ANALISIS PERBANDINGAN ALGORITMA NAIVE BAYES DAN KNN UNTUK KLASIFIKASI MULTI DATASET

[1]Eka bagus susanto (1410651097)
[2] Triawan adi cahyanto, M.Kom
[3]Reni umilasari S,pd Msi
Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember
Email : eekabagus@gmail.com

ABSTRAK

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia.Pada penelitian yang saya lakukan yang berjudul " Analisis Perbandingan Algoritma Naive Bayes Dan K-Nearest Neighbor Untuk Klasifikasi Multi Data Set "dengan membandingkan kedua metode antara Naive Baye dan K-Nearest Neighbor menggunakan 2 data set dengan varian jumlah data record dan atribut yang berbeda menunjukkan hasil rata-rata akurasi untuk Algoritma K-Nearest Neighbor pada pengujian K2, k3, k4 ,k5 ,k6 ,k7 ,k8 ,k9 ,k10 dan UTS yaitu 93,17% dan untuk naive bayes yaitu 78,38% dari rata rata akurasi pengujian tersebut Knn lebih unggul dari naive bayes.

Kata kunci : *Algoritma Naïve Bayes, k-Nearest Neighbor, Analisis, Klasifikasi, Perbandingan algoritma.*

BAB I PENDAHULUAN

1.1 Latar Belakang

Klasifikasi merupakan proses untuk menemukan model atau fungsi yang menjelaskan ataumembedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objekyang labelnya tidak diketahui. Model itu sendiri bisa berupa aturan "jika-maka", berupa *decision tree,formula matematis atau neural network*. Metode-metode klasifikasiantaralainC4.5,*RainForest,NaïveBayes,neuralnet work, genetic algorithm, fuzzy, case-based reasoning, dan k-NearestNeighbor* (Arriawati A S,2011).Dari beberapa metode klasifikasi tersebut, terdapat metode yang memiliki tingkat akurasitinggi yaitu *Naïve Bayes*. Hal tersebut diungkapkan olehHan and Kamber (2006) klasifikasi *Naïve Bayes* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam basis data dengan jumlah yang besar.Penelitian yang dilakukan oleh Luki Dwi Arisandi (2017) berjudul"*Perbandingan Algoritma Naïve Bayes dan C.45 Untuk klasifikasi*

multi data" Naïve Bayes menghasilkan akurasi 79,39%.

Selain *Naïve Bayes, algoritma K*-Nearest Neighbor atau dapat disingkat dengan K-NN adalah salah satu metode non parametrik yang digunakan dalam pengklasifikasian. Pada penelitian sebelumnya yang dilakukan oleh Nur Khotimah dan Deden Istiawan (2018) berjudul "Perbandingan Algoritma C4.5, Naïve Bayes dan K-Nearest Neighbour untuk Prediksi Lahan Kritis di Kabupaten Pemalang" Dari hasil perhitungan pengujian akurasi algoritma K-NN mampu menghasilkan akurasi sebesar 73,91% Dari kedua penelitian tersebut

menunjukkan bahwa algoritma Naïve Bayes dan algoritma K-NN memiliki tingkat akurasi yang tinggi pada proses klasifikasi data. Namun bagaimana jika kedua algoritma tersebut dibandingkan dengan menggunakan data yang sama. Bagaimana akurasi yang akan dihasilkan? Apakah algoritma Naïve Bayes lebih unggul dengan jumlah data yang banyak dibandingkan algoritma K-NN, karena K-Nearest Neighbour atau Naive bayes merupakan algoritma data mining yang sama sama baik dalam mengklasifikasi dan regesi data.

Oleh karena itu peneliti ingin melakukan sebuah penelitian dengan membandingkan algoritma Naïve Bayes dan algoritma K-NN menggunakan multi data yang peneliti ambil dari penyedia layanan dataset publik UCI (University of California, Irvine), dengan kriteria masing-masing data yang berbeda meliputi variabel,tipe data dan jumlah data, Hal ini bertujuan untuk menganalisa hasil akurasi dari kedua metode tersebut terhadap data yang diujikan.Berdasarkan pembahasan diatas peneliti mengambiljudul "***Analisis Perbandingan Algoritma Naïve Bayes dan K-NN untuk Klasifikasi Multi Data Set***".

BAB II TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Sebelum melakuan sebuah penelitian penulis terlebih dahulu melakukan tinjauan pustaka dari peneliti terdahulu yang berkaitan dengan algoritma Naive bayes dan KNN. Berikut adalah tabel beberapa penelitian yang terkait dengan masalah tersebut :

1. **Penelitian oleh (Nur Khotimah dan Deden Istiawan) Tahun 2018**

Penelitian yang pertama adalah penelitian yang dilakukan oleh Nur Khotimah dan Deden Istiawan (2018) berjudul "Perbandingan Algoritma C4.5, Naïve Bayes dan K-Nearest Neighbour untuk Prediksi Lahan Kritis di Kabupaten Pemalang" Dari hasil perhitungan pengujian akurasi algoritma K-NN mampu menghasilkan akurasi sebesar 73,91%

2. **Penelitian oleh (Puji Astuti) 2016**

Penelitian yang kedua adalah penelitian yang dilakukan .Penelitian yang dilakukan oleh Luki Dwi Arisandi (2017) berjudul"*Perbandingan Algoritma Naïve Bayes dan C.45 Untuk klasifikasi multi data*" Naïve Bayes menghasilkan akurasi 79,39%.

2.2 Klasifikasi

Klasifikasi merukapan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi ada dua pekerjaan utama yang dilakukan, yaitu (1) pembangunan model sebagai prototipe untuk disimpan sebagai memori dan (2) penggunaan modeltersebut untuk melakukan pengenalan/klasifikasi/prediksi pada suatu objek data lain agar diketahui kelas mana objek data tersebut dalam model yang sudah disimpannya.

Contoh aplikasi yang sering ditemui adalah pengkalsifikasian jenis hewan, yang mempunyai sejumlah atribut. Dengan atribut tersebut, jika ada hewan baru, kelas hewannya bisa langsung diketahui. Contoh lain adalah bagaimana melakukan diagnosis penyakit kulit kanker melanoma (Amaliyah et al, 2011), yaitu dengan melakukan pembangunan model berdasarkan data latih yang ada, kemudian menggunakan model tersebt untuk mengidentifikasi penyakit pasien baru sehingga diketahui apakah pasien tersebut menderita kanker atau tidak.

2.2.1 Pengkuran Kinerja Klasifikasi

Sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua set data dengan benar, tetapi tidak dapat dimungkiri bahwa kinerja suatu sistem tidak bisa 100% benar sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi (confusion matrix).

Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Tabel 2.2 merupakan contoh matriks konfusi yang melakukan klasifikasi masalah biner (dua kelas), hanya ada dua kelas, yaitu kelas 0 dan 1. Setiap sel fij dalam matriks menyatakan jumlah rekord/data dari kelas i yang hasil prediksinya masuk ke kelas j. Misalnya, sel f11 adalah jumlah data dalam kelas 1 yang secara benar dipetakan ke kelas 1, dan f10 adalah data dalam kelas 1 yang dipetakan secara salah ke kelas 0.

Tabel 2.1 Matriks konfusi untuk klasifikasi dua kelas

F_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 0
Kelas asli (i)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Keterangan :

f_{11}	Data dalam kelas 1 yang secara benar dipetakan ke kelas 1
f_{10}	Data dalam kelas 1 yang dipetakan secara salah ke kelas 0
f_{01}	Data dalam kelas 0 yang dipetakan secara salah ke kelas 1
f_{00}	Data dalam kelas 0 yang secara benar dipetakan ke kelas 0

Untuk menghitung akurasi digunakan formula

$$\text{Akurasi} = \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{jumlah prediksi yang dilakukan}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Untuk menghitung laju eror (kesalahan prediksi) digunakan formula

$$\text{Akurasi} = \frac{\text{Jumlah data yang diprediksi secara salah}}{\text{jumlah prediksi yang dilakukan}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Semua algoritma klasifikasi berusaha membentuk model yang mempunyai akurasi tinggi (laju eror yang rendah). Umumnya, model yang dibangun dapat memprediksi dengan benar pada semua data yang menjadi data latihnya, tetapi ketika model berhadapan dengan data uji, barulah kinerja model dari sebuah algoritma klasifikasi ditentkan.

2.3 Algoritma Naive Bayes

Bayes merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi (ketidaktergantungan) yang kuat (naif). Dengan kata lain, dalam Naive Bayes, model yang digunakan adalah "model fitur independen". Dalam Bayes (terutama Naive Bayes), maksud independensi yang kuat pada fitur adalah bahwa sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Contohnya, pada kasus klasifikasi hewan dengan fitur penutup kulit, melahirkan, berat, dan menyusui. Dalam dunia nyata, hewan yang berkembang biak dengan cara melahirkan dipastikan juga menyusui. Di sini ada ketergantungan pada fitur menyusui karena hewan yang menyusui biasanya melahirkan, atau hewan yang bertelur biasanya tidak menyusui. Dalam Bayes, hal tersebut tidak dipandang sehingga masing-masing fitur seolah tidak memiliki hubungan apa pun.prediksi Bayes didasarkan pada teorema Bayes dengan formula umum sebagai berikut:

$$P(H|E) = \frac{P(E|H)\ x\ P(H)}{P(E)} \text{.....(2.1)}$$

Penjelasan dari formula tersebut adalah sebagai berikut.

Parameter	Keterangan
P(HIE)	Probabilitas akhir bersyarat (conditional probability) suatu hipotesis H terjadi jika diberikan bukti (evidence) E terjadi.
P(ElH)	Probabilitas sebuah bukti E terjadi akan memengaruhi hipotesis H.
P(H)	Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apa pun
P(E)	Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain

Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati. Ada beberapa hal penting dari aturan Bayes tersebut, yaitu

1. Sebuah probabilitas awal/priori H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(HIE) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Contoh, dalam suatu peramalan cuaca untuk memperkirakan terjadinya hujan, ada faktor yang memengaruhi terjadinya hujan, yaitu mendung. jika diterapkan dalam Naive Bayes, probabilitas terjadinya hujan, jika bukti mendung sudah diamati, dinyatakan dengan

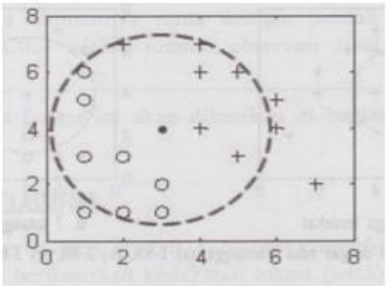
$$\mathrm{P(Hujan|Mendung)} = \frac{\mathrm{P(Mendung|Hujan)\ x\ P(Hujan)}}{\mathrm{P(Mendung)}}$$

P(HujanlMendung) adalah nilai probabilitas hipotesis hujan terjadi jika bukti mendung sudah diamanP(Mendung|Hujan) adalah probabilitas bahwa mendung yang diamati akan meme-v ngaruhi terjadinya hujan. P(Hujan) adalah probabilitas awal hujan tanpa memandang bukti apa pun, sementara P(Mendung) adalah probabilitas terjadinya mendung.

2.4 Algoritma K-Nearest –Neighbor

Algoritma prediksi K-NN diberikan pada 4.1 pada Algoritma tersebut ada sebuah data uji = (x',y') ,dimana x' adalah vektor/atribut data uji,sedangkan y' adalah label kelas data uji yang belum diketahui hitung jarak (atau kemiripan) data uji ke setiap data latih d(x',x),kemudian ambil K-tetangga terdekat pertama dalam Dz.setelah itu ,hitung jumlah data uang mengikuti kelas yang ada dai K-tetangga tersebut kelas dengan data terbanyak yang mengikutinya menjadi kelas pemenang yang diberikan sebagai label kelas pada data uji y'.

Salah satu masalah yang dihadapi K-NN adalah pemilihan nilai K yang tepat .cara voting mayoritas dari K-tetangga untuk nilai K yang besar bisa mengakibatkan distorsi data yang besar,seperti yang ditunjukan pada gambar 4.3 misalnya ,diambil K bernilai 13, pada Gambar 4.3, kelas 0 dimiliki oleh 7 tetangga yang jauh ,sedangkan kelas 1 dimiliki 6 tetangga yang lebih dekat. Hal ini mengakibatkan data uji tersebut akan terdistorsi sehingga ikut bergabung dengan kelas 0.Hal ini karena setiap tetangga tersebut mempunyai bobot yang sama terhadap data uji.sedangkan K yang terlalu kecil bisa menyebabkan algoritma terlalu sensitif terhadap noise(noise)



Gambar 2.1 K-Nearest Neighbor Dengan Nilai K yang besar

Untuk menangani masalah voting mayoritas tersebut biasanya ditambahkan pengguna bobotuntuk menghitung kandidat kelas yang sebaiknya diambil oleh data uji dari K –tetangga terdekat bobot dari setiap tetangga terdekat dihitung dengan formula

$$w_i = \frac{1}{d(x', x_i)}(41)$$

Formula yang bisa digunakan adalah

$$y^i = \arg max \sum(x_i, y_2) \in D_z w_i\ X\ (V = y_i)(4.2)$$

v merupakan umlah data yang masuk dalam kelas y_i. MATLAB menyediakan fungsi untuk melakukan klasifikasi dengan K-NN .sintaksisnya adalah

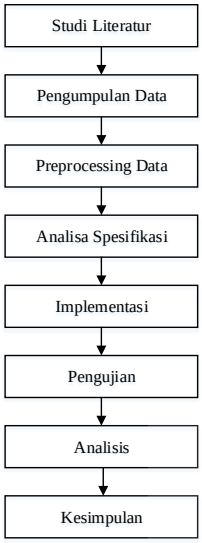
Class = knnclassify (sample,training,Group,k,distance ,rule)

2.5. Tool Weka Dalam Klasifikasi

Weka adalah aplikasi data mining open source berbasis Java. Aplikasi ini dikembangkan pertama kali oleh Universitas Waikato di Selandia Baru sebelum menjadi bagian dari Pentaho. Weka terdiri dari koleksi algoritma machine learning yang dapat digunakan untuk melakukan generalisasi /formulasi dari sekumpulan data sampling. Walaupun kekuatan Weka terletak pada algoritma yang makin lengkap dan canggih, kesuksesan data mining tetap terletak pada faktor pengetahuan manusia implementornya. Tugas pengumpulan data yang berkualitas tinggi dan pengetahuan pemodelan dan penggunaan algoritma yang tepat diperlukan untuk menjamin keakuratan formulasi yang diharapkan.

BAB III METODOLOGI PENELITIAN

Dalam mengerjakan Tugas Akhir ini diperlukan suatu tahapan penelitian untuk mendapatkan hasil yang maksimal. Untuk itu peneliti merencanakan suatu langkah-langkah yang dapat memaksimalkan dalam pengerjaan Tugas Akhir ini. Langkah-langkah tersebut adalah sebagai berikut :



Gambar 3.1 Tahap Penelitian

3.1 Studi Literatur

Dalam tahap ini peneliti mempelajari tentang semua data dan informasi yang berkaitan dengan algoritma Naive Bayes dan juga semua materi yang berhubungan dengan masalah yang akan dibahas, dalam penelitian ini refrensi diambil dari berbagai sumber, seperti buku, jurnal, e-book, serta sumber-sumber lain yang dinilai dapat memberi tambahan wawasan untuk penelitian ini.

3.2 Pengumpulan Data

Data yang peneliti gunakan dalam penelitian ini adalah data set publik yang disediakan oleh University of California, School of Information and Computer Science, pada Irvine (UCI) Machine Learning University, diunduh dari sits web http://archive.ics.uci.edu/ml/dataset.html. Data set tersebut antara lain sebagai berikut pada Tabel 3.1

Tabel 3.1 Data set

No	Dataset	Jumlah Data	Tipe Atribut	Jumlah Atribut	Jumlah Class
1	*bank Marketing*	4522	Categorial	11	2
2	*Tic-tac-toe Endgame*	958	Categorial	9	-

3.3 Preprocessing data

Pada tahap ini dilakukan pembersihan data dan perubahan format extensi data kedalam format yang bisa dibaca oleh *Weka* serta pengelompokan data kedalam masing-masing kategori.

3.3.1 Pembersihan Data

Pada tahap ini, dilakukan pembersihan terhadap data-data yan tidak lengkap, kosong atau *null*, data yang mengandung *noise*, dan data tidak konsisten. Pada tahap ini data yang bernilai *null* atau kosong, akan dibersihkan dengan cara dihapurs secara *manual*. Berikut rincian data yang akan dibersihkan :

3.3.2 Preprocesing data

Pada tahap ini dilakukan perubahan data untuk perhitungan metode K-nearest neighbour dimana metode ini membutuhkan data Numerik ,maka harus melakukan preprocessing data

age	job	marital	education	balance	housing	duration	campaign	pdays	previous	y
68	7	30	22	4189	0	897	2	7	1	yes
31	3	20	22	171	0	81	3	2	4	no
59	1	20	22	42	0	40	1	7	2	no
32	1	10	25	2536	1	958	6	6	5	yes
49	3	20	25	1235	0	354	3	15	2	yes
42	5	30	22	1811	1	150	1	9	1	no
78	7	30	21	229	0	97	1	14	1	yes
32	6	20	22	2089	1	132	1	11	0	yes
33	1	20	22	3086	1	765	1	30	2	yes
25	8	10	25	368	1	16	18	30	1	no

3.1.1 Analisis Data

Pada analisa data ini mempunyai tujuan yaitu untuk mengkategorikan data set ke dalam kategori tinggi, sedang dan rendah dan juga melihat informasi atribut dari masing-masing data set. Berikut tabel kategori tinggi, sedang dan rendah berdasarkan jumlah record dan jumlah variabel

Tabel 3.3 Kategori Jumlah record

Jumlah data	Kategori
<500	Rendah
500-1000	Sedang
>1000	Tinggi

Tabel 3.4 Kategori Jumlah variabel

Jumlah data	Kategori
<5	Rendah
5-10	Sedang
>10	Tinggi

Berdasarkan tabel kedua diatas, berikut 2 data set penelitian yang mana termasuk dalam kategori tinggi, sedang maupun rendah berdasarkan jumlah data dan variabel.

Tabel 3.5 Kategori data set penelitian

No	Dataset	Jumlah Data	Tipe Atribut	Jumlah Atribut	Jumlah Class
1	*bank Marketing*	4522	Categorial	11	2
2	*Tic-tac-toe Endgame*	958	Categorial	9	-

3.4 Analisa Spesifikasi

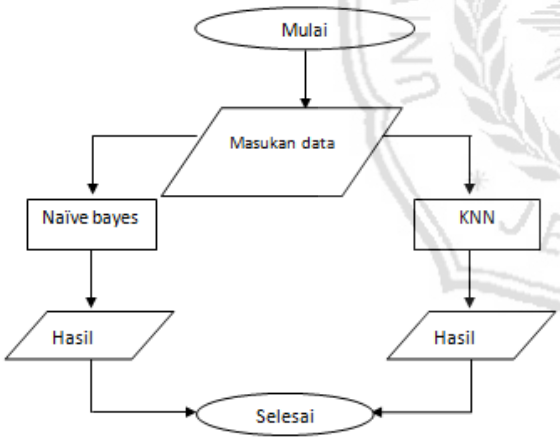
Pada analisis ini membahas tentang aspek – aspek yang di dinilai berpengaruh dalam pengujian data khususnya dalam Time Taken dalam proses data, antara lain sebagai berikut :

Tabel. 3.6 Spesifikasi Hardware & Softwere

Hardware	Software
1. Processor : Intel® Core™ i5 CPU M 560@ 2.67GHz 2. Operating System : Windows 10 Pro 3. Display Type : 14″ HD WLED 4. System Graphics: Integrated Graphic 5. RAM : 4 GB (2×2096) RAM 6. Hard Drive : 500GB HDD 7. Battery : 4 Cells 8. Weight : 3kg 9. Web-cam/HDMI/DVD-SM/VGA 10. LAN/BT/Card Reader/Speakers	1) Weka 3.8 2) Visio 2013

3.5 Implementasi

Implementasi *Weka* dilakukan dengan tahapan sebagai berikut :



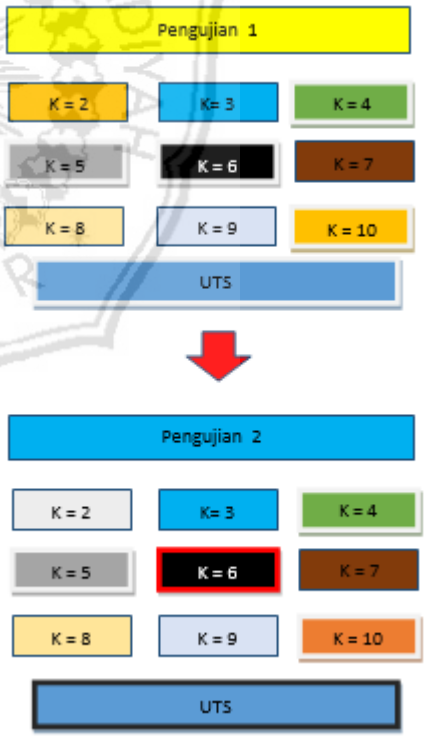
Gambar 3.1 Alur Pengujian dengan *Weka*

Berikut penjelasan dari gambar Alur pengujian diatas :

1. Import data
   Pada tahap ini, data di import melalui widget file, dengan memilih data set yang akan di implementasikan lalu menentukan target kelas sebagai label.
2. Penerapan metode *Naive bayes*
   Pada tahap ini, metode *Naive bayes* diterapkan untuk proses klasifikasi dengan memilih *Classsifier Naive bayes.*
3. Penerapan metode knn
   Pada tahap ini, metode knn diterapkan untuk proses klasifikasi dengan memilih *Classsifier* knn.
4. Hasil
   Pada tahap ini, hasil akurasi akan muncul pada bagian *Classsifier output* setelah data sudah pada tahap pemrosesan *Weka.*

3.6 Pengujian

Untuk menghasilkan nilai akurasi yang lebih variatif peneliti menggunakan dua teknik pengujian, teknik pengujian tersebut antara lain *Cross Validation* dan *Use training set. Cross Validation* merupakan teknik untuk menghasilkan sebuah akurasi dengan cara membagi data set ke dalam data testing dan data training. Dalam teknik ini dataset dibagi menjadi sejumlah K-buah partisi secara acak. Kemudian dilakukan sejumlah K-kali eksperimen, dimana masing-masing eksperimen menggunakan data partisi ke-K sebagai data testing dan memanfaatkan sisa partisi lainnya sebagai data training. Sedangkan teknik pengujian *Use training set* yaitu menggunakan seluruh data set sebagai data testing dan data training dalam proses perhitungan akurasi. Pada tahap ini pengujian dilakukan pada tiap-tiap data set yang telah di persiapkan dengan pengujian *Cross validation* dengan k=2 dan k= 5 dan pengujian *Use training set* dengan tujuan untuk mendapatkan nilai akurasi yang berbeda pada masing-masing data set dan tiap pengujian. Berikut gambar skenario pengujian data di bawah ini :



Gambar 3.2 Skenario Pengujian

Dari sekenario tersebut kemudian di implementasikan ke dalam pengujian menggunakan *Weka*. Dari pengujian tersebut kemudian akan menghasilkan akurasi dari masing-masing algoritma Naive Bayes dan C4.5 terhadap 10 data yang diujikan, jadi total pengujian yaitu 3 teknik uji x 10 data set x 2 algoritma = 60 kali uji coba. Lalu tahap selanjutnya yaitu analisa akurasi dengan menampilkan data akurasi kedalam tabel untuk mempermudah dalam proses analisa.

BAB IV HASIL DAN PEMBAHASAN

4.1 Implementasi dan Pengujian Algoritma Naïve Bayes dalam Weka

1. Pengujian Data 1 (*Bank marketing*)

Tabel 4.1 Hasil Pengujian Data 1

Tehnik pengujian	Akurasi
K=2	86,61%
K=3	86,86%
K=4	86,72%
K=5	86,75%
K=6	86,66%
K=7	86,64%
K=8	86,81%
K=9	86,81%
K=10	86,79%
UTS	87,01%

Pengujian pada data *Bank marketing* dengan algoritma Naïve Bayes menunjukkan hasil Akurasi untuk Teknik Pengujian *Cross Validation*dengan K =2 yaitu 86,61 % K = 3 yaitu 86 ,86 % K = 4 yaitu 86,72 % K = 5 yaitu 86,75% K= 6 yaitu 86,66 % K= 7 86,64% K= 8 yaitu 86,81% K = 9 86,81% dan pada K= 10 86,79 % selanjutnya pada UTS (*Use Training Set*) yaitu 87,01%.

2. Pengujian Data 2 (*Tic-Tac-Toe Endgame*)

Tabel 4.2 Hasil Pengujian Data 2

Tehnik pengujian	Akurasi
K=2	71,29%
K=3	70,87%
K=4	69,20%
K=5	69,72%
K=6	70,04%
K=7	69,51%
K=8	69,93%
K=9	70,14%
K=10	69,62%
UTS	69,83%

Pengujian pada data *Tic-Tac-Toe Endgame* algoritma Naïve Bayes menunjukkan hasil Akurasi untuk Teknik Pengujian *Cross Validation*dengan K = 2 yaitu 71,29% K = 3 yaitu 70,87% K = 4 69,20 % K = 5 yaitu 69,72% K= 6 70 ,04 % K = 7 yaitu 69,51% K = 8 yaitu 69,93 % K = 9 yaitu 70,14 dan pada K = 10 yaitu 69,62 %selanjutnya pada UTS (*Use Training Set*) yaitu 69,83% .

4.2 Implementasi dan Pengujian Algoritma K-Nearest Neighbor dalam Weka

1. Pengujian Data 1 (*Bank marketing*)

Tabel 4.3 Hasil Pengujian Data 1

Tehnik pengujian	Akurasi
K=2	87,87%
K=3	88,1%
K=4	88,34%
K=5	88,16%
K=6	88,01%
K=7	88,05%
K=8	88,34%
K=9	87,85%
K=10	88,16%
UTS	90,77%

Pengujian pada data *bank marketing* algoritma K-Nearest Neighbor menunjukkan hasil Akurasi untuk Teknik Pengujian *Cross Validation* dengan K = 2 yaitu 87,87% , K = 3 yaitu 88,1% ,K= 4 yaitu 88,34 % ,K = 5 yaitu 88,16% K = 6 yaitu 88,01 =% K=7 yaitu 88,05% K =8 yaitu 88,34% K=9 yaitu 87,85 % dan pada K = 10 yaitu 88,16% selanjutnya pada UTS (*Use Training Set*) yaitu 90,77%.

2. Pengujian Data 2 (*Tic-Tac-Toe Endgame*)

Tabel 4.4 Hasil Pengujian Data 2

Tehnik pengujian	Akurasi

K=2	94,88%
K=3	97,28%
K=4	97,18%
K=5	98,43%
K=6	98,43%
K=7	98,01%
K=8	98,64%
K=9	98,43%
K=10	98,74%
UTS	99,16%

Pengujian pada data *Tic-Tac-Toe Endgame* algoritma K-Nearest Neighbor menunjukkan hasil Akurasi untuk Teknik Pengujian *Cross Validation*dengan K = 2 yaitu 94,88% , K =3 yaitu 97,28 % ,K=4 yaitu 97,18% ,K=5 yaitu 98,43% ,K=6 yaitu 94,43% ,K=7 yaitu 98,01% ,K=8 yaitu 98,64% ,K=9 yaitu 98,43% dan pada K = 10 yaitu 98,74% , selanjutnya pada UTS (*Use Training Set*) yaitu 99,16%.

4.3 Analisis

Hasil dari pengujian terhadap seluruh data set di tampilkan dalam bentuk tabel untuk mempermudah dalam membandingkan algoritma Naïve Bayes dengan K-Nearest Neighbor dan peneliti melakukan analisa hasil Akurasi seluruh data set berdasarkan umlah data mulai dari terkecil sampai jumlah data terbesar. Berikut hasil perbandingan terhadap masing-masing data set berdasarkan teknik pengujian.

Tabel 4.5 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=2

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Akurasi Tertinggi
1	Bank marketing	4522	11	86,61%	87,87%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	71,29%	94,88%	K-Nearest Neighbor
	Rata-rata			78,95%	91.37%	

Tabel 4.6 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=3

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Akurasi Tertinggi
1	Bank marketing	4522	11	86,86%	88,1%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	70,87%	97,28%	K-Nearest Neighbor
	Rata-rata			78,86%	92,69%	

Tabel 4.7 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=4

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Akurasi Tertinggi
1	Bank marketing	4522	11	86,72%	88,34%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	69,20%	97,18%	K-Nearest Neighbor
	Rata-rata			77,96%	92,76%	

Tabel 4.8 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=5

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Tertinggi
1	Bank marketing	4522	11	86,75%	88,16%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	69,72%	98,43%	K-Nearest Neighbor
	Rata-rata			78,29%	93,68%	

Tabel 4.9 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=6

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Akurasi Tertinggi
1	Bank marketing	4522	11	86,66%	88.01%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	70,04%	98,43%	K-Nearest Neighbor
	Rata-rata			78,35%	93,22%	

Tabel 4.10 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=7

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Akurasi Tertinggi
1	Bank marketing	4522	11	86,64%	88,05%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	69,51%	98,01%	K-Nearest Neighbor
	Rata-rata			78,07%	93,03%	

Tabel 4.11 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=8

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Akurasi Tertinggi
1	Bank marketing	4522	11	86,81%	88,34%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	69,93%	98,64%	K-Nearest Neighbor
	Rata-rata			78,37%	93,49%	

Tabel 4.12 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=9

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Akurasi Tertinggi
1	Bank marketing	4522	11	86,81%	87,85%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	70,14%	98,43%	K-Nearest Neighbor
	Rata-rata			78,47%	93,14%	

Tabel 4.13 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=10

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Akurasi Tertinggi
1	Bank marketing	4522	11	86,79%	88,16%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	69,62%	98,74%	K-Nearest Neighbor
	Rata-rata			78,20%	93,45%	

Tabel 4.14 Analisa Akurasi pada Teknik Pengujian *Use Training Set* (UTS)

No	Dataset	Jumlah data	Jumlah atribut	Naïve Bayes	KNN	Tertinggi
1	Bank marketing	4522	11	87,01%	90,77%	K-Nearest Neighbor
2	Tic-Tac-Toe Endgame	958	9	69,83%	99,16%	K-Nearest Neighbor
	Rata-rata			78,42%	94.96%	

Pada penelitian diatas untuk pengujian K fold validation untuk Naïve bayes K2 tingkat akurasinya lebih baik dari K yang lainya yaitu sebesar 78,95 % untuk KNN UTS lebih baik dari K2, K3 , K4 , K5 , K6 , K7 , K8 , K10 yaitu memiliki akurasi sebesar 94,96 % sedangkan untuk naïve bayes yang memiliki tingkat akurasi terendah adalah K 4 sebesar 77 ,96% dan untuk K-Nearest Neighbor yaitu K2 sebesar 91.37%. setelah itu peneliti mengelompokkan hasil pengujian berdasarkan teknik pengujian dengan tujuan untuk mendapatkan hasil perbandingan dari seluruh data set yang telah di ujikan dengan algoritma Naïve Bayes dan K-Nearest Neighbor. Selanjutnya peneliti melakukan rata-rata pada masing-masing hasil akurasi algoritma untuk mendapatkan sebuah hasil dari penelitian.

Tabel 4.16 Kesimpulan Perbandingan Algoritma Naïve Bayes & K-Nearest Neighbor

	Teknik Pengujian	Naïve Bayes	K-Nearest Neighbor
Akurasi	K=2	78,95%	91,37%
	K=3	78,86%	92,69%
	K=4	77,96%	92,76%
	K=5	78,23%	93,68%
	K=6	78,35%	93,22%
	K=7	78,07%	93,03%
	K=8	78,37%	93,49%
	K=9	78,47%	93,14%
	K=10	78,20%	93,45%
	UTS	78,42%	94,96%
Rata Rata		78,38%	93,17%

Dari keseluruhan 2 data set yang telah di ujikan kedalam tool Weka, peneliti mendapatkan sebuah keseluruhan hasil pengujian dengan masing-masing Teknik Pengujian. K 2 untuk naïve bayes yaitu 78,95% untuk K-Nearest Neighbor 91,37% , K 3 untuk Naïve bayes yaitu 78,86 % dan KNN 92,69% K 4 untuk Naïve bayes 77,96 % K-Nearest Neighbor yaitu 92,76% , K 5 Naïve bayes 78,23% dan KNN 93,68% K 6 Naïve bayes 78,35% dan KNN yaitu 93,22% ,K 7 Naïve bayes 78,07% dan K-Nearest Neighbor 93,03 % ,K 8 naïve bayes 78,37 % dan K-Nearest Neighbor 93,49% K 9 Naïve bayes 78,47 dan % K-Nearest Neighbor 93,45 % K 10 Naïve bayes 78,20 % dan KNN 93,45 % untuk UTS Naïve bayes 78,42% dan K-Nearest Neighbor 94,17% ,maka didapat hasil rata rata dari pengujian K2, K3, K4, K5, K6, K7, K8, K9, K10 dan UTS yaitu naïve bayes 78,38% dan K-Nearest Neighbor 93,17%.

5.1 Kesimpulan

Berdasarkan penelitian yang sudah peneliti lakukan, dapat ditarik kesimpulan sebagai berikut :

1. Dari seluruh pengujian yang telah peneliti lakukan Pada teknikpengujian Cross Validation Maka didapat hasil rata rata dari pengujian K2, K3, K4, K5, K6, K7, K8, K9, K10 dan UTS yaitu naïve bayes 78,38% dan K-Nearest Neighbor 93,17%.

2. Berdasarkan Pengujian yang telah peneliti lakukan jika jumlah atribut lebih banyak didapatkan sebuah kesimpulan bahwa algoritma naïve bayes lebih akurat dari K-Nearest Neighbor namun dalam keselururuhan data K-Nearest Neighbor lebih akurat dalam Klasifikasi Data dibandingkan Naïve bayes.

3. Dengan menggunakan beberapa Teknik Pengujian algoritma K-Nearest Neighbor menghasilkan rata-rata akurasi yang tinggi pada proses klasifikasi.

5.2 Saran

Bagi peneliti - peneliti selanjutnya yang akan melakukan penelitian hampir serupa dan mengembangkan penelitian ini adalah :

1. Percobaan dengan mengunakan data dengan jumlah record, atribut dan parameter yang lebih banyak dan variatif agar hasil akurasi yang didapatkan bisa lebih berbeda dan lebih unggul mana diantara kedua algoritma.
2. Percobaan dengan menggunakan aplikasi selain Weka dalam analisa data dan mencoba menggunakan metode lain selain Algoritma K-Nearest Neighbor untuk perbandingan.

DAFTAR PUSTAKA

Prasetyo,Eko (2012). Data Mining konsep dan aplikasi menggunakan matlab.yogyakarta:Andi.

Dwi Luki arisandy(2017) Analisa perbandingan algoritma dan algoritma c4.5 untuk klasifikasi multi data

Nur Khotimah dan Deden Istiawan (2018) berjudul "Perbandingan Algoritma C4.5, Naïve Bayes dan K-Nearest Neighbour untuk Prediksi Lahan Kritis di Kabupaten Pemalang.

Wihardi yaya . (2013) . " K-fold Cross Validation".Diakses dari [http://blog.yayaw.web.id/riset/k-fold-crosss-validation] .

http://repository.telkomuniversity.ac.id/pustaka/94560/analisis-perbandingan-metode-k-nearest-neighbor-dan-naive-bayes-classifier-dalam-klasifikasi-teks.html

http://www.idx.co.id/id-id/beranda/perusahaantercatat/profilperusahaantercatat.aspx

http://www.cs.waikato.ac.nz/ml/people.html