ANALISIS PERBANDINGAN ALGORITMA NAIVE BAYES DAN ALGORITMA C4.5 UNTUK KLASIFIKASI MULTI DATA

[1]Dwi Luki Arisandy (1310651061)
[2] Victor Wahanggara, S.T, M.Kom
[3]Yeni Dwi Rahayu, S. ST., M. Kom
Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember
Email : lukiarie111088@gmail.com

ABSTRAK

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Metode Naïve Bayes memiliki tingkat akurasi tinggi pada saat proses klasifikasi. Hal tersebut diungkapkan oleh beberapa pendapat. Selain Naïve Bayes, algoritma C4.5 juga memiliki tingkat akurasi yang tinggi. Penelitian yang dilakukan oleh Puji Astuti (2016) berjudul "Komparasi Penerapan Algoritma C45, Knn Dan Neural Network Dalam Proses Kelayakan Penerimaan Kredit Kendaraan Bermotor" akurasi tertinggi yang dihasilkan C4.5 yaitu 92,89% dibandingka K-nn 77,78% dan Nn 91,1%. Dari kedua penelitian diatas menunjukkan bawa algoritma Naive Bayes dan algoritma C4.5 memiliki tingkat akurasi yang tinggi pada proses klasifikasi data. Pada penelitianyang saya lakukan dengan membandingkan kedua metode antara Naive Baye dan C4.5 menggunakan 10 data set dengan varian jumlah data record dan atribut yang berbeda menunjukkan hasil rata-rata akurasi untuk Algoritma C4.5 pada K=5 yaitu 79,28% dan Naive bayes 77,58%. Pada K=10 didapatkan sebuah hasil rata-rata akurasi C4.5 yaitu 79,38% dan Naive bayes 77%. Pada Use Training Set (UTS) didapatkan sebuah hasil rata-rata akurasi C4.5 yaitu 89,46% dan Naive bayes 79,39%. Dari haisl akurasi ketiga teknik pengujian tersebut didapatkan kesimpulan bahwa algoritma C4.5 lebih unggul untuk tingkat akurasi, namun untuk kecepatan proses klasifikasi Naive Bayes lebih unggul.

Kata kunci : *Algoritma Naïve Bayes, C4.5, Analisis, Klasifikasi, Perbandingan algoritma.*

BAB I PENDAHULUAN

1.1 Latar Belakang

Klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat dua proses yang dilakukan yaitu dengan membangun model untuk disimpan sebagai memori dan menggunakan model tersebut untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu data lain supaya diketahui di kelas mana objek data tersebut dimasukkan berdasarkan model yang telah disimpan dalam memori (Prasetyo, 2012). Algoritma Klasifikasi yang umum digunakan antara lain Artifficial Neural Network (ANN), Support Vector Machines (SVM), Decision Tree, Bayesian, dan sebagainya (Prasetyo, 2012).

Dari beberapa metode klasifikasi tersebut, terdapat metode yang memiliki tingkat akurasi tinggi yaitu Naïve Bayes. Hal tersebut diungkapkan oleh Han 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 Paulus Dian Wicaksana (2015) berjudul "Perbandingan Algoritma K-Nearest Neighbors dan Naive Bayes Untuk Studi Data Wisconsin Diagnosis Breast Cancer" Naive bayes menghasilkan akurasi 97.5% dan lebih tinggi dari K-Nearest Neighbor 95,6%.

Selain Naïve Bayes, algoritma C4.5 juga memiliki tingkat akurasi yang tinggi. Penelitian yang dilakukan oleh Puji Astuti (2016) berjudul "Komparasi Penerapan Algoritma C45, Knn Dan Neural Network Dalam Proses Kelayakan Penerimaan Kredit Kendaraan Bermotor" akurasi tertinggi yang dihasilkan C4.5 yaitu 92,89% dibandingka K-nn 77,78% dan Nn 91,1%.

Dari kedua penelitian diatas menunjukkan bawa algoritma Naive Bayes dan algoritma C4.5 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 Naive Bayes lebih unggul dengan jumlah data yang banyak dibandingkan algoritma C4.5? atau sebaliknya?

Oleh karena itu peneliti ingin melakukan sebuah penelitian dengan membandingkan algoritma Naive Bayes dan algoritma C4.5 menggunakan multi data yang peneliti ambil dari penyedia layanan dataset publik, dengan kriteria masing-masing data yang berbeda, hal ini bertujuan untuk menganalisa hasil akurasi dari kedua metode tersebut terhadap data yang diujikan.

Berdasarkan pembahasan diatas peneliti mengambil judul "Analisis Perbandingan Algoritma Naive Bayes dan Algoritma C4.5 Untuk Klasifikasi Multi Data".

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 C4.5. Berikut adalah tabel beberapa penelitian yang terkait dengan masalah tersebut :

1. **Penelitian oleh (Paulus Dian Wicaksana) Tahun 2015**

Penelitian yang pertama adalah penelitian yang dilakukan oleh Paulus Dian Wicaksana. Penelitian yang berjudul "Perbandingan Algoritma K-Nearest Neighbors dan Naive Bayes Untuk Studi Data "Wisconsin Diagnosis Breast Cancer " Metode yang digunakan pada peneliti ini adalah Algoritma Klasifikasi Naive Bayes dan K-Nearest Neighbor. Dengan menggunakan 683 data record Naive Bayes menghasilkan akurasi terbaik dibanding K-Nearest Neighbor. Dengan hasil akurasi Naive bayes 97,5% dan 95,8% pada K-Nearest eighbor.

2. Penelitian oleh (Puji Astuti) 2016

Penelitian yang kedua adalah penelitian yang dilakukan oleh Puji Astuti. Penelitian yang berjudul "Komparasi Penerapan Algoritma C45, Knn Dan Neural Network Dalam Proses Kelayakan Penerimaan Kredit Kendaraan Bermotor" Metode yang digunakan pada peneliti ini adalah Algoritma C4.5, Knn dan Neural Network. Data dalam penelitian ini 486 data maka digunakan metode cross validation untuk menguji tingkat akurasi. Untuk nilai akurasi model untuk metode C4.5 sebesar 92.8%, metode K-NN sebesar 77.78%, dan metode neural network sebesar 91.1%

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 model tersebut 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:

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

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

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

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 C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti Structured Query Language untuk mencari record pada kategori tertentu. pohon keputusan juga berguna untuk mengeksplorasi data, menemuka hubungan tersembunyi antara sejumlah calon variable input dengan sebuah variable target. Algoritma C4.5 ini mempunyai input berupa training samples dan samples. Training samples berupa data contoh yang akan digunakan untuk membangun sebuah pohon yang telah diuji kebenarannya. Sedangkan samples merupakan field-field data yang nanti akan kita gunakan sebagai parameter dalam melakukan klasifikasi data. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut.

a. Pilih atribut yang akan digunakan sebagai akar
b. Buatlah sebuah cabang untuk setiap nilai
c. Bagilah kasus dalam sebuah cabang
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama. Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5.

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan rumus seperti yang tertera dalam persamaan 1 berikut.

$$Entropy(S) = \sum_{i=1}^{n} -pi \; * \log_2 pi$$

Sehingga di sederhanakan :

$$Entropy(S) \equiv -p_+ log_2 p_+ - \; p_- log_2 p_-$$

Keterangan :

S : himpunan kasus
A : Atribut
n : jumlah partisi atribut
| Si | : jumlah kasus pada partisi ke –i
| S | : jumlah kasus dalam S

Sementara entropy dapat dilihat pada persamaan 2 berikut.

$$Gains(S, A) = Entropy - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si)$$

$$Gain(S, A) = \; Entropy(S) - \; \Sigma \frac{|Sv|}{|S|} \; Entropy(Sv)$$

Keterangan :
S : Himpunan kasus
A : Fitur
n : Jumlah partisi S
pi : Proporsi dari Si terhadap S

Untuk mengetahui analisis prediksi masa studi mahasiswa lulus tepat waktu dan tidak tepat waktu adalah dengan melakukan klasifikasi dari kumpulan data nilai mata kuliah umum tersebut. Salah satu metode klasifikasi adalah dengan membuat pohon keputusan. Berikut penjelasan lebih

terperinci mengenai tiap-tiap langkah dalam pembentukan pohon keputusan dengan menggunakan algoritma C4.5 untuk menyelesaikan permasalahan.

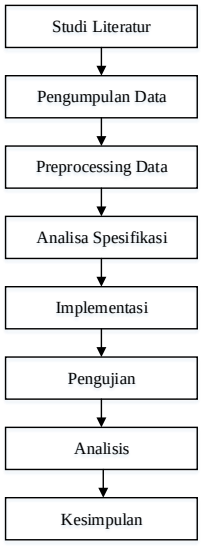
2.5. Tool Weka Dalam Klasifikasi

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.

Gambar 2.7 Halaman awal Weka

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 Penelitian

No	Dataset	Jumlah Data	Tipe Atribut	Jumlah Atribut	Jumlah Class
1	Balance Scale Weight & Distance	625	Categorial	5	3
2	Car Evaluation	1728	Categorial	7	4
3	Haberman's Survival Data	306	Integer	4	2
4	Iris Plants	150	Real	5	3
5	Qualitative_Bankruptcy	250	-	7	2
6	Tic-Tac-Toe Endgame	958	Categorial	9	2
7	Lenses	24	Categorial	4	3
8	Contraceptive Method Choice	1473	Categorial, Integer	10	3
9	Adult	30162	Categorial, Integer	15	2
10	Poker	25010	Categorial, Integer	11	10

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 :

Tabel 3.2 Jumlah Data Hasil Pembersihan

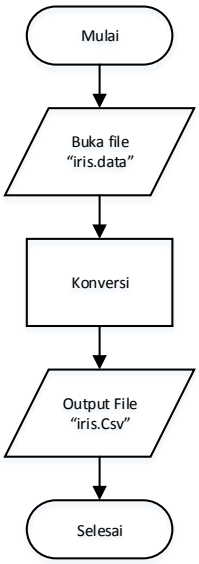
No	Dataset	Data Awal	Data tidak lengkap	Jumlah Data Bersih
1	Balance Scale Weight & Distance	625	-	625
2	Car Evaluation	1728	-	1728
3	Haberman's Survival Data	306	-	306
4	Iris Plants	150	-	150
5	Qualitative_Bankruptcy	250	-	250
6	Tic-Tac-Toe Endgame	958	-	958
7	Lenses	24	-	24

8	Contraceptive Method Choice	1473	-	1473
9	Adult	48842	18680	30162
10	Poker	25010	-	25010

3.3.2 Perubahan Format Data

Sebelum melakukan implementasi data set dengan *Weka*, semua format extensi data terlebih dahulu diubah kedalam format yang bisa dibaca oleh. Format awal data set pada saat di ambil dari UCI yaitu .data, setelah itu perubahan format data set dilakukan dengan menggunakan Microsoft excel dengan keluaran format extensi ".csv".

Berikut contoh alur proses konversi format data pada data set "iris.data" dengan menggunakan Microsoft excel



Gambar 3.2 Alur proses konversi format data

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 10 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	Jumlah Atribut	Kategori Jumlah data	Kategori Jumlah Atribut
1	Balance Scale Weight & Distance	625	5	Sedang	Sedang
2	Car Evaluation	1728	7	Tinggi	Sedang
3	Haberman's Survival Data	306	4	Rendah	Rendah
4	Iris Plants	150	5	Rendah	Sedang
5	Qualitative_Bankruptcy	250	7	Rendah	Sedang
6	Tic-Tac-Toe Endgame	958	9	Sedang	Sedang
7	Lenses	24	4	Rendah	Rendah
8	Contraceptive Method Choice	1473	10	Tinggi	Tinggi
9	Adult	30162	15	Tinggi	Tinggi
10	Poker	25010	11	Tinggi	Tinggi

3.4 Analisa Spesifikasi

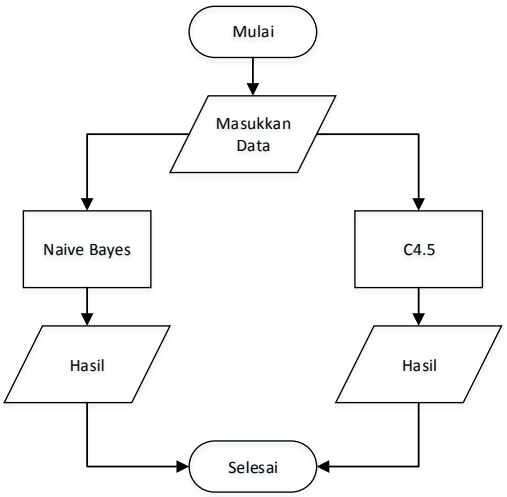
Pada analisis ini membahas tentang aspek – aspek yang di dinilai berpengaruh dalam pengujian data khususnya dalam Time Taken dalam proses data, antara lain sebagai berikut :

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 3.7 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 C4.5

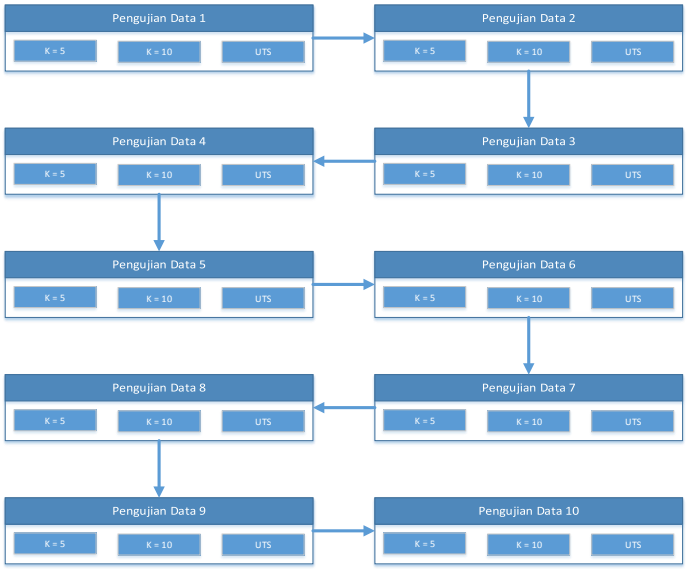
   Pada tahap ini, metode C4.5 diterapkan untuk proses klasifikasi dengan memilih *Classsifier* C4.5.

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=5 dan k= 10 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 3.8 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 Naive Bayes

1. Pengujian Data 1 (Balance Scale Weight & Distance)

Tabel 4.1 Hasil Pengujian Data 1

	Akurasi	Time Taken
K=5	89,60%	0 (s)
K=10	90,40%	0 (s)
UTS	90,88%	0 (s)

2. Pengujian Data 2 (Car Evaluation)

Tabel 4.2 Hasil Pengujian Data 2

	Akurasi	Time Taken
K=5	85%	0 (s)
K=10	86%	0 (s)
UTS	87%	0 (s)

3. Pengujian Data 3 (Haberman's Survival Data)

Tabel 4.3 Hasil Pengujian Data 3

	Akurasi	Time Taken
K=5	75%	0 (s)
K=10	75%	0 (s)
UTS	76%	0 (s)

Pengujian pada data *Haberman's Survival Data* dengan algoritma Naive bayes menunjukkan hasil Akurasi untuk Teknik Pengujian *Cross Validation* dengan K = 5 yaitu 75 % , Time Taken 0 (s) dan pada K = 10 yaitu 75% , Time Taken 0 (s) selanjutnya pada UTS (*Use Training Set*) yaitu 76%, Time Taken 0 (s).

4. Pengujian Data 4 (Iris Plants)

Tabel 4.4 Hasil Pengujian Data 4

	Akurasi	Time Taken
K=5	96%	0 (s)
K=10	96%	0 (s)
UTS	96%	0 (s)

5. Pengujian Data 5 (Qualitative_Bankruptcy)

Tabel 4.5 Hasil Pengujian Data 5

	Akurasi	Time Taken
K=5	99,20%	0 (s)
K=10	99,20%	0 (s)
UTS	100%	0 (s)

6. Pengujian Data 6 (Tic-Tac-Toe Endgame)

Tabel 4.6 Hasil Pengujian Data 6

	Akurasi	Time Taken
K=5	69,73%	0 (s)
K=10	70%	0 (s)
UTS	69,83%	0 (s)

7. Pengujian Data 7 (Lenses)

Tabel 4.7 Hasil Pengujian Data 7

	Akurasi	Time Taken
K=5	79%	0 (s)
K=10	75%	0 (s)
UTS	92%	0 (s)

8. Pengujian Data 8 (Contraceptive Method Choice)

Tabel 4.8 Hasil Pengujian Data 8

	Akurasi	Time Taken
K=5	48,95%	0 (s)
K=10	49,29%	0 (s)
UTS	49,76%	0 (s)

9. Pengujian Data 9 (Adult)

Tabel 4.9 Hasil Pengujian Data 9

	Akurasi	Time Taken
K=5	82,81%	0 (s)
K=10	82,85%	0 (s)
UTS	82,90%	0 (s)

10. Pengujian Data 10 (Poker)

Tabel 4.0 Hasil Pengujian Data 10

	Akurasi	Time Taken
K=5	49,96%	0,08 (s)
K=10	50%	0,07 (s)
UTS	49,95%	0,07 (s)

4.2. Implementasi dan Pengujian Algoritma C4.5

1. Pengujian Data 1 (Balance Scale Weight & Distance)

Tabel 4.11 Hasil Pengujian Data 1

	Akurasi	Time Taken
K=5	76,96%	0,12 (s)
K=10	76,64%	0,03 (s)
UTS	90,08%	0,01 (s)

2. Pengujian Data 2 (Car Evaluation)

Tabel 4.12 Hasil Pengujian Data 2

	Akurasi	Time Taken
K=5	91,55%	0,04 (s)
K=10	92,36%	0,01 (s)
UTS	96,29%	0,01(s)

3. Pengujian Data 3 (Haberman's Survival Data)

Tabel 4.13 Hasil Pengujian Data 3

	Akurasi	Time Taken
K=5	70,92%	0 (s)
K=10	71,89%	0 (s)
UTS	77,12%	0 (s)

4. Pengujian Data 4 (Iris Plants)

Tabel 4.14 Hasil Pengujian Data 4

	Akurasi	Time Taken
K=5	96%	0 (s)
K=10	96%	0 (s)
UTS	98%	0 (s)

5. Pengujian Data 5 (Qualitative_Bankruptcy)

Tabel 4.15 Hasil Pengujian Data 5

	Akurasi	Time Taken
K=5	98%	0 (s)
K=10	98%	0 (s)
UTS	99,60%	0 (s)

6. Pengujian Data 6 (Tic-Tac-Toe Endgame)

Tabel 4.16 Hasil Pengujian Data 6

	Akurasi	Time Taken
K=5	84,45%	0,01 (s)
K=10	85,07%	0,01 (s)
UTS	93,74%	0,01(s)

7. Pengujian Data 7 (Lenses)

Tabel 4.17 Hasil Pengujian Data 7

	Akurasi	Time Taken
K=5	83,33%	0 (s)
K=10	79,17%	0 (s)
UTS	91,67%	0 (s)

8. Pengujian Data 8 (Contraceptive Method Choice)

Tabel 4.18 Hasil Pengujian Data 8

	Akurasi	Time Taken
K=5	50,98%	0,12 (s)
K=10	53,22%	0,07 (s)
UTS	74,95%	0,05 (s)

9. Pengujian Data 9 (Adult)

Tabel 4.19 Hasil Pengujian Data 9

	Akurasi	Time Taken
K=5	85,63%	2,88 (s)
K=10	85,52%	2,5 (s)
UTS	88,26%	2,49 (s)

10. Pengujian Data 10 (Poker)

Tabel 4.20 Hasil Pengujian Data 10

	Akurasi	Time Taken
K=5	55,01%	6,48 (s)
K=10	55,90%	6,68 (s)
UTS	84,93%	7,43 (s)

4.3. Analisis

Hasil dari pengujian terhadap seluruh data set di tampilkan dalam bentuk tabel untuk mempermudah dalam membandingkan algoritma Naive bayes dengan C4.5 dan peneliti melakukan analisa hasil Akurasi seluruh data set berdasarkan jumlah data mulai dari terkecil sampai jumlah data terbesar. Berikut hasil perbandingan terhadap masing-masing data set berdasarkan teknik pengujian.

Tabel 4.21 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=5

No	Dataset	Jumlah data	Jumlah atribut	Naive bayes	C4.5	Akurasi Tertinggi	Time Taken Naive bayes	Time Taken C4.5
1	Lenses	24	4	79%	83,33%	C4.5	0	0
2	Iris Plants	150	5	96%	96%	-	0	0
3	Qualitative_Bankruptcy	250	7	99,20%	98%	Naive Bayes	0	0
4	Haberman's Survival Data	306	4	75%	70,92%	Naive Bayes	0	0
5	Balance Scale Weight & Distance	625	5	89,60%	76,96%	Naive Bayes	0	0,12
6	Tic-Tac-Toe Endgame	958	9	69,73%	84,45%	C4.5	0	0,01
7	Contraceptive Method Choice	1473	10	48,95%	50,98%	C4.5	0	0,12
8	Car Evaluation	1728	7	85%	91,55%	C4.5	0	0,04
9	Poker	25010	11	49,96%	55,01%	C4.5	0	6,48
10	Adult	30162	15	82,81%	85,63%	C4.5	0	2,88
	Rata-rata			**77,58%**	**79,28%**		**0**	**0,965**

Tabel 4.22 Analisa Akurasi pada Teknik Pengujian *Cross Validation* K=10

No	Dataset	Jumlah data	Jumlah atribut	Naive bayes	C4.5	Tertinggi	Time Taken Naive bayes	Time Taken C4.5
1	Lenses	24	4	75%	79,17%	C4.5	0	0
2	Iris Plants	150	5	96%	96%	-	0	0
3	Qualitative_Bankruptcy	250	7	99,20%	98%	Naive Bayes	0	0
4	Haberman's Survival Data	306	4	75%	71,89%	Naive Bayes	0	0
5	Balance Scale Weight & Distance	625	5	90,40%	76,64%	Naive Bayes	0	0,03
6	Tic-Tac-Toe Endgame	958	9	70%	85,07%	C4.5	0	0,01
7	Contraceptive Method Choice	1473	10	49,29%	53,22%	C4.5	0	0,07
8	Car Evaluation	1728	7	86%	92,36%	C4.5	0	0,01
9	Poker	25010	11	50%	55,90%	C4.5	0,07	6,68
10	Adult	30162	15	82,85%	85,52%	C4.5	0	2,5
	Rata-rata			**77%**	**79,38%**		**0**	**0,930**

Tabel 4.23 Analisa Akurasi pada Teknik Pengujian *Use Training Set* (UTS)

No	Dataset	Jumlah data	Jumlah atribut	Naive bayes	C4.5	Tertinggi	Time Taken Naive bayes	Time Taken C4.5
1	Lenses	24	4	92%	91,67%	Naive Bayes	0	0
2	Iris Plants	150	5	96%	98%	C4.5	0	0
3	Qualitative_Bankruptcy	250	7	100%	99,60%	Naive Bayes	0	0
4	Haberman's Survival Data	306	4	76%	77,12%	C4.5	0	0
5	Balance Scale Weight & Distance	625	5	90,88%	90,08%	Naive Bayes	0	0,01
6	Tic-Tac-Toe Endgame	958	9	69,83%	93,74%	C4.5	0	0,01
7	Contraceptive Method Choice	1473	10	49,76%	74,95%	C4.5	0	0,05
8	Car Evaluation	1728	7	87%	96,29%	C4.5	0	0,01
9	Poker	25010	11	49,95%	84,93%	C4.5	0,07	7,43
10	Adult	30162	15	82,90%	88,26%	C4.5	0	2,49
	Rata-rata			**79,39%**	**89,46%**		**0**	**1,000**

Peneliti mengelompokkan hasil pengujian berdasarkan teknik pengujian dengan tujuan untuk mendapatkan hasil perbandingan dari seluruh data set yang telah di ujikan

dengan algoritma Naive bayes dan C4.5. Selanjutnya peneliti rata-rata pada masing-masing hasil akurasi algoritma dan juga time taken untuk mendapatkan sebuah hasil dari penelitian.

4.4. Kesimpulan

Tabel 4.25 Kesimpulan perbandingan algoritma Naïve Bayes & C4.5

	T. Pengujian	Naïve Bayes	C4.5
Akurasi	K=5	77,58%	79,28%
	K=10	77%	79,38%
	UTS	79,39%	89,46%
Time Taken	K=5	0 s	0,965 s
	K=10	0 s	0,930 s
	UTS	0 s	1 s

Dari keseluruhan 10 data set yang telah di ujikan kedalam tool Weka, peneliti mendapatkan sebuah keseluruhan hasil pengujian dengan masing-masing Teknik Pengujian. Pada teknik pengujian Cross Validation dengan K=5 didapatkan sebuah hasil rata-rata akurasi C4.5 yaitu 79,28% dan Naive bayes 77,58% dimana C4.5 lebih unggul 1,7% dibandingkan Naive bayes. Namun untuk Time taken Algoritma Naives lebih cepat di bandingkan C4.5, Naive bayes 0 (s) dan C4.5 0,965 (s). Selanjutnya Pada teknik pengujian Cross Validation dengan K=10 didapatkan sebuah hasil rata-rata akurasi C4.5 yaitu 79,38% dan Naive bayes 77% dimana C4.5 lebih unggul 2,38% dibandingkan Naive bayes dan untuk Time taken Algoritma Naives lebih cepat di bandingkan C4.5, Naive bayes 0 (s) dan C4.5 0,930 (s). Pada teknik pengujian Use Training Set (UTS) didapatkan sebuah hasil rata-rata akurasi C4.5 yaitu 89,46% dan Naive bayes 79,39% dimana C4.5 lebih unggul 10,07% dibandingkan Naive bayes. Namun untuk Time taken Algoritma Naives lebih cepat di bandingkan C4.5, Naive bayes 0 (s) dan C4.5 1 (s). Time Taken yang dihasilkan oleh algoritma C4.5 dipengaruhi oleh banyaknya jumlah atribut, hal ini bisa dilihat pada 2 dataset yang menghasilkan time taken cepat yaitu pada Poker dan Adult. Tentu dapat dijadikan sebuah kesimpulan bahwa pada algoritma C4.5 yang mempengaruhi Time Taken yaitu Jumlah dari kelasnya juga mempengaruhi Time Taken untuk proses Klasifikasi data.

BAB V KESIMPULAN DAN SARAN

5.1 Kesimpulan

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

1. Dari seluruh pengujian yang telah peneliti lakukan Pada teknik pengujian Cross Validation dengan K=5 didapatkan sebuah hasil rata-rata akurasi C4.5 yaitu 79,28% dan Naive bayes 77,58%. Pada teknik pengujian Cross Validation dengan K=10 didapatkan sebuah hasil rata-rata akurasi C4.5 yaitu 79,38% dan Naive bayes 77%. Pada teknik pengujian Use Training Set (UTS) didapatkan sebuah hasil rata-rata akurasi C4.5 yaitu 89,46% dan Naive bayes 79,39%.
2. Berdasarkan Pengujian yang telah peneliti lakukan didapatkan sebuah kesimpulan bahwa algoritma C4.5 lebih akurat dalam Klasifikasi Data dibandingkan Naive Bayes, namun untuk Time Taken Naive Bayes lebih cepat dalam Kalasifikasi Data.
3. Dalam algoritma C4.5 yang mempengaruhi dalam cepat dan lambatnya sebuah proses (Time Taken) yaitu Jumlah Kelas pada sebuat dataset.
4. Dengan menggunakan beberapa Teknik Pengujian algoritma C4.5 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 C4.5 untuk perbandingan.

DAFTAR PUSTAKA

Berry, I. H. and Browne, M. 2006. Lecture Notes in DATA MINING. USA:

World Scientific.

Jawei Han, Micheline Kamber (2006). Concept and techniques. San Franscisco: Elsevier Inc.

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

Arrasyid, R.A. (2013). "Efektivitas Media *Software Excel* Dalam Model Pembelajaran *Contextual Teaching And Learning* Sebagai Upaya Meningkatkan Hasil Belajar Pokok Bahasan Kertas Kerja Mata Pelajaran Ekonomi Pada Siswa Kelas XI IPS Sma Negeri 2 Magelang Tahun Ajaran 2012/2013". *Skripsi*. Semarang : Pendidikan Ekonomi Universitas Negeri Semarang.

Wicaksana, P.D. (2015). "Perbandingan Algoritma K-Nearest Neighbors dan Naive Bayes Untuk Studi Data *"Wisconsin Diagnosis Breast Cancer*". *Skripsi.* Yogyakarta : Teknik Informatika Universitas Sanata Dharma.

Iskandar, Derick. (2013), “Perbandingan akurasi klasifikasi tingkat kemiskinan antara algoritma C4.5 dan Naïve Bayes Clasifier”.

Supriyanti, Wiwit. (2016) . “Perbandingan Kinerja Algoritma C4.5 Dan Naive Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa”.

Kamus Besar Bahasa Indonesia 2012-2015. “Arti Kata Analisa”, (Online), (Kamus Besar Bahasa Indonesia (KBBI) Online - arti kata analisa, diakses 8 Januari 2017)

Metode C4.5, (2015) “Contoh Soal algoritma C4.5 dan Perhitungannya”. di akses *[http://mardani-g4ul.blogspot.co.id/2011/12/contoh-soal-algoritma-c45.html]* (diakses 7 Januari 2017).

Nugraha, Erfan. (2010). “Perbedaan Mekanisme, Proses, Tinjauan, Analisis, dan Evaluasi”. Diakses dari [https://drummerfan.wordpress.com/2010/03/25/perbedaan-mekanisme-proses-tinjauan-analisis-dan-evaluasi/] (diakses 8 Januari 2017).

Wihardi yaya. (2013). “K-Folds Cross Validation”. Diakses dari [http://blog.yayaw.web.id/riset/k-folds-cross-validation] (diakses 16 Januari 2017).

Website ILMUSKRIPSI. http://www.ilmuskripsi.com

Website WEKA. http://www.cs.waikato.ac.nz/ml/weka/.