Kumpulan jurnaL Ilmu Komputer (KLIK)
Volume 02, No.02 September 2015
ISSN: 2406-7857

Penerapan K-Optimal Pada Algoritma Knn untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4

Mutiara Ayu Banjarsari[1], H. Irwan Budiman, Andi Farmadi[3]
[1,2,3]Prodi Ilmu Komputer FMIPA UNLAM
Jl. A. Yani Km 36 Banjarbaru, Kalimantan selatan
[1]Email: mutiaraayubanjarsari@gmail.com

Abstract

The data pile on a database of academic information systems at Computer Science Program of Mathematic and Natural Science Faculty of Lambung Mangkurat University is not fully utilized, although it can provide new information that has not been known before. Data mining techniques can be used to predict the timely graduation of students. The k-Nearest Nieghbor, a method to classify objects based on training data located closest to the object, was used in this study. Selection of the value of k in kNN algorithm became important because it would affect the performance of the algorithm kNN, therefore it was necessary to know how the value of k and the level of accuracy. The k-Fold Cross Validation method and Accuracy Test was used to determine the value of k-Optimal. The result showed that the value of k = 5 was defined as k-Optimal which was then be applied in the kNN algorithm for prediction of timely graduation of students based on the Grade Point Average up to 4th semester.

Keywords: *kNN, k-Optimal, Classification, Data mining, k-Fold Cross Validation method*

Abstrak

Tumpukan data pada database sistem informasi akademik Program Studi Ilmu Komputer FMIPA Unlam belum dimanfaatkan secara maksimal, padahal dari data tersebut dapat memberikan sebuah informasi baru yang belum diketahui sebelumnya. Teknik data mining dapat digunakan untuk memprediksi kelulusan tepat waktu mahasiswa. Penelitian menggunakan metode k-Nearest Nieghbor yang merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data training yang jaraknya paling dekat dengan objek tersebut. Pemilihan nilai k pada algoritma kNN menjadi hal yang penting karena akan mempengaruhi kinerja dari algoritma kNN, oleh karena itu perlu diketahui berapa nilai k dan tingkat akurasinya. Metode k-Fold Cross Validation dan Uji Akurasi digunakan untuk mengetahui nilai k-Optimal. Hasil yang didapat adalah nilai k=5 dengan tingkat akurasi sebesar 80.00% yang ditetapkan sebagai k-Optimal. Nilai k=5 diterapkan pada algoritma kNN untuk prediksi kelulusan tepat waktu mahasiswa berdasarkan IP sampai dengan semester 4.

Kata Kunci *: kNN, k-Optimal, Klasifikasi, Data mining, Sistem Informasi Akademik, Metode k-Fold Cross Validation*

1. PENDAHULUAN

Data akademik atau informasi yang ada pada suatu perguruan tinggi akan semakin bertambah seiring dengan berlangsungnya proses kegiatan akademik. Hal ini menciptakan kondisi yang membuat adanya suatu tumpukan data. Pada sistem informasi akademik program studi Ilmu Komputer FMIPA Universitas Lambung Mangkurat terdapat berbagai macam kumpulan data didalamnya. Sampai saat ini, data mahasiswa yang ada belum dimanfaatkan secara maksimal sehingga perlu diolah untuk menemukan sebuah informasi baru. Data yang digunakan adalah data IP mahasiswa mulai dari semester satu sampai dengan semester empat, menggunakan salah satu fungsi *data mining* yaitu fungsi klasifikasi dengan menggunakan algoritma *k-Nearest Neighbor* dengan harapan dapat memprediksi kelulusan tepat waktu mahasiswa sehingga dapat digunakan oleh pihak program studi untuk mencari solusi atau kebijakan dalam proses evaluasi pembelajaran di program studi ilmu komputer.

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Moertini, 2002). Teknik *data mining* merupakan sebuah proses ekstraksi informasi untuk menggali pengetahuan (*knowledge discovery*) dan menemukan pola (*pattern recognition*) pada tumpukan data dalam database yang biasanya berskala besar. (Larose, 2005).

Klasifikasi adalah proses penemuan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep yang bertujuan agar bisa digunakan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui (Han, 2006). Algoritma yang digunakan untuk melakukan fungsi klasifikasi adalah algoritma kNN. kNN adalah metode klasifikasi yang menentukan kategori berdasarkan mayoritas kategori pada *k-Nearest Neighbor* (Liu, 2007). kNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing (Wu, 2009).

Dalam Algoritma kNN terdapat salah satu parameter yaitu k. Berdasarkan jurnal "*The Top Ten Algorithms in Data mining*" (Wu, 2009) pemilihan nilai k menjadi hal yang penting karena akan mempengaruhi kinerja algoritma kNN. Nilai k yang terlalu kecil, maka hasil klasifikasi akan lebih terpengaruh oleh *noise*. Di sisi lain, jika nilai k terlalu 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*. Sampai saat ini masih belum diketahui nilai *k-Optimal* dalam algoritma kNN untuk studi kasus yang diteliti.

Kajian penelitian terdahulu dilakukan oleh Astrid Darmawan dari Universitas Komputer Indonesia pada tahun 2012 dengan judul Skripsi *Pembuatan Aplikasi Data Mining untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighborhood*. Penelitian yang telah dilakukan sebelumnya menjadi bahan acuan penulis untuk melakukan penelitian ini. Penelitian lain dilakukan oleh Emerensye S. Y. Pandie yang berjudul *Implementasi Algoritma Data mining K-Nearest Neighbour (kNN) dalam Pengambilan Keputusan Pengajuan Kredit*. Penelitian ini menggunakan metode k-fold cross validation dalam pencarian k terbaik.

Penelitian sekarang ini mencoba untuk mengetahui nilai k-Optimal dan tingkat akurasinya pada algoritma kNN untuk prediksi kelulusan tepat waktu mahasiswa berdasarkan IP sampai dengan semester 4. Data yang digunakan untuk penelitian yaitu data nilai IP mahasiswa program studi Ilmu Komputer FMIPA Unlam yang telah lulus dari angkatan 2006 hingga 2009. Data mahasiswa tersebut selanjutnya diklasifikasikan menjadi "Ya" dan "Tidak". Jika mahasiswa tersebut lulus tepat waktu (<=5 Tahun) maka masuk dalam klasifikasi "Ya", sebaliknya jika mahasiswa lulus selama >5 Tahun maka termasuk dalam klasifikasi "Tidak" atau tidak tepat waktu.

2. METODE PENELITIAN

2.1. Prosedur Penelitian

Penelitian ini pertama diawali dengan tahap identifikasi dan pengumpulan data, pencarian literatur, kemudian dilanjutkan dengan tahap persiapan dan pemilihan data, selanjutnya pembersihan data, pembentukan data baru, proses data mining, dan yang terakhir adalah evaluasi.

a. Identifikasi dan Pengumpulan Data

   Pada tahap ini dilakukan identifikasi terhadap penelitian yang akan dilakukan dan melakukan pengumpulan data yang sesuai dengan penelitian. Pada pengumpulan data, data yang digunakan adalah data IP mahasiswa angkatan 2006-2009 mahasiswa Ilmu Komputer FMIPA Unlam. Data ini didapatkan dari database Sistem Informasi Akademik Program Studi Ilmu Komputer FMIPA Unlam.

b. Pencarian Literatur

   Tahap ini adalah mencari literatur dari buku-buku maupun jurnal penelitian terdahulu tentang prediksi kelulusan tepat waktu mahasiswa, metode *data mining* yang digunakan yaitu *k-Nearest Neighbor*, dan algoritma untuk pencarian *k-Optimal* yaitu *k-Fold Cross Validation*.

c. Persiapan dan Pemilihan Data

   Melakukan persiapan terhadap data yang telah didapat seperti melihat struktur tabel yang ada pada database. Pemilihan data dilakukan karena tidak semua tabel serta data yang ada dalam database berhubungan dengan penelitian yang dilakukan, sehingga hanya data yang berkaitan dengan penelitian yang akan digunakan.

d. Pembersihan Data

   Tahap ini dilakukan untuk memastikan bahwa tidak ada data yang terduplikasi, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data. Data yang telah bersih dari kesalahan dapat mempermudah penelitian dan mencegah adanya kesalahan pada penelitian.

e. Pembentukan Data Baru

   Pembentukan data baru ini agar data yang didapat dan telah bersih dari kesalahan bisa dibentuk menjadi sebuah tabel baru yang sesuai dengan algoritma kNN untuk prediksi kelulusan tepat waktu mahasiswa serta mencari *k-Optimal* pada algoritma kNN.

f. Proses *Data Mining*

Tahap yang sangat penting dalam penelitian. Pada tahap ini ada beberapa tahap dilakukan yaitu:

1) Proses pencarian *k-Optimal* pada algoritma kNN menggunakan metode *k-Fold Cross Validation*.
2) Setelah mendapatkan nilai *k-Optimal*, nilai tersebut digunakan untuk di uji akurasinya menggunakan data real sebanyak 17 buah dari database sehingga akan diketahui berapa banyak data yang memiliki banyak ketepatan prediksi.
3) Tahap terakhir adalah melakukan prediksi kelulusan tepat waktu mahasiswa dengan menggunakan algoritma kNN dengan nilai k hasil dari k-Folds Cross Validation dengan variabel input yaitu IP sampai dengan semester 4.

g. Evaluasi

Tahap ini adalah tahap dimana pola informasi yang dihasilkan dari proses *data mining* ditampilkan dalam bentuk yang dapat dipahami oleh pihak yang berkepentingan.

2.2. Bahan Penelitian

Bahan yang diperlukan adalah data yang mendukung dalam penelitian ini. Penelitian ini menggunakan data IP tiap mahasiswa yang telah lulus dan memenuhi syarat untuk diterapkan pada algoritma kNN. Data yang digunakan adalah IP mahasiswa Ilmu Komputer FMIPA Unlam angkatan 2006 – 2009.

3. HASIL DAN PEMBAHASAN

3.1. *Data Selection*

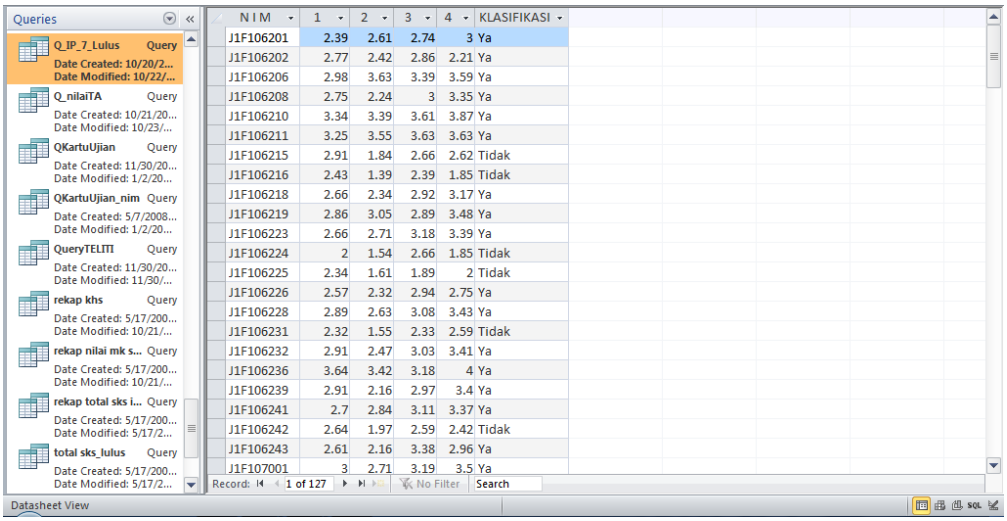
Tahap pemilihan data dilakukan untuk mendapatkan data yang sesuai dengan penelitian. Penelitian memerlukan data mahasiswa yang telah lulus dari angkatan 2006 hingga 2009. Dari tahap pemilihan data ini didapatkan data sebanyak 127. Data mahasiswa yang digunakan yaitu Nim dan data nilai IP Semester tiap mahasiswa. Data ini selanjutnya akan dilakukan perubahan bentuk tabel pada tahap transformasi, sehingga bentuk tabel akhir yang dibuat berisi *field* NIM, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, dan Klasifikasi.

3.2. *Data Pre-processing / Cleaning*

Pada penelitian ini tahap pembersihan data tidak dilakukan karna umumnya data pada database Sistem Informasi Akademik Program Studi Ilmu Komputer FMIPA Unlam sudah tidak ada duplikasi, kesalahan, dan *validation rules* pada database sudah sesuai dengan penelitian.

3.3. *Transformation*

Tahap ini merupakan proses untuk melakukan perubahan bentuk tabel terhadap data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Tabel yang akan digunakan pada penelitian berupa tabel dengan *field* atau kolom NIM, 1 (Nilai IP Semester 1), 2 (Nilai IP Semester 2), 3 (Nilai IP Semester 3), 4 (Nilai IP Semester 4), dan Klasifikasi.



NIM	1	2	3	4	KLASIFIKASI
J1F106201	2.39	2.61	2.74	3	Ya
J1F106202	2.77	2.42	2.86	2.21	Ya
J1F106206	2.98	3.63	3.39	3.59	Ya
J1F106208	2.75	2.24	3	3.35	Ya
J1F106210	3.34	3.39	3.61	3.87	Ya
J1F106211	3.25	3.55	3.63	3.63	Ya
J1F106215	2.91	1.84	2.66	2.62	Tidak
J1F106216	2.43	1.39	2.39	1.85	Tidak
J1F106218	2.66	2.34	2.92	3.17	Ya
J1F106219	2.86	3.05	2.89	3.48	Ya
J1F106223	2.66	2.71	3.18	3.39	Ya
J1F106224	2	1.54	2.66	1.85	Tidak
J1F106225	2.34	1.61	1.89	2	Tidak
J1F106226	2.57	2.32	2.94	2.75	Ya
J1F106228	2.89	2.63	3.08	3.43	Ya
J1F106231	2.32	1.55	2.33	2.59	Tidak
J1F106232	2.91	2.47	3.03	3.41	Ya
J1F106236	3.64	3.42	3.18	4	Ya
J1F106239	2.91	2.16	2.97	3.4	Ya
J1F106241	2.7	2.84	3.11	3.37	Ya
J1F106242	2.64	1.97	2.59	2.42	Tidak
J1F106243	2.61	2.16	3.38	2.96	Ya
J1F107001	3	2.71	3.19	3.5	Ya

Record: 1 of 127 No Filter Search

Gambar 1. Tabel data *training*

Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

3.4. *Data Mining*

Proses ini bertujuan untuk mendapatkan pola-pola dan informasi yang tersembunyi di dalam basis data yang telah melewati tahap transformation. Pada tahap ini dibuat tiga *form* utama yang digunakan untuk penelitian yaitu *form* k-Fold Cross validation, uji akurasi, dan prediksi kNN. Metode K-Fold Cross Validation dan Uji akurasi digunakan untuk mengetahui nilai k-Optimal. Setelah mendapatkan nilai k-Optimal, maka selanjutnya menggunakan nilai k tersebut untuk melakukan prediksi kelulusan tepat waktu mahasiswa berdasarkan IP sampai dengan semester 4.

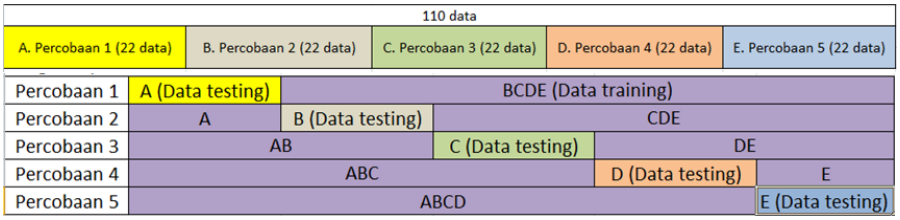
3.4.1. k-Fold Cross Validation

Dalam algoritma kNN sebuah data baru diklasifikasikan berdasarkan jarak data baru tersebut dengan tingkat kemiripan data baru terdekat terhadap data pola. Jumlah data tetangga terdekat ditentukan dan dinyatakan dengan k. Penentuan nilai k terbaik dapat ditentukan dengan optimasi parameter, misalnya dengan menggunakan *k-Fold Cross Validation* yang merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak (Pandie, 2012).

Menurut Fu (1994), *k-Fold Cross Validation* mengulang k-kali untuk membagi sebuah himpunan contoh secara acak menjadi k subset yang saling bebas, setiap ulangan disisakan satu subset untuk pengujian dan subset lainnya untuk pelatihan. Menurut Hastie et al (2008), dengan K=5 atau 10 dapat digunakan
untuk memperkirakan tingkat kesalahan yang terjadi, sebab data *training* pada setiap *fold* cukup berbeda dengan data training yang asli. Secara keseluruhan, 5 atau 10-fold cross validation sama-sama direkomendasikan dan disepakati bersama. Menghitung nilai akurasinya dapat dilakukan dengan menggunakan persamaan (Rodiyansyah, 2013) :

$$\text{Akurasi} = \frac{\textbf{Jumlah klasifikasi benar}}{\textbf{Jumlah data uji}} \boldsymbol{x}\ \mathbf{100}\%$$

Form ini merupakan proses untuk mendapatkan *k-Optimal* menggunakan metode *k-Fold Cross Validation* dengan pengulangan sebanyak 5 kali percobaan. Langkah metode *k-Fold Cross Validation* ini adalah dengan membagi data *training* yang berjumlah 110 menjadi 5 bagian yang sama yaitu 22 buah data tiap bagian data *testing*. Data *training* ini sebelumnya telah dilakukan pengacakan data. Tiap satu bagian percobaan dilakukan prediksi kNN dengan nilai k tertentu yang hasilnya dibandingkan dengan data *real.* Proses pembagian data untuk satu nilai k yang dimasukkan ke dalam perhitungan algoritma kNN:



Gambar 2. Pembagian data pada metode *k-Fold Cross Validation*

Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

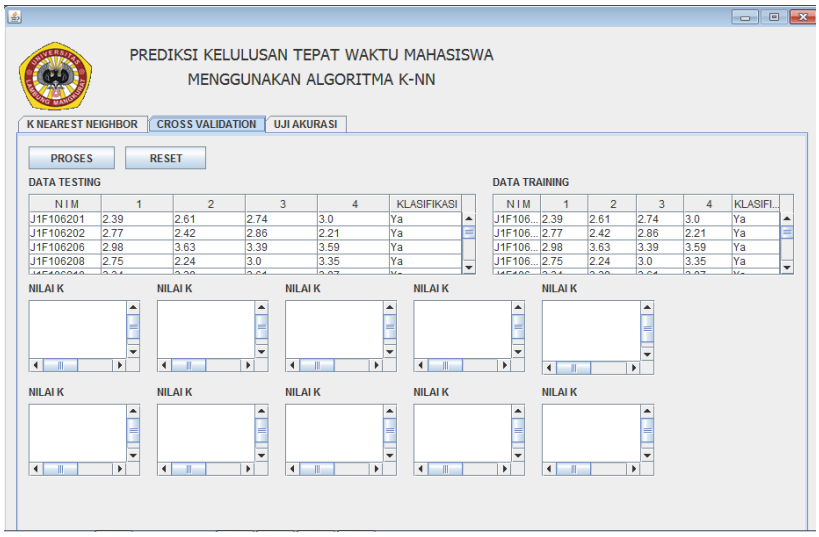
Setiap data *testing* dilakukan prediksi menggunakan algoritma kNN dengan setiap nilai k yang dimasukkan. Hasil klasifikasi prediksi kNN dibandingkan dengan klasifikasi data *real* dan dihitung jumlah prediksi yang tepat atau cocok dengan data *real*. Tingkat akurasi yang tinggi itulah yang terpilih menjadi nilai k terbaik atau *k-Optimal* (Banjarsari, 2015).

data ke-	Hasil prediksi sistem (knn)	Data real
1	Ya	Ya
2	Tidak	Ya
3	Ya	Ya
4	Tidak	Tidak
5	Tidak	Tidak
6	Ya	Ya
7	Ya	Ya
8	Ya	Ya
9	Ya	Tidak
10	Tidak	Tidak
11	Ya	Ya
12	Ya	Ya
13	Tidak	Tidak
14	Ya	Ya
15	Ya	Ya
16	Ya	Ya
17	Ya	Tidak
18	Ya	Ya
19	Tidak	Tidak
20	Tidak	Tidak
21	Tidak	Ya
22	Ya	Ya

Gambar 3. Contoh perbandingan antara data real dan hasil prediksi kNN

Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

Gambar 3. merupakan ilustrasi bagaimana hasil prediksi sistem dibandingkan dengan data *real* untuk tiap sekali percobaan, hasil prediksi yang salah ada 4, yg benar ada 18. Cara perhitungan percobaan 1 adalah dengan membagi jumlah prediksi yang tepat dengan jumlah data *testing*. Setelah percobaan1 selesai, maka dilanjutkan ke percobaan 2 sampai 5 dan dicari rata-ratanya, lalu dikali dengan 100%.



Gambar 4. Form k-Fold Cross Validation

Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

3.4.2. Uji Akurasi

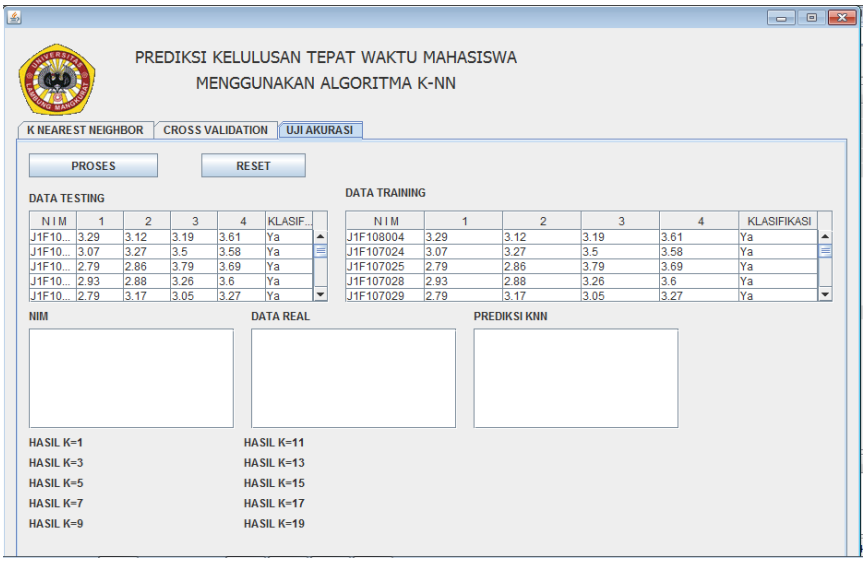
Form uji akurasi ini digunakan untuk mengetahui ketepatan dari nilai k terbaik yang didapatkan dari proses sebelumnya yaitu *k-Fold Cross Validation*. Perhitungan pada proses ini adalah dengan membandingkan antara data *real* dan hasil prediksi menggunakan kNN. Sebelumnya terdapat jumlah data mahasiswa sebanyak 127 data, namun untuk uji akurasi ini data *training* yang digunakan sebanyak 110 data dan sisanya yaitu 17 data digunakan sebagai bahan untuk uji akurasi ketepatan dari nilai k terbaik dari proses *k-Fold Cross Validation*.

Data	Data real	Hasil prediksi menggunakan k-NN
J1F108021	Tidak	Tidak
J1F108031	Tidak	Tidak
J1F108040	Ya	Ya
J1F108045	Ya	Ya
J1F108046	Ya	Ya
J1F108047	Ya	Ya
J1F108049	Ya	Ya
J1F108051	Ya	Ya
J1F108052	Ya	Ya
J1F108053	Ya	Ya
J1F108056	Ya	Ya
J1F108057	Ya	Ya
J1F108062	Tidak	Ya
J1F108201	Ya	Tidak
J1F108202	Ya	Ya
J1F108203	Ya	Ya
J1F108204	Ya	Ya

Gambar 5. Hasil perbandingan antara data real dan hasil prediksi kNN

Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

Gambar 5. memperlihatkan bahwa dari 17 data *real* terdapat 15 data yang hasil prediksinya tepat atau hasil antara prediksi kNN dan data *real* adalah sama.



Gambar 6. Form Uji Akurasi

Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

3.4.3. k-Nearest Neighbor

Algoritma *k-Nearest Neighbor* (kNN) merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. kNN termasuk algoritma *supervised learning* dimana hasil dari *query instance* yang baru diklasifikan berdasarkan mayoritas dari kategori pada kNN. Kelas yang paling banyak muncul itu yang akan menjadi kelas hasil klasifikasi. Tujuan dari algoritma ini adalah mengklasifikasikan objek baru berdasarkan atribut dan *training sample*.

Algoritma *k- Nearest Neighbor* menggunakan klasifikasi ketetanggaan (*neighbor*) sebagai nilai prediksi dari *query instance* yang baru. Algoritma ini sederhana, bekerja berdasarkan jarak terpendek dari *query instance* ke *training sample* untuk menentukan ketetanggaannya (Rizal, 2013). Langkah-langkah untuk menghitung metode *k-Nearest Neighbor* antara lain:

a. Menentukan parameter k
b. Menghitung jarak antara data yang akan dievaluasi dengan semua pelatihan
c. Mengurutkan jarak yang terbentuk
d. Menentukan jarak terdekat sampai urutan k
e. Memasangkan kelas yang bersesuaian
f. Mencari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi

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

Keterangan:
x_1 = Sampel data
x_2 = Data uji atau data testing
i = Variabel data
d = Jarak
p = Dimensi data

Contoh proses data mining menggunakan algoritma k-NN yaitu sebagai berikut:

Terdapat beberapa data yang berasal dari IP mahasiswa Program Studi Ilmu Komputer FMIPA Unlam yang telah lulus sebagai training data untuk diklasifikasikan dengan testing data menggunakan empat atribut yaitu IP dari semester satu sampai semester empat sehingga dapat menentukan kelulusan tepat waktu mahasiswa apakah mahasiswa tersebut lulus tepat waktu atau tidak.

Tabel 1. Data *Training*

Nim	Semester				Y = Klasifikasi
	1	2	3	4	
J1F106201	2.39	2.61	2.74	3	Ya
J1F106202	2.77	2.42	2.86	2.21	Ya
J1F106206	2.98	3.63	3.39	3.59	Ya
..	..	..	..	..	..
..	..	..	..	..	..
J1F106215	2.91	1.84	2.66	2.62	Tidak
J1F106216	2.43	1.39	2.39	1.85	Tidak

Keterangan :
1 : Nilai IP Semester 1
2 : Nilai IP Semester 2
3 : Nilai IP Semester 3
4 : Nilai IP Semester 4
Ya : Tepat Waktu (<= 5 Tahun)
Tidak : Tidak Tepat Waktu (> 5 Tahun)

Setelah pembuatan data *training* maka kita perlu adanya data *testing* berupa data IP mahasiswa yang akan diklasifikasikan, apakah lulus <= 5 tahun atau lulus > 5 tahun.

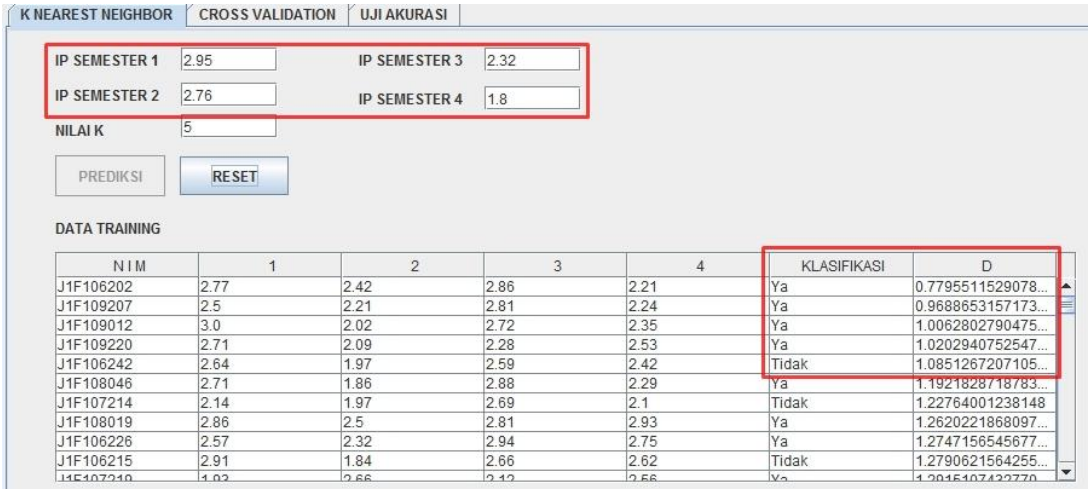
Tabel 2. Data *Testing*

N	Semester				Y= Klasifikasi
	1	2	3	4	
1	2.95	2.76	2.32	1.8	?

Data *testing* merupakan sekumpulan data IP mahasiswa yang akan diklasifikasikan dengan data *training*, apakah lulus <= 5 tahun atau lulus >5 tahun. Setelah ada data *testing* dan data *training*, lalu menentukan nilai k-nya, nilai k yang digunakan berdasarkan hasil pencarian k-Optimal didapatkan k=5. Selanjutnya adalah menghitung kuadrat jarak euclidean masing-masing objek
terhadap data training yang diberikan dengan menggunakan rumus perhitungan jarak : $d_I = \sqrt{\sum_{i=1}^{p}(x_{2i} - x_{1i})^2}$

N
1 $\sqrt{(2.39-2.95)^2+(2.61-2.76)^2+(2.74-2.32)^2+(3-1.8)^2} = 1.9525$
2 $\sqrt{(2.77-2.95)^2+(2.42-2.76)^2+(2.86-2.32)^2+(2.21-1.8)^2} = 0.6077$
3 $\sqrt{(2.98-2.95)^2+(3.63-2.76)^2+(3.39-2.32)^2+(3.59-1.8)^2} = 2.2598$
.. ..
.. ..
109 $\sqrt{(2.91-2.95)^2+(1.84-2.76)^2+(2.66-2.32)^2+(2.62-1.8)^2} = 1.636$
110 $\sqrt{(2.43-2.95)^2+(1.39-2.76)^2+(2.39-2.32)^2+(1.85-1.8)^2} = 2.1547$

Dari keseluruhan hasil perhitungan jarak dengan data training yang berjumlah 110 data, hasil perhitungan diurutkan mulai terkecil hingga terbesar. Setelah diurutkan, dilihat mayoritas klasifikasi yang muncul dari perhitungan jarak yang terkecil atau yang pertama sampai dengan yang kelima. Dapat dilihat pada gambar mayoritas klasifikasi yang muncul adalah "Ya", sehingga data testing yang ingin diklasifikasikan termasuk kedalam klasifikasi "Ya" atau tepat waktu.



NIM	1	2	3	4	KLASIFIKASI	D
J1F106202	2.77	2.42	2.86	2.21	Ya	0.7795511529078...
J1F109207	2.5	2.21	2.81	2.24	Ya	0.9688653157173...
J1F109012	3.0	2.02	2.72	2.35	Ya	1.0062802790475...
J1F109220	2.71	2.09	2.28	2.53	Ya	1.0202940752547...
J1F106242	2.64	1.97	2.59	2.42	Tidak	1.0851267207105...
J1F108046	2.71	1.86	2.88	2.29	Ya	1.1921828718783...
J1F107214	2.14	1.97	2.69	2.1	Tidak	1.22764001238148
J1F108019	2.86	2.5	2.81	2.93	Ya	1.2620221868097...
J1F106226	2.57	2.32	2.94	2.75	Ya	1.2747156545677...
J1F106215	2.91	1.84	2.66	2.62	Tidak	1.2790621564255...

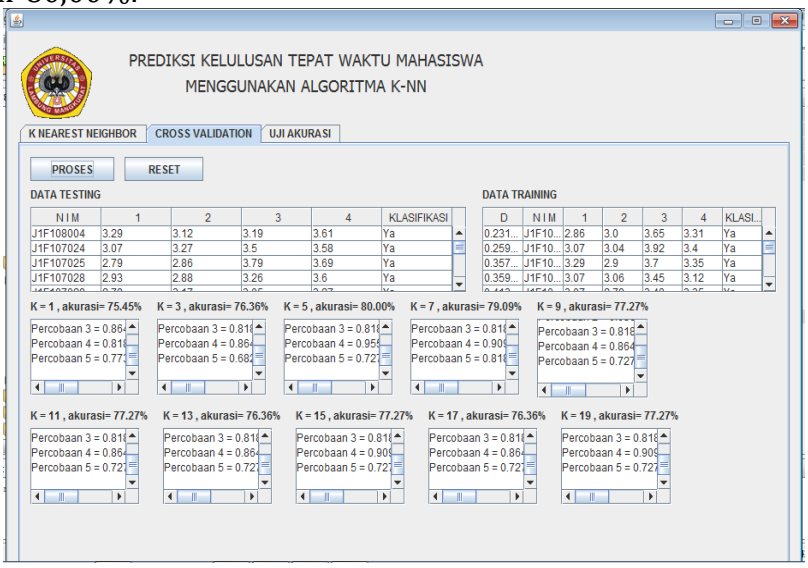
Gambar 7. Penentuan Hasil Prediksi dari Perhitungan Jarak

3.5. *Evaluation*

Pada tahap ini dilakukan pencarian k-Optimal pada algoritma kNN untuk prediksi kelulusan tepat waktu mahasiswa berdasarkan IP sampai dengan semester 4 menggunakan metode k-Fold Cross Validation dan Uji Akurasi. Setelah mendapatkan nilai k-Optimal maka dapat digunakan pada algoritma kNN untuk prediksi kelulusan tepat waktu mahasiswa berdasarkan IP sampai dengan semester 4.

3.5.1. k-Fold Cross Validation

Form *k-Fold Cross Validation* menghasilkan tingkat akurasi dari masing-masing nilai k yang digunakan yaitu 1,3,5,7,9,11,13,15,17, dan 19. Dari nilai k tersebut didapatkan bahwa hanya k=5 yang menghasilkan persentase tertinggi yaitu sebesar 80,00%.



Gambar 8. Hasil Pencarian k-Optimal dengan k-Fold Cross Validation

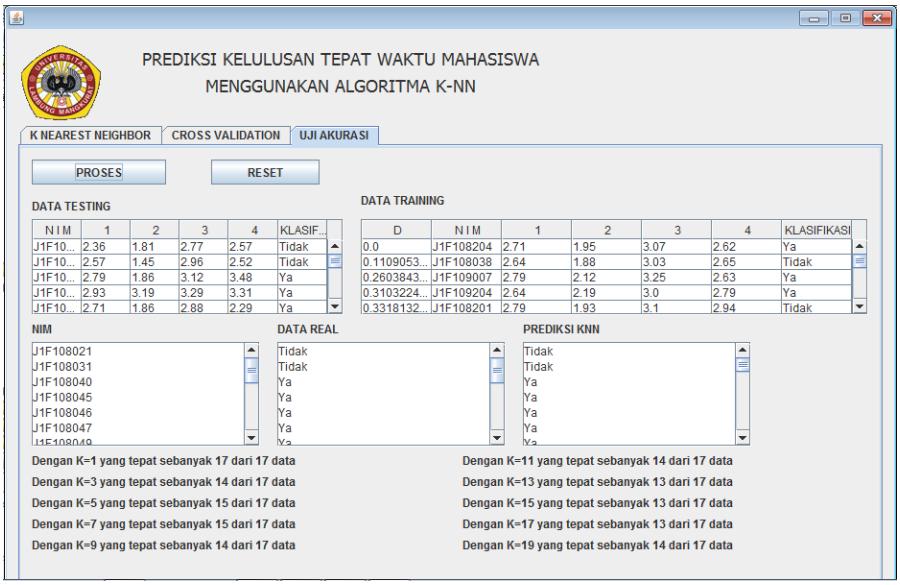
Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

3.5.2. Uji Akurasi

Form uji akurasi ini menggunakan data sebanyak 17 buah yang diambil dari *data training* untuk dilakukan perbandingan antara klasifikasi yang ada di data real dan hasil prediksi kNN, lalu dihitung berapa jumlah data yang tepat diprediksi oleh algoritma kNN. Berdasarkan hasil uji akurasi diketahui bahwa k=1 menghasilkan ketepatan sebanyak 17 dari 17 data.

Menurut Banjarsari (2015), hal ini disebabkan karena pada tahap uji akurasi, data yang diuji berasal dari data training itu sendiri sehingga jika menggunakan nilai k=1 yang mana hasil prediksi tersebut berasal dari perhitungan jarak yang hasil nya 0,0 yang didapatkan dari data yang diujikan tersebut. Sehingga dari perhitungan jarak yang terkecil tersebut jika menggunakan nilai k=1 maka hasilnya sama antara klasifikasi di data real dan prediksi kNN, itulah yang menyebabkan k=1 ini menghasilkan ketepatan sebanyak 17 dari 17 data. Dapat diambil kesimpulan bahwa nilai k=1 ini bukan satu-satunya faktor yang menentukan pemilihan k-Optimal, karena pada k=1 perhitungan jarak dan penentuan hasil prediksi dipengaruhi oleh perhitungan data. Selain itu, dari uji akurasi didapatkan nilai k=5 dan k=7 sama-sama menghasilkan ketepatan sebanyak 15 dari 17 data. Untuk memutuskan satu nilai k terbaik atau k-Optimal maka dilihat dari tingkat akurasinya dari kedua nilai k tersebut pada hasil pencarian menggunakan metode *k-Fold Cross Validation.* Didapatkan bahwa k=5 yang memiliki tingkat akurasi sebesar 80,00% sedangkan k=7 sebesar 79,09%.

Nilai k=5 inilah yang ditetapkan sebagai *k-Optimal* oleh peneliti, yang selanjutnya dapat digunakan untuk memprediksi kelulusan tepat waktu mahasiswa dengan variabel yang digunakan yaitu IP Semester 1, IP Semester 2, IP Semester 3, dan IP Semester 4.

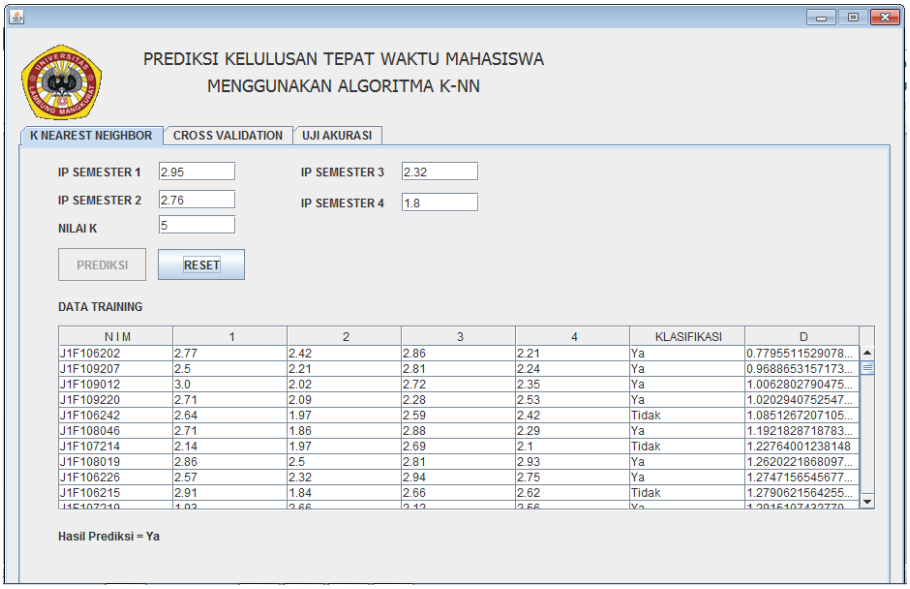


Gambar 9. Hasil Pencarian k-Optimal dengan Uji Akurasi

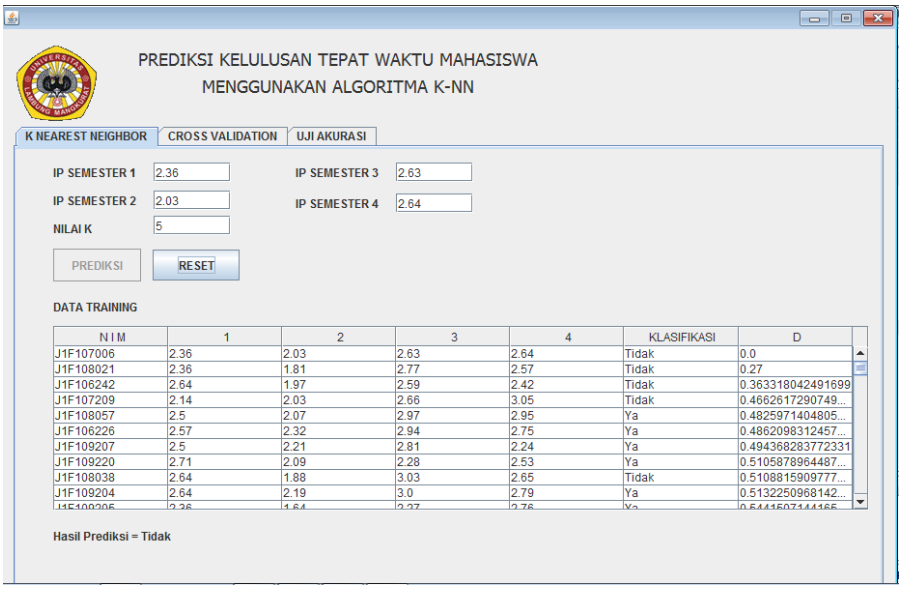
Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

3.5.3. Prediksi k-Nearest Neighbor

Pada tahap k-Fold Cross Validation dan Uji Akurasi telah didapatkan nilai k-Optimal yaitu k=5 yang digunakan sebagai nilai k pada algoritma kNN untuk prediksi kelulusan tepat waktu mahasiswa berdasarkan IP sampai dengan semester 4.



Gambar 10. Hasil Prediksi menggunakan data IP sembarang



Gambar 11. Hasil Prediksi menggunakan data IP dari data *training*

Sumber : Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waku Mahasiswa Berdasarkan IP Sampai Dengan Semester 4 (Studi Kasus : Mahasiswa Program Studi Ilmu Komputer FMIPA Unlam). 2015

4. KESIMPULAN

Kesimpulan yang diperoleh dari penelitian ini adalah:

a. Nilai k-Optimal pada algoritma kNN untuk prediksi kelulusan tepat waktu mahasiswa berdasarkan IP sampai dengan semester 4 adalah k=5.
b. Dari proses *k-Fold Cross Validation* didapatkan tingkat akurasi untuk k=5 pada algoritma kNN untuk prediksi kelulusan tepat waktu mahasiswa berdasarkan IP sampai dengan semester 4 adalah sebesar 80,00%.

DAFTAR PUSTAKA

[1] Banjarsari, Mutiara A. 2015. ***Pencarian k-Optimal pada Algoritma kNN untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Berdasarkan IP Sampai Dengan Semester 4*.** FMIPA Unlam : Banjarbaru

[2] Darmawan, Astrid. 2012. ***Pembuatan Aplikasi Data mining untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighborhood*.** Fakultas Teknik Dan Ilmu Komputer Universitas Komputer Indonesia : Bandung

[3] Fu L. 1994. ***Neural Network In Computer Intelligence***. Singapura : McGraw Hill.

[4] Han, J.,&Kamber, M. 2006. ***Data mining Concept and Tehniques***. San Fransisco : Morgan Kauffman.

[5] Hastie Trevor, Tibshirani Robert, dan Jerome Friedman. 2008. ***The Elements of Statistical Learning Data Mining, Inference, and Prediction***. California : Springer.

[6] Larose, D. T. 2005. ***Discovering Knowledge in Data***. New Jersey : John Willey & Sons, Inc.

[7] Liu, B., 2007 . ***Web Data mining: Exploring Hyperlinks, Contents, dan Usage Data*.** Berlin: Springer.

[8] Moertini, V. S. 2002. ***Data mining* sebagai solusi bisnis**. Integral, vol 7 no.1.

[9] Pandie, Emerensye S. Y. 2012. ***Implementasi Algoritma Data mining K-Nearest Neighbour (KNN) Dalam Pengambilan Keputusan Pengajuan Kredit*.** Jurusan Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana : Kupang

[10] Rizal, Azwar. 2013. ***Perbandingan Performa antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma Mutual Nearest Neighbor***. Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember. Surabaya.

[11] Rodiyansyah, S. Fajar dan Edi Winarko. 2013. ***Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification.*** Yogyakarta : Universitas Pendidikan Indonesia.

[12] Wu X, Kumar V. 2009. ***The Top Ten Algorithms in Data Mining***. New York: CRC Press.