis juga melakukan percobaan dalam optimasi metode tersebut dikarenakan hasil dari akurasi pengujian metode *Nive Bayes* masih kurang, sehingga peneliti menggunakan Algoritma Genetika sebagai *feature selection* untuk membantu mengoptimalkan hasil akurasi yang bermula dari 78,59% menjadi 84,29%, sehingga dapat ditarik kesimpulan menggunakan algoritma Genetika dan *Naïve Bayes* dapat digunakan untuk penentuan cerpen sesuai minat baca dan usia [5].

Penelitian Yaning wu, Song Huang dkk (2017) Prediksi kerusakan perangkat lunak berperan penting dalam mengidentifikasi modul-modul yang paling rusak sebelum pengujian perangkat lunak. Cukup banyak peneliti telah melakukan upaya besar untuk meningkatkan prediksi ketepatan. Namun, masalah data historis yang tidak memadai masih tersedia untuk dalam atau lintas proyek tetap tidak terselesaikan. Lebih lanjut, adalah praktik umum untuk menggunakan fungsi kerapatan probabilitas untuk fungsi normal distribusi dalam klasifikasi *Naïve Bayes* (NB). Namun demikian, setelah melakukan tes *Kolmogorov Smirnov*, Kami menemukan bahwa 21 metrik perangkat lunak utama tidak terdistribusi secara normal pada tingkat signifikansi 5%. Oleh karena itu, makalah ini mengusulkan *classifier Bayes* baru, yang mengembangkan *classifier* NB menjadi non-normal fungsi difusi informasi, untuk membantu memecahkan masalah kurangnya data pelatihan yang sesuai untuk yang baru proyek. Kami melakukan tiga percobaan pada 34 set data yang diperoleh dari 10 proyek sumber terbuka, hanya menggunakan Masing-masing 10%, 6,67%, 5%, 3,33%, dan 2% dari total data untuk pelatihan. Empat klasifikasi terkenal Algoritma juga termasuk untuk perbandingan, yaitu *Regresi Logistik*, *Naïve Bayes*, *Random Tree* dan Mendukung *Mesin Vektor*. Semua hasil eksperimen menunjukkan efisiensi dan kepraktisan yang baru penggolong. [6]

Penelitian S. A. Mustofa, Aida, dkk (2019) membahas tentang teknik bagaimana Diagnosis penyakit Parkinson yang dilakukan oleh spesialis melibatkan beberapa pemeriksaan pemikiran kita, dan fisik. Spesialis menganalisa sejumlah gejala dan tanda ketika memeriksa kondisi sistem saraf seseorang, Diagnosa dengan cara peninjauan literature medis dan faktor genetik orang yang menderita penyakit parkinson. Cara untuk mengatasi diagnosis terkini untuk penyakit Parkinson bergantung pada analisis gangguan suara. Metodologi ini mencakup penggalan set fitur suara orang yang direkam kemudian dimanfaatkan untuk *mezin learning*, dimana teknik tersebut digunakan untuk mengidentifikasi kasus Parkinson yang benar berdasarkan suara seseorang. Penulis mencoba untuk meningkatkan diagnosis penyakit Parkinson dengan menguji beberapa evaluasi fitur dan metode *mezin learning* salah satunya dengan pemanfaat data mining klasifikasi berdasarkan analisis gangguan suara. Tujuan dari penelitian ini adalah untuk menemukan solusi optimal dalam mengatasi masalah tersebut dengan mengusulkan Pendekatan Evaluasi Fitur Berganda (MFEA) baru dari sistem multi-agen dan menerapkan lima skema klasifikasi independen diantaranya meliputi: *Decision Tree*, *Naive Bayes*, *Neurel Network*, *Randem Forests*, dan *Support Vector Machine* pada diagnosis Parkinson sebelum dan sesudah menerapkan MFEA, setelah itu mengevaluasi akurasi diagnosis hasil. Metodologi tes meliputi 10 kali lipat validasi silang untuk mengevaluasi pembelajaran metode ods dan lacak variasi dalam kinerja mereka. Hasil pengujian menunjukkan bahwa MFEA dari sistem multi-agen menemukan serangkaian fitur terbaik dan meningkatkan kinerja klasifikasi. Tingkat rata-rata peningkatan dalam akurasi diagnostik dari klasifikasi adalah *Dicision Tree* 10,51%, *Naive Bayes* 15,22%, *Neural Network* 9,19%, *Random*

Forests 12,75%, dan Support Vector Machine 9,13%. Ini hasil menunjukkan bahwa MFEA membuat peningkatan yang signifikan pada hasil diagnosis klasifikasi, sehingga dapat disimpulkan bahwa metode yang tepat untuk menganalisis penyakit parkison adalah Naïve Bayes, dikarenakan dari hasil ujian metode tersebut mendapatkan akurasi yang paling tinggi dengan akurasi 15,22%. [7]

Penelitian H. Muhammad, C. A. Prasojo, dkk (2017) membahas tentang optimasi *Naïve Bayes Classifier* dengan PSO dari data iris. Penelitian tersebut melakukan klasifikasi dengan cara pembobotan atribut yang timal menggunakan PSO. Hasil diperoleh dari nilai *fitness* tertinggi, dan dilakukan 2 pengujian dengan jumlah partikel serta kombinasi parameter, jumlah partikel yang diuji 10 hingga 50 dengan pengujian 3 kali dengan hasil nilai rata-rata *fitness* tertinggi 97,39 dengan partikel 50, pada pengujian kombinasi parameter, membangkitkan nilai kombinasi dengan cara random dan diuji sebanyak 3 kali, dengan hail *fitness* 97,39 [8]

Penelitian Admaja dwi herlambang, S. H. Wijoyo (2019) melakukan penelitian di SMK TIK dan membahas tentang klasifikasi sumber belajar mata pelajaran produktif dari media internet, penelitian tersebut melakukan klasifikasi berdasarkan kriteria ciri esensial setiap mata pelajaran produktif di SMK TIK, pengelompokkan dengan algoritma *Naïve Bayes* yang digunakan untuk dokumen teks dan menganggap semua atribut tidak berhubungan, Tujuan penelitian ini adalah untuk mendeskripsikan hasil klasifikasi dan evaluasi kualitas klasifikasi sumber belajar berbasis teks dengan menggunakan algoritma *Naïve Bayes* dan pengujian menghasilkan nilai akurasi tertinggi 81,48%, sedangkan nilai akurasi terendah sebesar 79,63%. [9]

Dari kajian relevan yang ada, terdapat perbedaan indicator Age, Kridit (Default), Housing, Loan, Marital, Education, Contact, Balance, Duration, Campaign, Pdays, Previous, Putcome, Job, Serta Label Yes atau no, dan kasus dengan penerapan algoritma *Naïve Bayes*, sehingga penulis menarik kesimpulan bahwa dari penelitian yang ada *Naïve Bayes* merupakan metode yang cepat dalam melakukan perhitungan, agoritma yang sederhana dan akurasi yang tinggi namun memiliki kelemahan pada seleksi atribut sehingga peneliti ingin memperbaiki metode tersebut dengan menerapkan algoritma Genetika mampu menjadi seleksi fitur untuk meningkatkan hasil dari proses *metode Naïve Bayes*.

## LANDASAN TEORI

### *Naïve Bayes*

*Naïve Bayes* merupakan algoritma klasifikasi yang terbukti memiliki akurasi dan kecepatan yang tinggi ketika diaplikasikan pada database yang besar dan metode ini digunakan dalam statistika untuk menghitung peluang dari sebuah hipotesis berdasarkan atribut, kelas dan nilai probabilitas tertinggi yang dimiliki sebuah database [10], dengan rumus:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Dimana:

X : Data yang belum diketahui kelasnya

H : Hipotesis data X merupakan suatu kelas spesifik

P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X

P(H) : Probabilitas hipotesis H

P(X|H) : Probabilitas X Berdasarkan kondisi H

P(X) : Probabilitas dari X

Dapat disederhanakan sebagai berikut :

$$P(H|X) = P(X|H)P(H) \quad (2)$$

*Naïve Bayes* terdapat kelemahan dalam penentuan atribut, dimana masih terjadi kesalahan sehingga kurang optimal dalam menentukan hasil estimasi probabilitas, untuk mengatasi tersebut diperlukan cara dengan metode pembobotan yang menghasilkan peningkatan akurasi dari metode *Naïve Bayes* [11]

### *Fiture Selection*

Dalam penggunaan Mechine Learning terdapat proses *feature selection* untuk pembelajaran algoritma, akurasi didapatkan dari kontribusi dimensi angka yang sedikit dari subset dalam proses *feature selection* akan memfilter sisa dari dimensi yang tidak berkepentingan sehingga akan meningkatkan performa dan akurasi dari sebuah metode algoritma [12]

### *Optimasi*

Optimasi merupakan sebuah proses yang digunakan untuk menyelesaikan masalah agar mendapatkan kondisi yang paling menguntungkan dari sudut pandang, masalah tersebut berkaitan dengan variabel dari kumpulan data yang berkaitan. Yang di maksud dengan kondisi yang paling menguntungkan merupakan pencarian nilai minimum maupun maksimum tergantung yang diinginkan [13].

### *Optimasi*

Optimasi merupakan sebuah proses yang digunakan untuk menyelesaikan masalah agar mendapatkan kondisi yang paling menguntungkan dari sudut pandang, masalah tersebut berkaitan dengan variabel dari kumpulan data yang berkaitan. Yang di maksud dengan kondisi yang paling menguntungkan merupakan pencarian nilai minimum maupun maksimum tergantung yang diinginkan [13].

### Algoritma Genetika

Adalah salah satu algoritma pencarian dengan cara seleksi alam dan genetika alam, genetika dimulai dengan sekumpulan individu yang membentuk populasi, dan satu individu adalah satu solusi, dari populasi akan berubah menjadi iterasi baru, dan pada iterasi terakhir salah satu populasi akan kembali menjadi solusi yang terbaik untuk mengatasi masalah yang terjadi, sehingga dapat diproseskan perubahan yang terjadi [14], Trevino dan Falciani (2006) ada 7 tahap dalam penerapan algoritma genetika yaitu :

- 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 permasalahan. Gen pada algoritma genetika berisi variabel yang ingin dioptimumkan, pada penelitian ini gen berisi variabel independen.

$$POP = \text{round}\{\text{random}(N_{Ipop}, N_{Ibits})\} \quad (3)$$

Dimana:

POP = merupakan gen yang berisi pembulatan dari bilangan acak yang dibangkitkan

( $N_{Ipop}$ ) = (jumlah populasi)

( $N_{Ibits}$ ) = (jumlah gen dalam tiap kromosom)

- Masing-masing kromosom dalam populasi dievaluasi kemampuannya dengan menggunakan fungsi *fitness*. Pada penelitian ini fungsi *fitnesnya* berupa kesalahan klasifikasi.

$$a + 2b + 3c = 30 \tag{4}$$

Keterangan:

a, b, c, merupakan varuabel yang digunakan untuk gen-gen dalam membuat *chromosome*. Variabel a memiliki bilangan integer 0 sampai 30. b, c, dan d memiliki bilangan integer 0 sampai 10.

Rumus probabilitas

$$P(i) = \frac{Fitness(i)}{Total\ Fitness} \tag{5}$$

- c. 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 fitness terkecil yang dipilih sebagai solusi permasalahan dari penelitian ini, karena fungsi fitness yang digunakan adalah tingkat kesalahan klasifikasi.

$$Q(i) = \frac{1}{Fitness(i)} \tag{6}$$

- d. Memilih kromosom dengan nilai fitness yang optimum yang dijadikan orang tua.

$$Ci = C(n) + P(n + 1) \tag{7}$$

- e. Mengkombinasikan informasi genetika yang ada dalam replikasi orang tua melalui pindah silang. Dua induk secara random dipilih dan digunakan untuk membentuk dua kromosom baru. bilangan acak R antara 0 sampai dengan 1, bilangan acak dipilih sesuai dengan jumlah cromosom:

$$R = R(i) \tag{8}$$

- f. Melakukan mutasi untuk memperkenalkan unsur gen baru pada kromosom secara acak.

$$C[x - 1] < R < C[x] \tag{9}$$

- g. Tahapan diulangi dari tahapan 2 sampai kromosom yang memberikan nilai fitness paling optimum atau sudah mencapai konvergen [15].

## ANALISA DAN HASIL

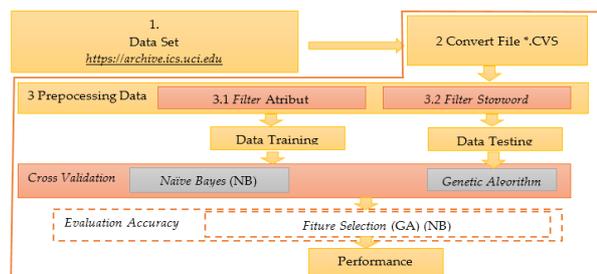
Penelitian ini melakukan eksperimen dengan maksud untuk mendapatkan penunjang keputusan dalam memprediksi nasabah bank yang ingin membeli produk bank yaitu deposito dengan menggunakan metode Naïve Bayes dan di optimalkan dengan metode Genetika data yang akan digunakan merupakan data primer dan skunder yang didapat dari data public UCI WEKA dengan pengujian menggunakan tools Rapidminer 9.0 adapun tahapan sebagai berikut:



Gambar 1. Tahapan Penelitian

Dari gambar 1. Pengumpulan data ditentukan dengan pencريان data ke Website Uci Weka, dan data tersebut diintegrasikan kedalam data set, setelah pengumpulan data, pada pengolahan data, data yang sudah di jadikan data set dilakukan penyeleksian atribut, sehingga tertrasformasikan kedalam bentuk yang dibutuhkan, setelah data terkumpul dan diolah data akan diberi model dan membagi menjadi dua data training dan data testing, setelah mendapatkan metode penulis akan melakukan

Melakukan perhitungan dengan masingmasing *Algoritma Naïve Bayes* dan Genetika yang akan diulang beberapa kali sampai menghasilkan nilai akurasi tertinggi. Tahap evaluasi merupakan tahap akhir dari rangkaian kegiatan penelitian ini. Setelah melakukan tahap pengujian model maka akan menghasilkan nilai akurasi dan AUC. Kemudian dari hasil itu dievaluasi, dari hasil evaluasi itu dapat ditarik kesimpulan dari hasil penelitian ini dapat dilihat pada gambar 2.



Gambar 2. Kerangka Pemikiran

Dari gambar 2. dapat dijelaskan kerangka pemikiran dalam penelitian ini tahapan awal melakukan pengumpulan data set yang didapat dari data public UCI WEKA dan di unduh pada laman <https://archive.ics.uci.edu>., dari data yang ada penulis merubah data yang awalmula berbentuk excel merubah menjadi \*.cvs, tahapan selanjutnya melakukan praprocessing data dengan pemilihan atribut yang digunakan dan pembersihan tanda baca pada record atribut, penulis membagi data tersebut menjadi 2 kelompok yaitu: data training (data asli) dan testing (sampel data yang diambil dari data training tetapi *record* yang ada di atribut dirubah dan tidak diberikan label, kedua data tersebut akan diuji dan mendapatkan sebuah prediksi dari metode naïve bayes dan kombinasi *naïve bayes* genetika, hasil pengujian akan dibandingkan dengan melihat performance yang lebih tinggi.

**Pengumpulan Data**

Data yang digunakan menggunakan data UCI WEKA, terdiri dari 4522 record, 16 atribut dan 1 atribut target, yang akan digunakan sebagai output data konsumen telemarketing, yang dapat dilihat pada tabel 1.

Tabel 1. Tabel Data Konsumen Telemarketing

| No. | Variabel | Atribut | Keterangan   |
|-----|----------|---------|--|
| .   | A        | Age     | Merupakan atribut yang akan di katagorikan menjadi 3 dengan menggunakan referensi WHO tentang klasifikasi age.<br>Child (<=14), young – adult (>=15 and <=49), old (>=50)  |
| 2.  | A2       | Job     | Atribut dari konsumen dengan jenis pekerjaan yang dikatagorikan: “Admin”, “unkown”, “unemployed”, “management”, “housemaid”, “entrepreneur”, “student”, “blue-collar”, “selfemployed”, “retaired”, “technician”, “services”. |
| 3.  | A3       | Marital | Atribut status perkawinan dengan katagori: “married”, “divorced”, “single”.  |
| -   | -        | -       | -  |
| 17. | A17      | Output  | Output = Atribut dengan targen yang akan mendepositokan, katagori yaes atau no, dan data di ubah menjadi numeric untuk mempermudah pengerjaan yaitu Yes = 2 dan No = 1.  |

Untuk melakukan pengujian di persiapkan data yang akan digunakan dengan cara pengambilan data secara acak

Tabel 2. Data Training

| Data Training |          |           |          |   |   |           |
|---------------|----------|-----------|----------|---|---|-----------|
| No.           | 1        | 2         | 3        | - | - | 28        |
| Age           | 70       | 63        | 60       | - | - | 37        |
| Job           | Retired  | Retired   | Services | - | - | Entrepre  |
| Marital       | Married  | Married   | Married  | - | - | Married   |
| Education     | Primary  | Secondary | Tertiary | - | - | Secondary |
| Balance       | 324      | 1495      | 4256     | - | - | 2971      |
| Housing       | No       | No        | Yes      | - | - | No        |
| Loan          | No       | No        | No       | - | - | No        |
| Contact       | Celluler | Celluler  | Celluler | - | - | Celluler  |
| Day           | 15       | 16        | 16       | - | - | 17        |
| Month         | Nov      | Nov       | Nov      | - | - | Nov       |
| Duration      | 78       | 138       | 200      | - | - | 361       |
| Campaign      | 1        | 1         | 1        | - | - | 2         |
| Pdays         | 96       | 22        | 92       | - | - | 188       |
| Previous      | 7        | 5         | 4        | - | - | 11        |
| Putcome       | Success  | Success   | Success  | - | - | other     |
| Y             | No       | No        | Yes      | - | - | No        |

Data tabel 2 merupakan data atribut dengan Age, Kridit (Default), Housing, Loan, Marital, Education, Contact, Balance, Duration, Campaign, Pdays, Previous, Putcome, Job, Serta Label Yes atau no, yang dipilih secara acak yang akan dilakukan transformasi pada data yang telah dipilih sehingga sesuai dengan proses data mining dalam bentuk MS. Excel karena data akan digunakan dalam tools rapid miner dan di tentukan bebrapa atribut yang akan digunakan.

**Pengujian data dan hasil**

Pada tahap ini dilakukan pengujian data training sesuai tujuan penelitian yaitu untuk menerapkan teknik klasifikasi menggunakan metode optimasi *naive bayes* dengan genetika untuk mendapatkan keputusan dalam penentuan nasabah untuk melakukan deposito.

Tabel 3. (a) Data Testing Dan (b) data Training

| Data Testing |          |
|--------------|----------|
| No.          | 1        |
| Age          | 22       |
| Job          | Student  |
| Marital      | Single   |
| Education    | Primary  |
| Balance      | 3472     |
| Housing      | No       |
| Loan         | No       |
| Contact      | Celluler |
| Day          | 19       |
| Month        | Apr      |
| Duration     | 529      |
| Campaign     | 1        |
| Pdays        | 181      |
| Previous     | 4        |
| Putcome      | Success  |
| Y            | ?????    |

(a)

| Data Training |          |           |   |           |
|---------------|----------|-----------|---|-----------|
| Indikator     | No       |           |   |           |
|               | 1        | 2         | - | 28        |
| Age           | 70       | 63        |   | 37        |
| Job           | Retired  | Retired   |   | Entre     |
| Marital       | Married  | Married   |   | Marr      |
| Education     | Primary  | Secondary |   | Seco<br>y |
| Balance       | 324      | 1495      |   | 2971      |
| Housing       | No       | No        |   | No        |
| Loan          | No       | No        |   | No        |
| Contact       | Celluler | Celluler  |   | Cellu     |
| Day           | 15       | 16        |   | 17        |
| Month         | Nov      | Nov       |   | Nov       |
| Duration      | 78       | 138       |   | 361       |
| Campaign      | 1        | 1         |   | 2         |
| Pdays         | 96       | 22        |   | 188       |
| Previous      | 7        | 5         |   | 11        |
| Putcome       | Success  | Success   |   | other     |
| Y             | No       | No        |   | No        |

(b)

Tabel 3. Merupakan tabel training dan testing, tabel (a) menjadi tabel target yang akan diprediksi untuk mendapatkan sebuah label dari pengujian metode yang dapat dilihat pada tabel testing atribut y yang masih ditanyakan labelnya, (b) sebagai tabel asli atau tabel training, dari tabel tabel training akan dilakukan perhitungan masing-masing label dengan menggunakan kriteria yes dan no tahapan sebagai berikut:

- Tahap 1 Menghitung masing-masing kelas dan label  
 $Y (Y = \text{Yes}) = 19/28$   
 $y (Y = \text{No}) = 9/28$
- Tahap 2 Menghitung kasus yang sama dengan kelas yang sama

Tabel 3. (a) Pengujian atribut poutcome (b) Pengujian atribut previous (c) Data training

| No. | poutcome | y                   |
|-----|----------|---------------------|
| 1   | success  | no                  |
| 2   | success  | no                  |
| 3   | success  | yes                 |
| 4   | failure  | no                  |
| 5   | unknown  | yes                 |
| 6   | success  | yes                 |
| 7   | failure  | no                  |
| 8   | success  | yes                 |
| 9   | failure  | yes                 |
| 10  | unknown  | yes                 |
| -   | -        | -                   |
| -   | -        | -                   |
| 28  | other    | No                  |
| Hsl |          | Success = 7<br>7/19 |
| Hsl |          | Success = 2<br>2/9  |

(a)

| No. | previous | y                                  |
|-----|----------|------------------------------------|
| 1   | 7        | no                                 |
| 2   | 5        | no                                 |
| 3   | 4        | yes                                |
| 4   | 2        | no                                 |
| 5   | 0        | yes                                |
| 6   | 2        | yes                                |
| 7   | 5        | no                                 |
| 8   | 3        | yes                                |
| 9   | 1        | yes                                |
| 10  | 0        | yes                                |
| -   | -        | -                                  |
| -   | -        | -                                  |
| 28  | 11       | No                                 |
| Hsl |          | Previous 4 = 3<br>Yes = 19<br>3/19 |
| Hsl |          | previous = 2<br>No = 9<br>2/9      |

(b)

| Atribut   |              | Statu<br>s<br>Labe<br>l<br>Yes<br>atau<br>No | Jumlah               |                    | Hasi<br>l |
|-----------|--------------|--|----------------------|--------------------|-----------|
| Nama      | Nilai        |  | Nilai<br>Atribu<br>t | Nilai<br>Labe<br>l |           |
| Pdays     | 181          | Yes  | 0                    | 19                 | 0/19      |
| Pdays     |              | No   | 0                    | 9                  | 0/9       |
| campaign  | 1            | Yes  | 13                   | 19                 | 13/19     |
| campaign  |              | No   | 4                    | 9                  | 4/9       |
| duration  | 529          | Yes  | 0                    | 19                 | 0/19      |
| duration  |              | No   | 0                    | 9                  | 0/19      |
| contact   | cellula<br>r | Yes  | 19                   | 19                 | 19/19     |
| contact   |              | No   | 8                    | 9                  | 8/9       |
| Loan      | No           | Yes  | 18                   | 19                 | 18/19     |
| Loan      |              | No   | 8                    | 9                  | 8/9       |
| Housing   | No           | Yes  | 13                   | 19                 | 13/19     |
| Housing   |              | No   | 7                    | 9                  | 7/9       |
| Balance   | 3472         | Yes  | 0                    | 19                 | 0/19      |
| Balance   |              | No   | 0                    | 9                  | 0/9       |
| Default   | No           | Yes  | 19                   | 19                 | 19/19     |
| Default   |              | No   | 9                    | 9                  | 9/9       |
| education | Primar<br>y  | Yes  | 1                    | 19                 | 1/19      |
| education |              | No   | 1                    | 9                  | 1/9       |
| Marital   | Single       | Yes  | 9                    | 19                 | 9/19      |
| Marital   |              | No   | 2                    | 9                  | 2/9       |
| Job       |              | Yes  | 2                    | 19                 | 2/19      |

|     |         |     |   |    |      |
|-----|---------|-----|---|----|------|
| Job | Student | No  | 0 | 9  | 0/9  |
| Age | 22      | Yes | 0 | 19 | 0/19 |
| Age |         | No  | 0 | 9  | 0/9  |

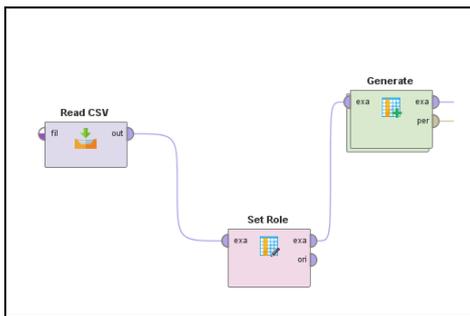
(c)

Tabel 3 merupakan perhitungan kasus yang sama dengan kelas sama dengan contoh tabel (a) menghitung jumlah atribut putcome dengan kasus success dan kelas yes, serta kasus *success* dan kelas no. tabel (b) menghitung jumlah atribut previous dengan kasus 4 dan kelas yes, serta kasus 4 dan kelas no. tabel (c) merupakan kelanjutan perhitungan atribut yang lain sampai semua atribut dihitung.

c. Tahap 3 Mengalikan semua hasil kasus  
 $7/19 \times 2/9 \times 3/19 \times 0/9 \times 0/19 \times 0/9 \times 13/19 \times 4/9 \times 0/19 \times 0/9 \times 19/19 \times 8/9 \times 18/19 \times 8/9 \times 13/19 \times 7/9 \times 0/19 \times 0/9 \times 19/19 \times 9/9 \times 1/19 \times 1/9 \times 9/19 \times 2/9 \times 2/19 \times 0/9 \times 0/19 \times 0/9 = 0$

Dari hasil perhitungan dengan metode naïve bayes menghasilkan nilai 0 yang memiliki arti bahwa data testing akan diprediksi (Yes)

Pengujian model yang dilakukan dalam memprediksi nasabah bank dengan algoritma *naïve bayes* dengan bantuan tools *rapidminer* dapat menentukan nilai *accuracy* dan AUC. Didalam proses penentuan nilai *accuracy* penulis menggunakan cross validation dengan desain model Gambar 3 :

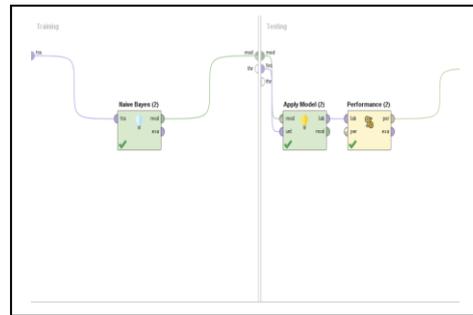


Gambar 3. Model Validasi

Pada penelitian ini penulis akan menerapkan metode naïve bayes

| Accuracy: 90.27% +/- 0.90% (micro average: 90.27%) |                 |
|--|-----------------|
|  | class precision |
| pred no  | 92.56%          |
| pred yes   | 61.95%          |
| class recall                                       | 40.31%          |

sebagai pengambil keputusan untuk menentukan nasabah bank yang akan membeli produk deposito bank pada gambar 3.



Gambar 4. Pengujian data menggunakan metode *Naïve Bayes*

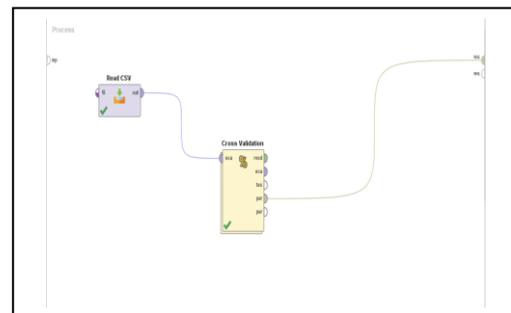
Sehingga dari pengujian tersebut didapatkan sebuah hasil yang berupa *accuracy*, *precision*, dan nilai *recall* dari data *training* dengan bantuan tools *rapidminer*, yang dapat dilihat pada gambar 5.

|              | true no | true yes | class precision |
|--------------|---------|----------|-----------------|
| pred no      | 3660    | 281      | 93.34%          |
| pred yes     | 340     | 260      | 43.33%          |
| class recall | 91.50%  | 49.90%   |                 |

Gambar 5. Hasil *Accuray* dari pengujian Metode

Dari gambar 5. Dapat disimpulkan bahwa dari pengujian menggunakan metode *naïve bayes* mendapatkan akurasi 86,71%, dengan demikian peneliti melakukan optimasi dengan metode Genetika.

Pengujian metode *naïve bayes* dengan genetika sebagai untuk menentukan nasabah bank yang akan membeli produk deposito pada gambar 6.



Gambar 6. Pengujian data menggunakan metode *Naïve Bayes* dan Genetika

Sehingga dari pengujian tersebut didapatkan sebuah hasil yang berupa *accuracy*, *precision*, dan nilai *recall* dari data *training* dengan bantuan tools *rapidminer*, yang dapat dilihat pada gambar 7.

Dari gambar 7. menunjukkan bahwa pengujian dengan algoritma *Naive Bayes* dan Genetika menunjukkan bawa mengalami peningkatan *accuray* dengan peningkatan 4% dari hasil algoritma *Naive Bayes*, acuracy Naive bayes dan Genetika 90.27% dengan demikian peneliti melakukan optimasi dengan metode Genetika.

d. Penggunaan Algoritma *Naive Bayes*

Input : NB(C, D)

V ← Banyaknya atribut Nasabah Bank (17 jenis)

N ← Banyaknya Kelas (2 Kelas)

For each  $c \in C$  do

$N_c \leftarrow$  Count Docs In Class (D, c)

prior [c] ←  $N_c/N$

$L_c \leftarrow$  Count Case Of All Docs in Class (D, c)

For each  $t \in v$  do

$T_{ct} \leftarrow$  Count Case Of Therm ( $L_c, t$ )

For each  $t \in$

do  $condprob [t][c] \leftarrow \left( \frac{t_{ct} + 1}{(\sum_{t'} - (T_{ct'} + 1))} \right)$

V, prior, condprob

NB (C, V, Prior, condprob, d)

W ← Extract Case From Doc (V, d)

For each  $c \in C$  do

Score[c] ← Log prior [c] do

Score[c] ← Log condprob[t] [c]

Return  $arg\ max_{c \in C} Score [c]$

Keterangan :

Untuk masing-masing atribut Yes merupakan himpunan Kelas, lalu menghitung jumlah Yes didalam kelas yang sama di keseluruhan atribut, sehingga mendapatkan nilai prior yes dan kasus yang sama, hitung semua kelas dalam 1 kasus, menghitung jumlah No didalam kelas yang sama pada keseluruhan atribut, sehingga mendapatkan nilai prior no pada kasus yang sama, menghitung kondisi probabilitas Yes dan No dan mendapatkan kelompok atribut yang dipilih dan dilakukakan secara berula

Dilakukan penilain jika  $v \leq$  titik potong di interval, nilai v masuk interval pertama sehingga frekuensi di interval tersebut ditambah. Lakukan penambahan  $count [c][label\ kelas]$  dan rekam interval keberapa yang mengalami perubahan, Jika  $v >$  titik potong di interval pertama maka lanjut ke interval selanjutnya dan lakukan hal yang sama di poin 10. Jika nilai error tidak berubah dari error sebelumnya maka keluar dari sistem dan jika berubah maka keadaan TRUE terpenuhi.

e. Langkah-langkah penyelesaian dengan algoritma genetika:

Membangkitkan Populasi

$$/ POP = round\{random(N_{ipop}, N_{ibits})\} \dots (10)$$

POP = merupakan gen yang berisi pembulatan dari bilangan acak yang dibangkitkan

( $N_{ipop}$ ) = (jumlah populasi)

( $N_{ibits}$ ) = (jumlah gen dalam tiap kromosom)

Prediksi (Yes) dan (No), dengancara dalam menyelesaikan masalah kombinasi. Pada persamaan  $a+2b+3c= 30$ , kita cari nilai a, b, c, dan d yang memenuhi persamaan diatas. menggunakan algoritma genetika untuk menyelesaikan permasalahan.

$$c1 = [07\_13\_17]$$

$$c2 = [03\_05\_07]$$

Evaluasi nilai fitness

Dengan rumus yaitu  $a+2b+3c= 30$ , akan menghasilkan nilai fitness pada setiap kromosom,

$$\begin{aligned} \text{Fitness C1} &= 1/(1 + |(a+3b+2c) - 30|) \\ &= 1/(1+|((7+(3*13)+(2*17))-30)|) \\ &= 0,019608 \end{aligned}$$

$$\begin{aligned} \text{Fitness C2} &= 1/(1 + |(a+3b+2c) - 30|) \\ &= 1/(1 + |((3)+(3*5)+(2*7))-30|) \\ &= 0,333333 \end{aligned}$$

$$\text{Total nilai fitness C1+C2} = 0,019608 + 0,333333 = 0,352941$$

Rumus untuk mencari probabilitas:  $P[i] = \text{fitness}[i] / \text{total\_fitness}$

$$P[1] = 0,019608/0,352941 = 0,055556$$

$$P[2] = 0,333333/0,352941 = 0,944444$$

Probabilitas diatas dapat kita lihat kalau kromosom ke 2 yang mempunyai fitness paling besar maka kromosom tersebut mempunyai probabilitas untuk terpilih pada generasi selanjutnya lebih besar dari kromosom lainnya.

Menentukan kromosom induk

Untuk proses seleksi kita gunakan roulette wheel, untuk itu kita harus mencari dahulu nilai kumulatif probabilitasnya:

$$\begin{aligned} C[1] &= 0,055556 \\ C[2] &= 0,055556 + 0,944444 = 1 \end{aligned}$$

Langkah selanjutnya adalah dengan menggunakan bilangan acak R antara 0 sampai dengan 1, bilangan acak dipilih sesuai dengan jumlah *chromosom*:

$$R[1] = 0,066667$$

$$R[2] = 0,25$$

Memilih kromosom ke x sebagai *parent* dengan syarat  $C[x-1] < R < C[x]$ . Angka acak  $R[1] <$  nilai kumulatif dari C2, sehingga C2 nanti akan dilakukan *crossover* dengan C1. Hasil seleksi

Roulette Wheel pada populasi ini untuk *crossover* menjadi:

$$C1 \text{ Menjadi } C2 = [03\_05\_07]$$

$$C2 \text{ Menjadi } C2 = [03\_05\_07]$$

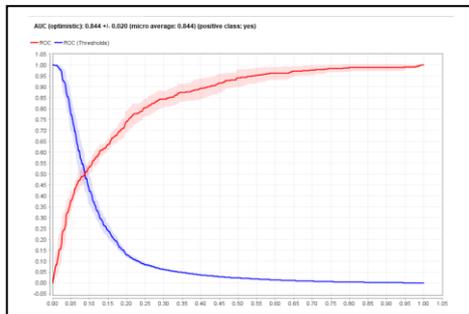
kromosom-kromosom ini akan mengalami proses yang sama seperti generasi sebelumnya yaitu proses evaluasi, seleksi, *crossover* dan mutasi yang kemudian akan menghasilkan kromosom-kromosom baru untuk generasi yang selanjutnya. Proses ini akan berulang sampai sejumlah generasi yang telah ditetapkan sebelumnya.

Pengujian *Cross Validation* didapatkan akurasi 90,27% dibanding dengan akurasi pada pengujian *Naive Bayes* dengan nilai 86,71%, algoritma genetika berhasil mengoptimalkan metode *Naive Bayes* sehingga metode *Naive Bayes* dan genetika akan digunakan untuk penentuan nasabah bank yang akan membeli produk deposito. Dan dapat dirumuskan dengan *accuracy* dari *confusion matrix* adalah sebagai berikut:

$$\begin{aligned} a) \text{ Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\ &= \frac{210+3871}{210+3871+311+129} = \frac{4081}{4521} \\ &= 0,9027 = 90,27\% \end{aligned}$$

$$\begin{aligned}
 \text{b) Sensitivity} &= \frac{TP}{TP+FP} \\
 &= \frac{210}{210+129} = \frac{210}{339} \\
 &= 0,6195 = 61,95\% \\
 \text{c) Specificity} &= \frac{TN}{TN+FP} \\
 &= \frac{3871}{3871+311} = \frac{3871}{4182} \\
 &= 0,9256 = 92,56\% \\
 \text{d) PPV} &= \frac{TP}{TP+FP} \\
 &= \frac{210}{210+311} = \frac{210}{521} \\
 &= 0,4031 = 40,31\% \\
 \text{e) NPV} &= \frac{TN}{TN+FN} \\
 &= \frac{3871}{3871+129} = \frac{3871}{4000} \\
 &= 0,9678 = 96,78\%
 \end{aligned}$$

Hasil dari pengujian mendapatkan sebuah kurva roc pada gambar 8. Dimana curva tersebut mengepresikan *Confusion Matrix* dari sebuah tabel data set.



Gambar 8. Kurva Roc Naïve Bayes dan Genetika

Gambar 8. Merupakan kurva ROC yang memberikan penjelasan tentang hubungan uji sensitifitas dengan spesifisitas, pada kurva roc dapat diketahui terdapat 2 sumbu, yaitu sumbu x dan y, dengan persamaan y (TP) True Positive dan x (FP) False Positive, sedangkan warna biru menunjukkan ROC (Thershold) dan di nyatakan sebagai (TP) dikarenakan memiliki nilai 1 sedangkan warna merah disebut ROC dinyatakan sebagai (FP) karena memiliki titik 0, pada ROC (Thershold) memiliki titik 1 dan ROC memiliki titik 0 menunjukkan bahwa prediksi dinyatakan tidak benar (*false*), jika nilai titik antara ROC (*Thershold*) dan ROC (0,1) maka prediksi dinyatakan sempurna atau benar (*True*) baik kasus positif maupun negatif

## KESIMPULAN DAN SARAN

Berdasarkan dari hasil pengujian metode, didapatkan sebuah model untuk melakukan optimasi data dengan menggunakan algoritma Genetika terbukti dengan peningkatan hasil akurasi dari klasifikasi Penentuan Nasabah Bank yang akan Berdeposito dengan *feature selection*, dengan hasil nilai 86,71% menjadi 90,27%, Peningkatan akurasi yang dihasilkan tentunya belum sempurna, sehingga perlu adanya upaya-upaya lain dalam meningkatkan akurasi yang dihasilkan.

Terdapat beberapa saran untuk penelitian selanjutnya, yaitu sebagai berikut. Perlu dilakukan

eksperimen dengan menggunakan algoritma optimasi lain seperti ACO, PSO, dan yang lainnya. Pemilihan parameter untuk optimasi juga sangat berpengaruh terhadap tingkat akurasi yang dihasilkan, sehingga perlu adanya penetapan parameter yang sesuai dengan model yang diusulkan. Selain itu, terdapat optimasi peningkatan akurasi kembali dengan menggunakan metode lain yang sesuai, sehingga dimungkinkan untuk diterapkannya konsep model hybrid lain.

## DAFTAR PUSTAKA

- [1] C. Ou, C. Liu, J. Huang and N. Zhong, "On Data Mining for Direct Marketing," *Springer-Verlag Berlin*, p. PP. 491–498., 2003.
- [2] D. T. LAROSE, *Data Mining Methods and Models*, ISBN-13 978-0-471-66656-1, America: Department of Mathematical Sciences Central Connecticut State University, 2006.
- [3] D. C. P. Buani, "Optimasi Algoritma Naïve Bayes dengan Menggunakan Algoritma Genetika untuk Prediksi Kesuburan (Fertility)," *Evolusi*, vol. 4, no. 1, pp. 54 - 63, 2016.
- [4] R. Wati, "Penerapan Algoritma Genetika Untuk Seleksi Fitur Pada Analisis Sentimen Review Jasa Maskapai Penerbangan Menggunakan Naive Bayes," *Evolusi*, vol. 4, no. 1, pp. 25 - 31, 2016.
- [5] O. Sumantri and M. Kambali, "Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika," *JNTETI*, vol. 6, no. 3, pp. 301 - 3016, 2017.
- [6] Y. Wu, S. Huang, H. Ji, C. Zheng and C. Bai, "A Novel Bayes Defect Predictor Based on Information Diffusion Function," *Knowledge-Based Systems*, vol. 144, pp. 9-15, 2017.
- [7] S. A. Mostafa, A. Mustapha and M. a. Muhammad, "Examining multiple feature evaluation and classification methods," *Cognitive Systems Research*, vol. 54, pp. 90-99, 2019.
- [8] H. Muhammad, C. A. Prasoj, N. A. Sugianto, L. Surtiningsih and I. Cholissodin, "Optimasi Naïve Bayes Classifier Dengan Menggunakan Particle Swarm Optimization Pada Data Iris," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, vol. 4, no. 3, pp. 180-184, 2017.
- [9] A. D. Herlambang and S. H. Wijoyo, "Algoritma Naïve Bayes Untuk Klasifikasi Sumber Belajar Berbasis Teks Pada Mata Pelajaran Produktif Di Smk Rumpun Teknologi Informasi Dan Komunikasi," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, vol. 6, no. 4, pp. 431-436, 2019.
- [10] B. "Penerapan Algoritma Naïve Bayes Untuk Mengklasifikasi Data Nasabah Asuransi," *TECHSI: Jurnal Penelitian Teknik Informatika*, vol. 3, no. 2, pp. 127 - 146, 2013.
- [11] J. C. M. J. C. N. A. Zaidi, "Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting," *Journal of Machine Learning Research*, vol. 1, no. 14, pp. 1947 - 1988, 2013.
- [12] S. Kusuma dewi, *Artificial Intelligent*, Yogyakarta: Graha Ilmu, 2003.
- [13] Z. Zukhri, *Algoritma Genetika Metode Komputasi untuk Menyelesaikan Masalah Optimasi*, Yogyakarta: Andi Offset, 20014.

- [14] A. M. D. Anita, *Konsep Kecerdasan Buatan*, Yogyakarta: Andi Offset., 2006.
- [15] V. d. F. Trevino, "GALGO: an R package for multivariate variable selection using genetic algorithms," *Bioinformatics*, vol. 22, p. 1154–1156., 2006.