Optimasi *K-Nearest Neighbour* dengan Algoritma Genetika

Ibnu Dwi Lesmono[1], Ardian Dwi Praba[2]

Abstract— Data mining is a field that are merged from several fields which brings together techniques from machine learning, pattern recognition, statistical databases, and visualization for information retrieval problem recognition and large databases. Classification is the process of the invention of the model which describe and distinguish classes or the concept that aims to be used to predict the class of the object label kelasya is not yet known. Many algorithms that can be used in solving the problem of classification in data mining one is K Nearest neighbour (K-NN). K-NN is the algorithm that aims to find new patterns in the data with the use of existing data patterns with the new data. The k-NN algorithm is one of the most widely used in the classification of learning based on terawasi, but the K-NN has problems on the determination of the optimal K parameter resulting in lower accuracy. In this paper will be discussed how the influence of K-NN algorithm if combined with genetic algorithm for parameter optimization of K..

Intisari— Data mining adalah bidang yang digabung dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin pengenalan pola, statistik database, dan visualisasi untuk pengenalan permasalahan pengambilan informasi dan database yang besar. Klasifikasi adalah proses penemuan model yang menggambarkan dan membedakan kelas atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari object yang label kelasya belum diketahui. Banyak algoritma yang bisa digunakan dalam menyelesaikan masalah klasifikasi dalam data mining salah satunya adalah K Nearest neighbour (K-NN). K-NN adalah algoritma yang bertujuan untuk menemukan pola baru dalam data dengan menggunakan pola data yang sudah ada dengan data yang baru. K-NN adalah salah satu algoritma yang paling banyak digunakan dalam klasifikasi berdasarkan pembelajaran terawasi, namun K-NN memiliki masalah pada penentuan parameter K yang optimal sehingga menghasilkan akurasi yang rendah. Dalam makalah ini akan dibahas bagaimana pengaruh algoritma K-NN jika digabung dengan algoritma genetika untuk optimasi parameter K.

Kata Kunci: Klasifikasi, Optimasi, K Nearest Neighbour, Algoritma Genetika.

I. PENDAHULUAN

Data mining adalah istilah yang diciptakan untuk menggambarkan proses pengolahan data besar yang mempunyai hubungan dan pola yang menarik [4]. Dalam perkembanganya data mining sering digunakan dalam bidang ekonomi, statistik, peramalan asumsi bahwa pola dan data dapat dicari secara otomatis, divalidasi, dan digunakan untuk prediksi.

Klasifikasi adalah proses untuk menemukan model yang menjelaskan atau membedakan kelas data dengan tujuan dapat memperkirakan kelas dari suatu objek yang labelnya belum diketahui. Beberapa algoritma yang bisa digunakan untuk klasifikasi diantaranya adalah decision tree, RainForest, Naive Bayes, Neural Network, dan K Nearest Neighbour (K-NN) [8]. K-NN adalah algoritma yang baik ketika diterapkan pada dataset yang besar dan terbukti lebih stabil jika dibandingkan dengan algoritma lainya seperti decision tree dan neural network [10]. K-NN termasuk dalam kategori algoritma supervised learning yaitu bertujuan untuk menemukan pola baru dalam data dengan menggunakan pola data yang sudah ada dengan data yang baru [15]. Tujuan dari algoritma K-NN adalah untuk mengklasifikasi objek baru berdasarkan attribute dan training sample [3] dimana hasil sample uji yang baru akan diklasifikasikan berdasarkan mayoritas dari kategori yang ada pada K-NN. Algoritma K-NN menggunakan klasifikasi ketetanggan sebagai nilai prediksi dari sample uji data yang baru [15]. Cara kerja algoritma K-NN adalah dengan menentukan jarak pada pengujian data testing dengan data training berdasarkan nilai terkecil dari nilai ketetanggaan terdekat [7].

K Nearest Neighbour adalah algoritma yang efektif dan telah banyak digunakan untuk mengelompokan pola [13] dan termasuk dalam salah satu dari 10 algoritma terbaik dalam data mining [18]. K-NN merupakan algoritma klasifikasi yang paling popular dan terbukti lebih stabil jika dibanding dengan yang lainya seperti jaringan syaraf dan C4.5 [10]. K-NN merupakan algortma yang efektif dan sederhana namun mempunyai masalah pokok yaitu kompleksitas kesamaan sampelnya sangat besar [6] dan sulitnya menentukan nilai K yang optimal [17]

Beberapa peneliti telah melakukan pengembangan dengan menggabungan beberapa metode untuk meningkatkan optomasi parameter K pada algoritma K-NN diantara adalah penelitian yang dilakukan oleh Liaw, Leou, dan Wu [12] dengan menggabungkan algoritma K-NN dengan Orthogonal Search Tree (OST). Penelitian lain juga dilakuan untuk mengoptimalkan penentuan parameter K dengan metode Receiver Operating Characteristics (ROC) [5].

Pada penelitian kali ini akan dibahas tentang optimasi parameter K pada K-NN dengan menggunakan algoritma

[1]*Program Studi Manajemen Informatika, AMIK BSI Purwokerto, Jl.HR.Bunyamin No.106 Banyumas Indonesia (telp:0281-642848); e-mail: ibnu.idl@bsi.ac.id*

[2]*Program studi sistem informasi STMIK NUSA MANDIRI Jakarta, Jln. Jln Damai No.8 Warung Jati Barat Jakarta Selatan 12540 Indonesia (telp: 021-8839513);e-mail:* ardian.ddw@bsi.ac.id

genetika. Algoritma genetika adalah suatu metode pencarian yang sangat efektif untuk menyelesaikan permasalahan optimasi dengan mekanisme evolusi.

II. KAJIAN LITERATUR

A. Tinjauan Studi

Beberapa peneliti telah melakukan penelitian untuk menemukan nilai K yang optimamal pada algoritma K-NN dengan beberapa metode. Seperti penelitian yang dilakukan oleh Liaw dan kawan-kawanya [12] yang mencoba mengoptimalkan parameter K dengan metode *Orthogonal Search Tree* (OST). Dalam penelitian tersebut dataset yang digunakan adalah dataset public yaitu Statlog (*Lansat Satellite*) yang diambil dari UCI. Pertama simpul OST akan membuat *d* dan berisi semua titik data dari data yang ditetapkan. Selain itu, panjang sisa harus didefinisikan untuk setiap data titik dan digunakan dalam memeriksa apakah titik data dalam *node* daun mungkin menjadi K-NN dari titik permintaan dan dapat ditolak nilai awal proses. *Total computation time* pada penelitian ini lebih cepat dibandingkan dengan algoritma yang lain yaitu sebesar 441, sedangkan algoritma LB *Tree* 4594, dan *Ball Tree* 2475.

Penelitian yang dilakukan oleh (Laily dan Suciana, 2014) dengan menggunakan metode *Backward Elimination* untuk mengimprove algoritma *K Nearest Neighbour*. Dataset yang digunakan pada penelitian ini adalah dataset jantung yang diambil dari UCI *Mechine Learning*. *Algoritma Backward Eliminaton* digunakan untuk menghilangkan *atribute-atribute* yang tidak relevan didasarkan pada model *regresi linear*. Setelah melakukan pengujian dengan menggunakan dataset jantung, maka didapat hasil akurasi *K Nearest Neighbour* sebesar 88.62%, sedangkan untuk K *Nearest Neighbour* yang telah dioptimasi dengan *Backward Elimination* adalah sebesar 89.55%. sehingga bisa disimpulkan bahwa akurasi Algoritma K-NN yang diimprove dengan *Backward Elimination* lebih besar dibandingkan K-NN normal.

Penelitian yang lain dilakukan oleh (muis dan purwanto, 2015) yang mengimprove algoritma *K Nearest neighbour* dengan algoritma *forward selection*. Dataset yang digunakan dalam penelitian ini diambil dari badan pengawasan perdagangan berjangka komoditi. Data yang diperolah berupa data harian *time series univariate*. Pengujian dalam metode ini menggunakan *Root Mean Squared Error* (RMSE). Dengan menggunakan RMSE akan terlihat seberapa besar perbedaan hasil yang akan dihitung. Dari hasil penelitian yang dilakukan terilihat bahwa K-NN yang telah diimprove dengan forward seletion berhasil melakukan seleksi variable dengan lebih baik dengan angka RMSE 6328,376 sedangkan K-NN biasa nilainya adalah 7062,539.

B. Tinjuan Pustaka

1) Data Mining

Data mining merupakan extraksi untuk mendapatkan informasi penting yang sifatnya implisit yang sebelumnya tidak diketahui dari suatu data [16]. Data mining adalah proses pengambilan pengetahuan dari volume data yang besar yang disimpan dalam basis data atau informasi yang disimpan dalam *repository* [3]. Data mining adalah bidang yang digabung dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistic database, dan visualisasi untuk pengenalan permasalahan pengambilan informasi dari database yang sangat besar yang sangat [9].

2) Algoritma Klasifikasi

Algoritma klasifikasi adalah metode pembelajaran data untuk memprediksi nilai dari suatu atribut. Algoritma klasifikasi akan menghasilkan sekumpulan aturan yang akan digunakan sebagai indikator untuk memprediksi kelas dari data yang ingin diprediksi [11]. Menurut Han dan Kamber [3] klasifikasi adalah proses penemuan model yang menggambarkan dan membedakan kelas atau konsep yang bertujuan agar bisa untuk memprediksi kelas objek yang labelnya belum diketahui. Tujuan dari algoritma klasifikasi adalah untuk menemukan relasi antara beberapa variable yang tergolong dalam kelas yang sama. Relasi tersebut akan digambarkan dengan aturan-aturan agar dapat memprediski kelas dari data yang atributnya sudah diketahui.

Proses pembuatan model klasifikasi dikelompokan dalam tiga tahapan [11]

1. Tahap pembelajaran
   Pada tahap inu algoritma diterapkan kedalam data contoh untuk mendapatkan relas data dalam setiap kelasnya. Tahap ini akan membentuk model yang berisikan aturan-aturan atribut dalam menentukan kelas data.
2. Tahap pengujian
   Tahapan pengujian adalah tahap penerapan aturan-aturan yang sudah terbentuk pada tahapan pembelajaran. Dalam tahap ini, aturan model yang dimiliki akan diterapkan pada setiap atribut dalam data. Pengujian data dilihat kecocokan antara kelas yang diprediksi dengan kelas pada data sebenarnya.
3. Tahap prediksi
   Pada tahap prediksi model yang dihasilkan benar-benar diterapkan pada data yang belum diketahui kelasnya. Penilaian algoritma klasifikasi biasanya dilihat dari akurasi. Akurasi adalah ketepatan model dalam memprediksi kelas data.

Disamping proses pembuatan model klasifikasi diatas, menurut Gorunescu [2] ada empat komponen klasifikasi.

1. *Class*
   Variable dependen yang berupa katagorikal yang merepresentasikan label yang terdapat pada objek, missal: penyakt jantung, resiko kredit, jenis gempa.
2. *Predictor*
   Variable independen yang direpresentasikan oleh karakteristik (*attribute*) data. Misal: merekok, tekanan darah, gaji.

3. *Training dataset*
   Satu set data yang berisi nilai dari kedua komponen diatas yang digunakan untuk menentukan kelas yang cocok berdasarkan *predictor.*
4. *Testing dataset*
   Berisi data baru yang akan diklasifikasi oleh model yang telah dibuat dan akurasi klasifikasi akan dievaluasi.

3) K Nearest Neighbour

K Nearest Neighbour (K-NN) termasuk kelompok *instance-based learning.* Algoritma ini merupakan salah satu teknik *lazy learning*. Cara kerja algoritma K-NN dilakukan dengan mencari kelompok K objek dalam data *training* yang paling dekat (mirip) dengan objek pada data baru atau data *testing* [18]. Algoritma K-NN merupakan suatu metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembejalajaran yang jaraknya paling dekat dengan objek tersebut. K-NN merupakan salah satu metode pengklasifikasi data berdasarkan similaritas dengan label data [9]. Algoritma K-NN merupakan algoritma yang sangat efektif namun ketepatan algoritma ini sangat dipengaruhi oleh ada atau tidaknya fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi. Riset terhadap algoritma ini sebagian besar membahas bagaimana cara meningkatkan akurasi sehingga performa klasifikasi menjadi lebih baik.

Pada fase pembelajaran, algoritma ini hanya melakukan penyimpanan vector-vektor fitur dari klasifikasi dari data pembelajaran. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk data set (yang klasifikasinya tidak diketahui). Jarak dari vector yang baru ini terdapat seluruh vector data pembelajaran dihitung, dan sejumlah k buah yang paling dekat diambil. Titik yang baru klasifikasinya diprediksi termasuk pada klasifikasi terbanyak dari tititk-titik tersebut.

Nilai K yang terbaik untuk algoritma ini tergantung pada data secara umumnya, nilai k yang tinggi akan mengurangi efek *noise* pada klasifikasi, tetapi membuat batasan antara setiap klasifikasi menjadi lebih kabur. Nilai k yang bagus dapat dipilih dengan optimasi parameter, misalnya dengan menggunakan *cross-validation*. Kasus khusus dimana klasifikasi diprediksi berdasarkan data pembelajaran yang paling dekat (dengan kata lain, k=1) disebut algoritma *nearest neighbor*.

K-NN adalah metode yang melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. *Nearest Neighbour* adalah sebuah pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dan kasus lama yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada. Untuk mendefinisikan jarak antar dua titik pada data *training* (x) dan titik pada data testing (y) maka digunakan rumus *Euclidean.*

Penggunaann algoritma K-NN banyak digunakan untuk menangani masalah dalam bidang ilmiah dan rekayasa perangkat lunak seperti pengenalan pola, pengenalan objek, pengelompokan data dan lain-lain. Untuk menentukan jumlah data atau tetangga terdekat ditentukan oleh user yang dinyatakan dengan K.

$$dist\ (x1, x2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2} \qquad (1)$$

Penjelasan dari rumus diatas adalah $x1=(x_{11},x_{12}..x1_n)$ dan $x2=(x_{21},x_{22},.x_{2n})$. dengan kata lain, untuk setiap atribut numerik, kita mengambil perbedaan antara yang sesuai nilai-nilai atribut yang dalam vector x1 dan x2 dari matriks dengan ukuran dimensi. Akar kuadrat diambil dari akumulasi jumlah total jarak. Biasanya, kita menormalkan nilai masing-masing atribut sebelum menggunakanya. Prinsip kerja K-NN adalah mencari jarak terdekat antara data yang dievaluasi dengan k tetangga terdekatnya dalam data pelatihan. Persamaan perhitungan untuk mencari *euclidean* dengan d adalah jarak dan p adalah dimensi data.

$$d_i = \sqrt{\sum_{i=1}^{p}(x_{1i} - x_{2i})^2} \qquad (2)$$

Keterangan:
x1= sample data uji
x2= data uji
d = jarak
p = dimensi data

Pseudecode algoritma K-NN

***Input**: D, the set of training objects, the test object, **z**, which is a vector of attribute values, and L, the set of classes used to label the objects*
***Output**: $C_z \epsilon$ L, the class of □*
Foreach** object y ϵ D **do
*Compute d**(z,y)**, the distance between **z** and **y**;*
end
Select N $\subseteq$ D, the set (neighborhood) of k closest training objects for □;
$c_z = argmax \sum_{y \epsilon N} I\ (v = classs(c_y));$
where $I^{v \epsilon L}(.)$ is an indicator function the returns the value 1 if its argumnets is true and 0 otherwise

Antara z dan semua dataset pelatihan untuk menentukan daftar tetangga terdekat. Maka menetapkan kelas z dengan mengambil kelas mayoritas tetangga terdekat dari objek. Kompleksitas penyimpanan algoritma adalah 0(n), dimana n adalah jumlah pelatihan objek. Kompleksitas waktu juga 0(n), karena jarak perlu dihitung anatara target dan setiap objek pelatihan. Namun, ada waktu diambil untuk pembangunan klasifikasi model, misalnya, algoritma c 4.5. dengan demikian K-NN berbeda dari teknik klasifiasi lain yang membutuhkan banyak waktu untuk membuat sebuah model.

4) Algoritma Genetika

Algoritma genetika adalah salah satu diantara algoritma evolusi yang sangat popular dari segi keragaman aplikasi mereka. Sebagian besar masalah-masalah optimasi

yang terkenal telah dicoba oleh algoritma genetika. Selain itu, banyak algoritma-algoritma optimasi lain yang didasarkan pada algoritma genetika atau memiliki beberapa kemirian yang kuat [19]

Genetic algorithm (GA), dikembangkan oleh John Holland dan kolaboratornya di tahun 1960-an dan 1970-an, adalah model atau abstraksi evolusi biologi berdasarkan Charles Darwin tentang teori seleksi alam. *Operator genetic* ini membentuk bagian penting dari algoritma genetika sebagai strategi pemecah masalah. Banyak varian dari algoritma genetika telah dikembangkan dan diterapkan kedalam berbagai masalah optimasi seperti *Traveling Salesmen Problem* (TSP).

Konsep algoritma genetika sama dengan konsep sistem biologi. Algoritma genetika dimulai dari sebuah dataset solusi atau populasi. Setiap individu dipopulasi merepresentasikan solusi dari suatu permasalahan. Algoritma genetika merupakan gabungan atau kumpulan dari konsep evolusi alam, yang biasanya digunakan untuk inisialisasinya secara aturan acak pada populasi [4]

Dalam data mining, algoritma genetika biasa digunakan untuk melakukan evaluasi terhadap nilai *fitness* pada sebuah algoritma. Nilai *fitness* yang semakin besar akan menghasilkan optimasi yang lebih baik.

Algoritma genetika
Objective function $f(x)$, $x = (x_{1,\dots}x_1)^T$
Encode the solution into chromosomes (strings)
Define fitness F (eg, $F \propto f(x)$ *) for mazimization*
Generate the initial population
initialize the probabilities of crossover (p_c*) and mutation (*p_m*)*
while (t < max number of generations)
generate new solution by crossover and mutation
Mutate with a mutation probability p_m
Accept the new solution if their fitness increase
Select the current best for the next generation (elitism)
update t=t+1
end while
Decode the result and visualization

Algoritma genetika memeiliki tiga operator genetic utama, yaitu: *crossover*, *mutation*, dan *selection*. Ketiga operator tersebut memiliki peran yang masing-masing berbeda.

1. *Crossover*: bagian dari solusi yang lain dalam kromosom atau solusi representasi. Peran utamanya adalah untuk menyediakan pencampuran solusi dan konvergensi di dalam ruang
2. *Mutation*: adalah bagian dari salah satu solusi secara acak, yang meningkatkan keragaman populasi dan menyediakan mekanisme untuk mencari titik yang optimal.
3. *Selection of the fittest*: adalah pilihan yang paling kuat, menggunakan solusi dengan fitness yang paling tinggi untuk lolos pada generasi berikutnya. Hal tersebut terus dilakukan sampai mendapatkan pilihan solusi yang paling baik.

5) T-Test

T-test adalah membandingkan hubungan antar dua variable yaitu, variable respon dan variable prediktor [9]. T-Test bertujuan untuk mengetahui besarnya pengaruh masing-masing variable independen secara individual (parsial) terhadap variable independen. Hipotesis nol adalah hipotesis yang menyatakan tidak adanya pengaruh atau perbedaan antar dua variable, sedangkan hipotesis alternatif adalah hipotesis yang menyatakan adanya pengaruh atau perbedaan antara dua buah variable. Jika nilai P < 0,05 maka menunjukan adanya pengaruh atau perbedaan yang signifikan.

III. METODE PENELITIAN

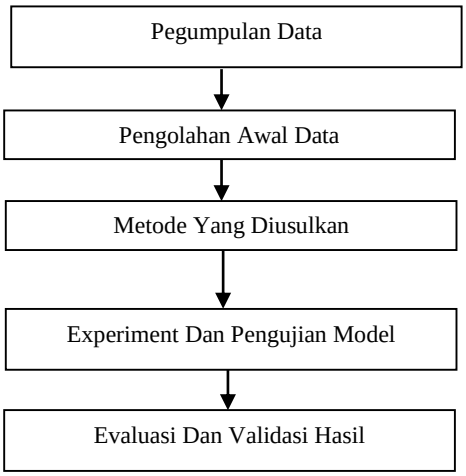
Penelitian merupakan kegiatan terencana yang bertujuan untuk meberikan kontribusi kepada pengetahuan. Metode yang digunakan dalam penelitian ini adalah menggunakan metode optimasi. Dari jenis optimasi yang digunakan yaitu algoritma genetika sebagai metode untuk optimasi penentuan K pada algoritma *K Nearest Neighbour* sehingga akurasinya bisa meningkat. Dalam penelitian ini dipilih algoritma K-NN karena termasuk algoritma yang efektif dan telah banyak digunakan untuk pengelompokan pola [13]. Penelitian merupakan kegiatan terencana yang bertujuan untuk memberikan kontribusi kepada pengetahuan [1]. Penelitian merupakan salah satu cara untuk mendapatkan realitas kebenaran.

A. Rancangan Penelitian

Metode penelitian yang akan dibahas dalam paper ini seperti pada Gambar 1 yang didalamnya berisi tahapan sebagai berikut:

1. Pengumpulan data
   Pengumpulan data pada penelitian ini diawali dengan mengumpulkan data. Data didapatkan dari dataset yang telah dihunakan oleh para peneliti lain (dataset publik).
2. Pengolahan awal data
   Pengolahan awal data meliputi penyaringan data kedalam bentuk yang dibutuhkan, serta pengelompoka dan penentuan atribut data.
3. Metode yang diusuklkan
   Mengajukan usulan metode yang akan digunakan. Metode yang diusulkan dalam penelitian ini adalah algoritma genetika untuk optimasi parameter K pada K-NN.
4. Experiment dan pengujian model
   Untuk melakukan pengujian model, dilakukan dengan menggunakan Rapid Miner. Dari algoritma yang telah ditentukan maka dataset akan diolah sehingga menghasilkan model yang diharapkan.
5. Evaluasi dan validasi hasil

Setelah melakukan experiment terhadap semua dataset dengan model yang diusulkan, maka hasil akan dievaluasi dan ditarik kesimpulan dari penelitian ini.



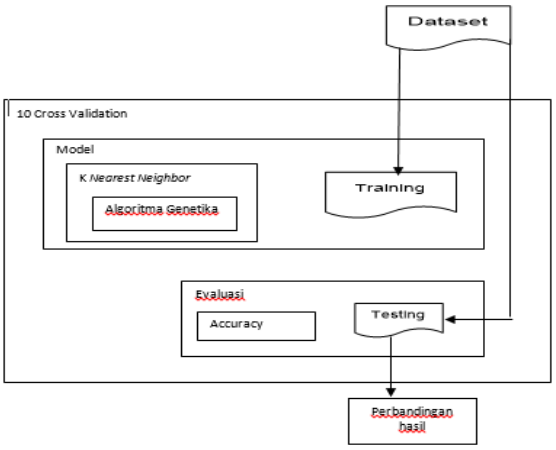
Gbr. 1 Metode Penelitian

B. Pengumpulan Data

Dataset yang akan digunakan dalam penelitian ini ada 3 jenis dan semuanya termasil dalam dataset public yaitu Tic-Tac-Toe Endgame, iris dan Image Segmentation yang diambil atau didownload langsung dari UCI *Repository*. Alasan penelitian ini mengambil dataset public agar peneliti lain yang akan melakukan pengujian bisa mengambil dataset yang sama yang ada pada alamat https://archive.ics.uci.edu/ml/datasets.html. mayoritas penelitian saat ini cenderung menggunakan dataset publik untuk menguji metode yang akan diusulkan dalam sebuah penelitian.

C. Metode yang diusulkan

Metode yang diusulkan dalam penelitian ini adalah penerapan algoritma genetika untuk optimasi parameter K pada algoritma K Nearest Neighbour. Seperti pada Gambar 3.2 untuk *proposed method* menggunakan algoritma genetika untuk optimasi parameter K dan untuk pengujian akurasi menggunakan 10 *Cross Fold Validation*.



Gbr .2 Metode Usulan

Metode yang diusulkan diawali dengan membagi dataset menjadi data training dan data testing dengan menggunakan 10 *fold cross validation*, yaitu dengan membagi data menjadi dua bagian yaitu 90% untuk proses *training* dan 10% untuk proses *testing*. Data dilatih dan diuji menggunakan algoritma *K Nearest Neighbour* dengan parameter K yang telah dioptimalkan dengan algoritma genetika. Setelah melakukan training dan testing, K-NN juga menghitung nilai akurasi dimana semakin besar presentasi akurasinya maka semakin bagus tingkat klasifikasinya.

D. Experiment dan pengujian model

Tahapan dalam penelitian ini adalah sebagai berikut:

1. Menyiapkan dataset untuk experiment
2. Melakukan validasi dengan menggunakan *x-validation*
3. Melakukan uji data *training* dan *testing* terhadap K-NN dengan optimasi K dengan menggunakan algoritma genetika.
4. Melakukan pencatatan akurasi tertinggi pada algoritma K-NN dan K-NN yang telah dioptimasi menggunakan algoritma genetika.

Untuk melakukan penelitian ini diperlukan eksperiment dan proses pengujian model yang diusulkan. Proses experiment dan pengujian model menggunakan tiga dataset yang diambil dari UCI. Semua dataset akan diuji dengan algoritma yang ada di rapidminer. Spesifikasi komputer ang digunakan dalam penelitian ini dapat dilihat pada Tabel .I

TABEL .I
SPESIFIKASI KOMPUTER

Processor	Intel Core i3-3217U 1.80 GHz
RAM	2 GB
Hardisk	500 GB
Sistem Operasi	Windows 10
Aplikasi	Rapid Miner

a. Evaluasi Dan Validasi

Dalam data mining *cross validation* adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi [12]. Dengan menggunakan *cross validation* akan dilakukan percobaan sebanyak k. data yang digunakan dalam percobaan ini adalah data *training* untuk mencari nilai akurasi secara keseluruhan.

Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi. Dalam penelitian ini nilai k yang digunakan berjumlah 10 atau 10-*fold cross validation*, tiap percobaan akan menggunakan satu data *testing* dan k-1 bagian akan menjadi data *training*, kemudian data *testing* itu akan ditukar dengan satu buah data *training* sehingga untuk tiap percobaan akan didapatkan data *testing* yang berbeda-beda.

Adapun model yang dihasilkan akan diuji dengan menggunakan *cross validation* untuk mengetahui tingkat akurasi, *cross validation* digunakan untuk menghindari *overlapping* pada data *testing*. Tahapan *cross validation*

terdiri dari proses pembagian data menjadi k subset yang berukuran sama kemudian menggunakan setiap subset untuk data testing dan sisanya untuk data *training*. Semakin tinggi nilai akurasi, semakin baik pula model yang dihasilkan.

Model validasi yang digunakan pada penelitian ini adalah 10 *fold cross validation.* 10 *fold cross validation* digunakan untuk mengukur kinerja model prediksi. Setiap dataset secara acak dibagi menjadi 10 bagian dengan ukuran yang sama. Selama 10 kali, 9 bagian untuk melatih model (data *training*) dan 1 bagian digunakan untuk menguji (data *testing*) yang lainya setiap kali dilakukan pengujian. Pegukuran pada evaluasi kinerja klasifikasi bertujuan untuk mengetahui seberapa akurat model klasifikasi dalam prediksi kelas dari suatu baris data [3]. Istilah yang biasa digunakan adalah *true positive* (TP) adalah jumlah data yang diprediksi *positive* dan sesuai dengan kenyataan, *true negative* (TN) adalah jumlah data yang diprediksi *negative* dan sesuai dengan kenyataan. *False positive* (FP) adalah jumlah data yang diprediksi positive namun tidak sesuai dengan kenyataan, *false negative* (FN) adalah jumlah data yang diprediksi *negative* namun tidak sesuai dengan kenyataanya.

Rumus untuk menghitung *accurascy* adalah

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (3)$$

Disamping mengukur akurasi kita juga mengukur hasil experiment dengan *Area Under Curve* (AUC) untuk mengukur hasil akurasi indikator dari performa model klasifikasi. Hasil akurasi dapat dilihat dengan melakukan perbandingan klasifikasi menggunaakan curva *Receiver Operating Characteristic* (ROC) dari hasil *confusion matrix.* ROC menghasilkan dua garis dengan bentuk true positif yang ditandai dengan garis vertical dan false positive yang ditandai dengan garis horiozontal [11]. ROC adalah grafik antara sensitivitas true positive rate pada sumbu X dan sumbu Y.

TABEL II
CONFUSION MATRIX

Class		**Actual**	
		True	False
Predic	True	True Positif(TP)	False Negative (FN)
	False	False Positive (FP)	True negative (TN)

Dalam klasifikasi untuk mengukur keakuratan dari tes diagnostic menggunakan *area under curve* (AUC). Semakin tinggi nilai AUC makan hasilnya semakin baik. Nilai maksimal dari AUC adalah 1,00. Keteranga dan symbol AUC bisa dilihat pada table 3.4 dibawah ini.

TABEL III
KETERANGAN DAN SIMBOL AUC

Nilai AUC	**Klasifikasi**	**Simbol**
0,90 – 1,00	Excellent Clasification	
080, - 0,90	Good Classification	
0,70 – 0,80	Fair Classification	
0,60 – 0,70	Poor Classification	
< 0,60	Failure	

IV. HASIL DAN PEMBAHASAN

Dalam pembahasan ini akan ditampilkan hasil pengujian yang dilakukan untuk megetahui performa algoritma *K Nearest Neighbor* jika dioptimasi dengan algoritma genetika. Experiment ini dimulai dengan menyiapkan dataset yang akan digunakan yaitu Image Segmentation, iris, dan Tic-Tac-Toe Endgame.

TABEL IV
DATASET UCI *MACHINE LEARNING REPOSITORI*

Dataset	Jumlah atribut	Jumlah record	Jumlah kelas
Iris	5	150	3
Image Segmentation	20	210	7
Tic-Tac-Toe Endgame	10	958	2

Selanjutnya dataset yang tersedia diterapkan pada algoritma *K Nearest Neighbour*. Pada pengujian ini akan terlihat hasil dari algoritma K Nearest Neighbour yang belum dioptimasi dengan algoritma genetika. Hasil pengujian bisa dilihat pada tabel 5 di bawah ini.

TABEL V
HASIL PENGUJIAN DATASET MENGGUNAKAN K-NN

Metode	**Dataset**	**Akurasi**
K-NN	Iris	97,00%
K-NN	Image Segmentation	82,38%
K-NN	Tic Tac To	84,86%

Pada Tabel V terlihat akurasi paling tinggi terdapat pada dataset iris dengan tingkat akurasi mencapai 97,00%, sedangkan akurasi paling rendah ada pada dataset Image Segmentation dengan akurasi 82,38%. Salah satu penyebab akurasi yang rendah pada algoritma K Nearest Neighbour adalah penentuan K yang kurang optimal sehingga akurasi menjadi rendah.

Pada pengujian kedua algoritma K Nearest Nearest neighbor akan diimprove dengan algorima genetika yang berfungsi untuk mengoptimasi parameter k sehingga bisa

meningkatkan akurasi pada K Nearest Neighbour. Hasil pengujian kedua yang menggabungkan antara K Nearest Neighbour dengan algoritma genetikabisa dilihat pada tabel VI

TABEL VI.
HASIL PENGUJIAN DATASET MENGUNAKAN ALGORITMA K-NN DAN GA

Metode	Dataset	Akurasi
K-NN dan GA	Iris	99,00%
K-NN dan GA	Image Segementation	92,86%
K-NN dan GA	Tic Tac To	87,36%

Dari data yang terdapat pada tabel 6 terlihat bahwa semua akurasi dari masing-masing dataset mengalami kenaikan nilai akurasi setelah algoritma genetika digunakan untuk optimasi pada K *Nearest Neighbour*. Peningkata akurasi palig besar terjadi pada dataset image segmentation yang awalnya hanya 82,38% naik sebesar 10% menjadi 92,86%.

Ada beberapa tahapan yang ada pada algoritma genetika dalam mengoptimalkan parameter K pada algoritma K-NN. Pada tahap peratama adalah pengkodean yaitu proses kodefikasi atas solusi dari permasalahanya. Hasil dari pengkodean berbentuk string yang merupakan representasi dari suatu kromosom. Tahap kedua adalah selection yaitu menentukan kromosom mana yang tetap tinggal pada generasi berikutnya. Tahap ketiga crossover yang akan menghasilkan kromosom baru yang menggantikan kromosam lama.tahap ketiga adalah mutation yang memungkinkan terjadinya kromosm baru secara unpredictable. Proses terakhir adalah decoding yaitu mengambil hasil kromosom terbaik untuk memberikan nilai yang optimal.

Untuk melihat tingkat perbedaan akurasi dalam penelitian ini akan digunakan t-Test. Metode ini adalah metode yang paling umum dala statistik tradisional (Maletic & Marcus, 2005). Untuk menjamin hasil penelitian ini, dalam menguji hubungan antara pengguna metode K-NN dengan K-NN dan GA. Ada atau tidaknya perbedaan antara keduanya membutuhkan pengujian, salah satunya menggunakan t-Test [9].

Tabel VII.
Paired Two-tailed t-Test dengan menggunakan K-NN dan K-NN GA dan dataset iris

	84,86	97,33
Mean	80,13833333	98,162
Variance	9,710056667	0,8575
Observations	6	6
Pearson Correlation	-0,119195952	
Hypothesized Mean Difference	0	
df	5	
t Stat	-13,1592316	
P(T<=t) one-tail	2,26266E-05	
t Critical one-tail	2,015048373	
P(T<=t) two-tail	4,52531E-05	
t Critical two-tail	2,570581836	

TABEL VII
PAIRED TWO-TAILED T-TEST DENGAN MENGGUNAKAN K-NN GA DAN K-NN PSO DAN DATASET IRIS

	97,33	*96,67*
Mean	98,12333333	97,11
Variance	0,825866667	0,11616
Observations	6	6
Pearson Correlation	-0,46595422	
Hypothesized Mean Difference	0	
df	5	
t Stat	2,237472649	
P(T<=t) one-tail	0,037726819	
t Critical one-tail	2,015048373	
P(T<=t) two-tail	0,075453638	
t Critical two-tail	2,570581836	

Pada Tabel VIII menunjukan hasil t-Test pada algoritma K-NN dan K-NN GA dengan menggunakan dataset iris menunjukan hipotesis nol ditolak (hipotesis alternatif) dengann nilai P <0,05 yaitu 2,26266E-05. Pada table 4.3 hasil t-Test pada algoritma K-NN GA dan K-NN PSO dengan dataset iris menunjukan hipotesis nol ditolak (alternatif) yaitu dengan nilai P < 0,05 yaitu0,037726819.

Hasil t-Test dengan hopitesis nol ditolak (hipotesis alternatif) tersebut menunjukan bahwa antara penggunaan metode K-NN dan K-NN GA menunjukan adanya pengaruh dan perbedaan yang signifikan. Metode K-NN GA menghasilkan akurasi yang lebih baik dibandingkan dengan metode K-NN yang biasa. Hal ini berarti bahwa algoritma genetika yang diusulkan bisa menjadi alat yang efektif untuk meningkatkan kinerja algoritma K-NN.

V. KESIMPULAN

Penelitian dengan menggunakan algoritma genetika untuk mengoptimalkan penentuan parameter K pada algoritma *K Nearest neighbor* sehingga akurasinya meningkat. Hasil penelitian mendapatkan nilai akurasi tertinggi pada dataset iris dengan akurasi sebesar 99.00%. Secara umum semua akurasi pada dataset menjadi meningkat ketika parameter K pada *K Nearest Neighbour* dioptimasi dengan algoritma genetika.

Dari hasil penelitian ini bisa disimpulkan bahwa algoritma genetika mampu menaikan akurasi pada *K Nearest Neighbour* dengan mengoptimalkan parameter k. Namun algoritma genetika membutuhkan waktu yang lama dalam menentukan nilai k yang optimal sehingga proses klasifikasi berjalan lambat. Untuk penelitian kedepan disarankan antara lain , 1) mengatasi masalah waktu yang lama pada saat proses klasifikasi, 2) mengatasi masalah *noisy* data yang bisa mengakibatkan akurasi menurun, 3) mencoba algorima lain untuk melakukan optimasi sehingga hasil yang dicapai bisa leih baik dari penelitian-penelitian yang ada.

REFERENSI

[1] Dawson, C. W. (2009). *Projects in Computing and Information Systems*.

[2] Gorunescu, F. (2011). *Data mining: concepts and techniques*. *Chemistry &* http://doi.org/10.1007/978-3-642-19721-5

[3] Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. *Soft Computing* (Vol. 54). http://doi.org/10.1007/978-3-642-19721-5

[4] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*.

[5] Hassan, R., Hossain, M. M., & Bailey, J. (2008). Improving k -Nearest Neighbour Classification with Distance Functions Based on Receiver Operating Characteristics, 489–504.

[6] Jiang, D., Pei, J., & Li, H. (2013). Mining search and browse logs for web search: A Survey. *ACM Trans. Intell. Syst. Technol.*, *4*(4), 1–37. http://doi.org/10.1145/2508037.2508038

[7] Jianhong Wu, Chaokun Ma, G. G. (2014). *data clustering theory, algorithms, and apllications*. *Igarss 2014*. http://doi.org/10.1007/s13398-014-0173-7.2

[8] Larose, D. T. (2004). *Discovering Knowledge in Data*. *Statistics* (Vol. 1st). http://doi.org/10.1002/0471687545

[9] Larose, D. T. (2006). *Data Mining Methods \& Models*. *Spring*. http://doi.org/10.1002/0471756482

[10] Lei, Y., & Zuo, M. J. (2009). Gear crack level identification based on weighted K nearest neighbor classification algorithm. *Mechanical Systems and Signal Processing*, *23*(5), 1535–1547. http://doi.org/10.1016/j.ymssp.2009.01.009

[11] Lewandowski, C. M. (2009). *Data mining and optomization for decision making*. *Business Intelligence* (Vol. 1). http://doi.org/10.1017/CBO9781107415324.004

[12] Liaw, Y., Leou, M., & Wu, C. (2010). Fast exact k nearest neighbors search using an orthogonal search tree, *43*, 2351–2358. http://doi.org/10.1016/j.patcog.2010.01.003

[13] Liu, H., & Zhang, S. (2012). Noisy data elimination using mutual k-nearest neighbor for classification mining. *Journal of Systems and Software*, *85*(5), 1067–1074. http://doi.org/10.1016/j.jss.2011.12.019

[14] Maletic, J. I., & Marcus, A. (2005). Data Cleansing Data Mining and Knowledge Discovery Handbook. Data Mining and Knowledge Discovery Handbook. http://doi.org/10.1007/0-387-25465-x_2

[15] Mitsa, T. (2010). *Temporal Data Mining*. *Clinics in Laboratory Medicine* (Vol. 28). http://doi.org/10.2165/11537630-000000000-00000

[16] Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining*. http://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

[17] Won, J., & Friel, N. (2015). Ef fi cient model selection for probabilistic K nearest neighbour classi fi cation, *149*, 1098–1108.

[18] Wu, X., & Kumar, V. (2009). *The Top Ten Algorithms in Data Mining*. Retrieved from http://books.google.com/books?hl=en&lr=&id=_kcEn-c9kYAC&pgis=1

[19] Yang, X.-S. (2014). Nature-Inspired Optimization Algorithms. Nature-Inspired Optimization Algorithms. http://doi.org/10.1016/B978-0-12-416743-8.00007-5

Ibnu Dwi Lesmono, M. Kom, Lulus Pasca Sarjana Magister Ilmu Komputer Pada Tahun 2013 Konsentrasi E-Bussines STMIK Nusa Mandiri.

Ardian Dwi Praba, M.Kom , Lulus Pasca Sarjana Magister Ilmu Komputer Pada Tahun 2015 Konsentrasi Manajemen Sistem Informasi STMIK Nusa Mandiri.