

## **PREPROCESSING DATA UNTUK SISTEM PERAMALAN TINGKAT KEDISIPLINAN MAHASISWA**

Henderi<sup>1</sup>  
Rizal Loa Wanda<sup>2</sup>

Email : *henderi@raharja.info* , *rizal.loa@raharja.info*

### **ABSTRAK**

*Data mining merupakan sebuah proses penggalian atau penemuan informasi baru dengan mencari pola tertentu dari sejumlah besar database yang telah tersedia. Teknik ini dapat membantu dalam pemanfaatan kembali data tersebut. Penelitian ini bermaksud melakukan penggalian informasi terhadap data absensi online di Perguruan Tinggi Raharja. Penggalian informasi dilakukan untuk meramalkan tingkat kedisiplinan mahasiswa. Penelitian ini bertujuan untuk mendapatkan set data yang berkualitas dan siap untuk dilakukan proses data mining untuk digunakan dalam proses peramalan tingkat kedisiplinan mahasiswa. Penelitian ini dilakukan dengan tahapan melakukan preprocessing data yang terdiri dari pembersihan, ekstraksi, transformasi data, dan pemilihan atribut. Preprocessing dilakukan terhadap set data absensi online semester genap 2014/2015 dan set data induk mahasiswa. Hasil preprocessing mendapatkan, 1.836 record mahasiswa. Setiap record berisi variabel y sebagai atribut target yaitu kedisiplinan mahasiswa, dan 8 variabel x sebagai atribut prediktor yang diasumsikan memiliki pengaruh terhadap variabel y.*

**Kata Kunci:** *Data mining, preprocessing, peramalan, tingkat kedisiplinan mahasiswa*

### **PENDAHULUAN**

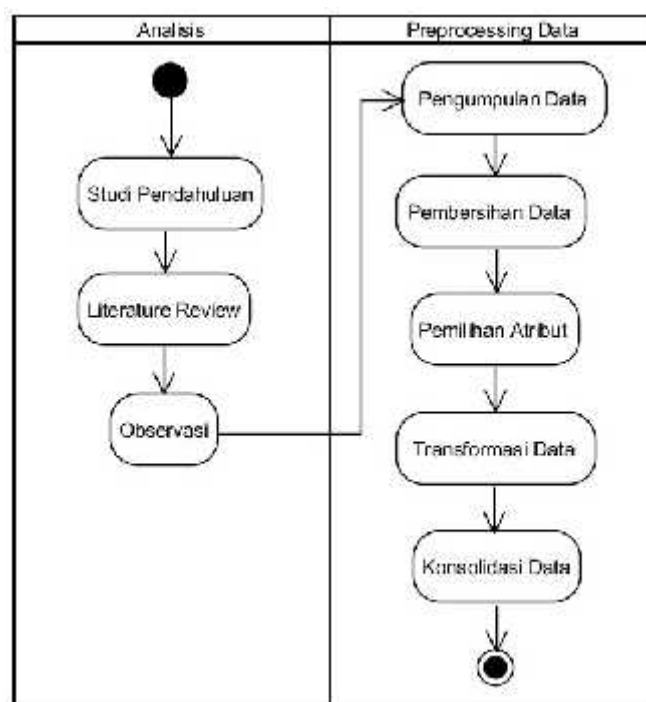
Perkembangan informasi di era digital ini terus meningkat setiap waktunya. Dampak perkembangan informasi tersebut ikut memberikan pengaruh yang besar terhadap peningkatan volume data yang disimpan dalam database. Data yang tersimpan setiap harinya menyebabkan penumpukan data yang besar. Jika hal tersebut tidak dimanfaatkan, maka data tersebut hanya akan menjadi data yang tidak terpakai di media penyimpanan. Tan mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar [1]. Dengan *data mining*, penumpukan data dalam *volume* besar dapat dimanfaatkan kembali sebagai penunjang keputusan ataupun untuk keperluan prediksi di masa depan. Namun pada nyatanya, data dalam *volume* besar adalah hal yang kotor. [2] *Preprocessing* diperlukan karena data dunia nyata umumnya tidak lengkap (kurang nilai atribut, kurang atribut tertentu yang menarik, atau hanya berisi data agregat), kotor / *noise* (mengandung kesalahan atau *outlier*), dan tidak konsisten (mengandung perbedaan dalam kode atau nama). Hal ini memiliki pengaruh terhadap hasil dari proses data mining. Kualitas data yang digunakan untuk proses *data mining*, berbanding lurus dengan hasil yang akan didapat.

Data Absensi Online (AO) di Perguruan Tinggi Raharja, adalah data yang cukup potensial untuk digali pengetahuannya. Data yang terdapat di AO dapat digunakan kembali sebagai salah satu sumber data dalam proses data mining untuk meramalkan tingkat kedisiplinan mahasiswa. Mengingat data dalam dunia nyata adalah hal yang

kotor begitu juga dengan data AO, maka diperlukannya sebuah proses yang dapat meningkatkan kualitas dari data yang akan diolah. Salah satu proses yang ada dalam *Knowledge Discovery in Database*(KDD) dapat mengatasi masalah tersebut, proses itu adalah *Preprocessing* data.

## METODOLOGI PENELITIAN

*Preprocessing* data untuk sistem peramalan tingkat kedisiplinan mahasiswa dibagi menjadi dua tahap aktivitas utama. Aktivitas pertama yaitu analisis yang terdiri dari studi pendahuluan, literature review, dan observasi. Observasi dilakukan di Perguruan Tinggi Raharja pada bagian RME yang terkait dengan objek penelitian yaitu data absensi online. Sedangkan aktivitas kedua adalah *preprocessing* data yang terdiri dari pengumpulan sumber data, pembersihan data, pemilihan atribut, transformasi data dan konsolidasi data yaitu replikasi data dari lebih dari set data untuk menjadi set data yang baru.



**Gambar 1.** Diagram Alir Metodologi Penelitian

## LITERATURE REVIEW

Terdapat beberapa penelitian terkait mengenai penelitian ini. Diantaranya yang dilakukan oleh Astuti, dkk [3] mengenai prediksi siswa yang berpotensi melakukan ketidakdisiplinan menggunakan Algoritma *Naïve Bayes Classifier*. [3] Penelitiannya menggunakan data asli siswa yang berjumlah 920 *record* sebagai sumber data dan atribut yang dipilih berjumlah 16 diantaranya 15 merupakan atribut prediktor dan 1 atribut target (ketidakdisiplinan). Sedangkan Mustafidah[4] menggunakan model Regresi untuk mengetahui hubungan antara motivasi belajar dengan tingkat kedisiplinan

agar dapat diprediksi. Sumber data dan atribut yang digunakan pada penelitiannya[4] berasal dari angket yang disebar pada 127 mahasiswa dan atribut yang digunakan hanya 2 yaitu motivasi belajar (variabel x) dan tingkat kedisiplinan (variabel y).

Penelitian terkait berikutnya yaitu penelitian Elisa [5] mengenai faktor-faktor yang mempengaruhi kedisiplinan. Objek penelitiannya [5] adalah kedisiplinan kerja karyawan yang bekerja di PT Suka Fajar Pekanbaru dengan sumber data berasal dari wawancara dan kuesioner dengan sampel sebanyak 74 orang responden menggunakan teknik cluster sampling. Penelitiannya [5] menjelaskan bahwa variabel bebas (*independent variable*) adalah faktor kompensasi, sanksi hukum dan kepemimpinan, sedangkan variabel terikatnya (*dependent variable*) adalah kedisiplinan kerja karyawan. Tujuan penelitiannya [5] adalah untuk mengetahui faktor-faktor yang mempengaruhi kedisiplinan kerja karyawan, dan untuk mengetahui faktor dominan yang mempengaruhi kedisiplinan kerja. Namun pada penelitiannya belum melalui proses data mining, yang sebenarnya dapat dikembangkan kembali menggunakan data mining. Sebagai contoh untuk meramalkan kedisiplinan kerja, sehingga tidak terbatas hanya untuk mengetahui faktor yang mempengaruhi kedisiplinan saja.

Penelitian berbeda telah dilakukan Ginting, dkk [6], telah membahas mengenai prediksi masa studi mahasiswa berdasarkan data nilai akademik. Penelitiannya menggunakan pendekatan yang ada di KDD, dan juga Algoritma C4.5 dalam penentuan masa studi mahasiswa, serta pohon keputusan (*decision tree*) untuk melihat kemungkinan mahasiswa yang lulus lebih dari 8 semeseter. Serupa dengan penelitian sebelumnya[6], di penelitian Trayasiwi [7] juga membahas mengenai prediksi kelulusan mahasiswa dan menggunakan pendekatan yang ada di KDD. Namun perbedaannya terletak pada metode yang digunakannya yaitu menggunakan metode klusterisasi serta algoritma *K-Means*. [7] berpendapat bahwa data yang telah diklusterisasi tersebut menghasilkan kategori prediksi kelulusan mahasiswa berdasarkan lama atau tidaknya waktu kelulusan dan tinggi rendah IPK yang diperoleh mahasiswa pada setiap kluster.

## PEMBAHASAN

### A. Sumber Data

Pada penelitian ini, sumber data yang digunakan berasal dari dua sumber yang berbeda. Sumber pertama yaitu data Absensi Online (selanjutnya disebut AO) yang sudah tidak digunakan kembali (data historis). Pada studi kasus ini, data AO yang digunakan yaitu data pada semester genap tahun akademik 2014/2015 dengan hanya mengambil perkuliahan kelas teori yang memiliki bobot 2 dan 3 sks. Sumber data AO ini terdiri dari 338 kelas teori dengan tenaga pengajar sebanyak 119 dosen. Dan setiap data kelas tersebut (Gambar 2) berisi *record* waktu kehadiran mahasiswa disetiap pertemuan dengan jumlah maksimal 14 kali pertemuan disetiap semesternya.

Sumber data kedua yaitu data mahasiswa yang terdapat di SiS+ (*Students iLearning Services*) Perguruan Tinggi Raharja. Sumber data kedua ini diambil dari *web service* yang telah *on-line* (Gambar 3), sehingga data tersebut dapat diambil dimana saja dan kapan saja selama terhubung dengan internet.

SI200A - Keamanan Komputer - 11 30 17 10 - 2 Sks  
02012 - Maimunah, M.Kom

		Perganti: Tidak Tidak Tidak Tidak Tidak Tidak Tidak Tidak Tidak Tidak Tidak Tidak Tidak														
		Check In: 15:31 12:34 15:02 14:58 15:37 8:05 7:54 12:40 12:12 15:29 12:21 15:18 12:43														
		Check Out: 17:10 17:10 17:10 17:10 17:10 17:10 17:10 17:10 17:10 17:10 17:10 17:10 17:10														
No	NIM	Nama	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
1	1431431643	Aditya Abdol Hafim	15:36	15:32	15:41	15:43	15:42	15:34	15:41	15:35	15:39	15:45	15:40	15:42	15:28	
2	1222874232	Aditya Giyartono	15:35		15:35	15:37	16:20	15:34	15:41	15:32	15:44	15:35	15:30		15:28	
3	1222474124	Aqas Gunawan	15:36	15:42	15:35	15:37	15:56	15:34	15:41	15:32	15:44	15:32	15:44	15:46	15:28	
4	1422477893	Alfa Stefan	15:36		15:33	15:43		15:34			15:53	15:51				
5	1431431499	Alifio Gaih Wicaksono	15:36	15:32	15:41	15:43	15:42	15:34	15:42	15:36	15:38	15:46	15:41	15:37	15:28	
" -- data -- "																
46	1121459672	Tiara Nurhidia														
47	1431431191	Tommy Joppy Shite	15:40	15:31	15:37	15:46	15:48	15:55	15:45	15:34	15:44		15:41	15:38	15:30	
48	1431431371	Wardi Supriyadi	15:40	15:31	15:34	15:43	15:48	15:35	15:43	16:01	15:43	15:35		15:30	15:30	
49	1431333224	Wawan	15:40	15:31	16:06	15:35	16:00				15:45	15:51		16:21		
50	1431333206	Yasinta Adnanah	15:42	15:31	15:50	15:55	15:48	15:35	15:45	15:35	15:43	15:49	15:31	15:32	15:31	

Gambar 2. Sumber Data AO

Daftar WEB SERVICE		Struktur tbl_TMMasiswa	
Show 10 entries Search: <input type="text"/>		Show: Count   Data / Jeda	
No	Nama Tabel	No	Nama Field
1	tbl_MUlayaKuliah	1	NIM
2	tbl_MDosen	2	NIPR
3	tbl_MI orm	3	NamaDipari
4	tbl_MGolonganDosen	4	NamaRelakang
5	tbl_MGrading	5	NamaPanggilan
6	tbl_MIlilan	6	TempatLahir
7	tbl_MJabatan	7	Tanggalahir
8	tbl_MJadwalKURS	8	KIP
9	tbl_MJenisKelas	9	KodeNegara
10	tbl_MJenisRuang	10	JenisKelamin
		11	StatusPerkawinan

Gambar 3. Sumber Data Mahasiswa dari Web Service Raharja

## B. Pembersihan Data

Pembersihan data pada studi kasus ini diantaranya adalah pembersihan data kelas dan pembersihan *record* mahasiswa. Pembersihan kelas dilakukan jika terdapat kelas yang diduga abnormal maka kelas tersebut dihapus dan pembersihan *record* mahasiswa jika mahasiswa tersebut tidak hadir satu kali pun selama 14 pertemuan perkuliahan.

## 1. Pembersihan Kelas

Ada 338 kelas teori yang digunakan dalam proses pembersihan, terdapat 4 kelas yang dikategorikan sebagai data anomali data. Diidentifikasi sebagai data anomali karena absensi penuh disetiap pertemuan dan waktu absensi ada di waktu yang sama.

AT 10171 - Akuntansi 1 - 08/00-09/16 - 1 Sks

No	NIM	Nama	Perganti													
			Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	Tidak	
Check In			8:00	7:50	7:48	7:32	8:17	7:53	7:55	7:39	7:40	7:41	8:00	7:49	7:51	8:00
Check Out			9:10	9:10	9:10	9:10	9:10	9:10	9:10	9:10	9:10	9:10	9:10	9:10	9:10	9:10
			I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
1	1312483749	Anis Khoiratul Misa	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:49	7:41	8:11	8:00	8:25	8:24	8:00
2	1412483054	Ayu Irfan Mardevani	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:49	7:41	8:11	8:00	8:25	8:24	8:00
3	1412483121	Chaeenisa Savitri	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:49	7:41	8:11	8:00	8:25	8:24	8:00
4	1312476375	Chelsia Indria Pranasari	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:49	7:41	8:11	8:00	8:25	8:24	8:00
5	1112483086	Dery Andika	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:49	7:41	8:11	8:00	8:25	8:24	8:00
6	1112483301	Devi Septiana	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:25	8:24	8:00
7	1312477393	Dian Triyandayu	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:25	8:24	8:00
8	1312483673	Dimas Sanggul Palanra	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:25	8:24	8:00
9	1412483208	Faelan Ardiansyah	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:25	8:24	8:00
10	1412483245	Faisal Haris Ridwan	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:25	8:24	8:00
11	1412483353	Faris Supriyansih	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:24	8:00
12	1312483750	Fani Yunita Andia	8:12	8:08	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:24	8:00
13	1112483051	Hanika Yuzman	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:24	8:00
14	1312476371	Haniroo Widjaja	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:24	8:00
15	1412483479	Jessica Kristasari	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:24	8:00
16	1412483514	Manny Huzamzah	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:24	8:00
17	1412483206	Mika Tyuliana Kanna	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:25	8:00
18	1312483522	Muhammad Arif Nordin	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:25	8:00
19	1412483655	Nar Farida	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:25	8:00
20	1412483899	Nar Taswiah	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:50	7:41	8:11	8:00	8:27	8:25	8:00
21	1112477956	Nar Rabhat Aditya Pratama	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:51	7:41	8:11	8:00	8:27	8:25	8:00
22	1112482230	Nurhalikah	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:51	7:41	8:11	8:00	8:27	8:25	8:00
23	1312483532	Pandi Susana	8:12	8:09	8:00	8:00	8:00	8:16	8:15	7:51	7:41	8:11	8:00	8:27	8:25	8:00
24	1412483393	Priyadita Bima Henyandio	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:51	7:41	8:11	8:00	8:27	8:25	8:00
25	1312477021	Reni Walandari	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:51	7:41	8:11	8:00	8:27	8:25	8:00
26	1412483921	Rizka Permata Devi	8:12	8:09	8:00	8:00	8:00	8:16	8:14	7:51	7:41	8:11	8:00	8:27	8:25	8:00
27	1412483017	Sita Ayu Iestari	8:12	8:09	8:00	8:00	8:00	8:16	8:15	7:51	7:41	8:11	8:00	8:27	8:25	8:00
28	1412483023	Tjati Sujarna	8:12	8:09	8:00	8:00	8:00	8:16	8:15	7:51	7:41	8:11	8:00	8:27	8:25	8:00
29	1112483011	Vani Nurmalasari	8:12	8:09	8:00	8:00	8:00	8:16	8:15	7:51	7:41	8:11	8:00	8:27	8:25	8:00

Gambar 4. Contoh Kelas yang Abnormal

Gambar 4 adalah contoh data anomaly karena absensi terisi penuh selama 14 kali pertemuan dengan waktu absen bersamaan. Data ini diragukan kebenaran absensinya, dan jika data kelas tersebut digunakan dapat mengakibatkan ketidakakuratan pada hasil dari proses data mining.

## 2. Perhitungan Variabel Y dan Pembersihan Record Mahasiswa

Perhitungan atau *counting* variabel y (kehadiran mahasiswa) diperlukan untuk mengetahui jumlah kehadiran mahasiswa dikelas. Selain untuk pembersihan *record* mahasiswa, perhitungan variabel y juga akan digunakan untuk menghasilkan rata-rata kehadiran yang akan disebut kedisiplinan mahasiswa. Pada studi kasus ini telah ditentukan untuk pembersihan *record* mahasiswa adalah jika pada *record* mahasiswa dikelas tersebut 0 (nol) kehadiran (Gambar 5) selama 14 kali pertemuan, maka *record* tersebut dihapus dan tidak digunakan.

26	1303385381	Muhammad Fauzan Al-Muhammad	15:42			15:48	15:49	15:51	15:44				15:52	15:52	15:49	8
27	1303437207	Rower Adhya Primaya	15:36	15:39	15:37	15:35	15:37	15:34	15:44	15:34	15:4	15:35	15:34	15:36	15:30	12
28	1303474235	Rizki Izzah	15:34	15:43	15:36	15:43	15:35	15:34	15:44	15:33	15:4	15:33	15:31	15:30	15:30	12
29	1303381134	Pradko Solwinayah Muband	15:34		15:05	15:14	15:39				15:4	15:50	15:57	15:53	15:50	9
30	1303477328	Rahman Nurvaika	15:36	15:36	15:49	15:43										4
41	1303475431	Riza Zulfan Pramaia	15:34	15:43	15:36	15:35	15:47	15:35	15:44	15:33	15:43	15:33	15:31	15:30	15:30	12
42	1303481476	Ridho Fudji Anugrah Mulyah														0
43	1303481477	Ridwan Muli Ramadji	15:34	15:34	15:32	15:43	15:48	15:50	15:45	15:36	15:40	15:35		15:38	15:30	12
44	1302475286	Rizki Rianda		15:35	15:44	15:43		15:35	15:45	15:39	15:42	15:45			15:30	9
75	1303381122	Sawing Mita Nur Ustama	15:45	15:41	15:39	15:45	15:48	15:45	15:45	15:37	15:43	15:51	15:35	15:42	15:49	13
76	1303481422	Sara Nurinda														0
17	1303481420	Tomy Lippy Elhan	15:40	15:32	15:32	15:36	15:38	15:33	15:45	15:34	15:45		15:41	15:38	15:30	12

\* Record Berwarna Merah Akan Dihapus

Gambar 5. Contoh Record Mahasiswa yang 0 Kehadiran

### 3. Penggabungan Record Untuk Mencari Rata-Rata Variabel Y

Satu mahasiswa dapat mengambil banyak mata kuliah (kelas) disetiap semester, sehingga hal yang perlu dilakukan adalah penggabungan (*Grouping*) dari setiap kehadiran mahasiswa (variabel y) di kelas-kelas mata kuliah sesuai dengan NIM dan nama mahasiswa.

NIM	Nama	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	A
1302475733	A. Andri Setianto	17:25	17:18	17:25		18:55	17:24	17:55	18:48	18:37		17:10					9
1302475733	A. Andri Setianto	17:53	17:31	17:18	18:40	17:37	18:40	17:25	18:15	17:55	17:23		17:10	17:10	17:10		13
1302475733	A. Andri Setianto	18:33	18:22	18:24	18:20	18:29	18:22	18:38	18:11	18:20	18:32	18:43	18:20	18:12			13
1302475733	A. Andri Setianto	18:39	17:40	17:37	18:27	18:32	18:31	17:14			17:46						8
1302475733	A. Andri Setianto		20:21	19:20	19:38	20:19	20:10	19:36	19:21			19:44	20:07				9
1302475733	A. Andri Setianto	17:10	17:10	17:53	17:42	17:32	18:30	17:30	18:10	19:34	17:54	17:21	18:16	18:12			18
1302475733	A. Andri Setianto	19:21	17:39	17:20	17:33	18:48	17:30	17:44	18:49	19:04	17:52	18:47	18:50				12
1302475733	A. Andri Setianto	19:20	20:18	19:28	20:31	20:27	19:30	19:44	20:20			19:42		19:37	19:36		11
1302477313	Aap Kurtubi	13:00	12:58	13:03	12:59	13:07	12:48	13:04	14:19	13:07	13:00	13:09	12:56				12
1302477313	Aap Kurtubi		13:07	13:04		13:41	13:44	13:02	13:01	13:06	12:52	14:26	13:29			13:00	11
1302477313	Aap Kurtubi	8:35	8:30	9:15	8:29	8:20	8:52	9:46	8:00	8:27	8:00	8:00	9:09	9:15			18
1302477313	Aap Kurtubi	13:00	13:08	13:12		13:21	13:20	13:18	13:00	12:59	13:04	13:24		14:07			11
1302477313	Aap Kurtubi		11:20	11:43	11:33	11:21	11:35	11:47	11:32	11:28	12:03	11:52		11:54	12:03		19
1302477313	Aap Kurtubi	15:38	15:41	15:37	15:45	15:33	15:33	15:44	15:38	15:40	15:30	15:27	15:26	15:41			18
1302477313	Aap Kurtubi	8:32	8:28	8:22	8:36	8:25	8:46	8:26	8:24	8:24	8:27	8:00	8:45	8:52			19
1302477313	Aap Kurtubi	9:49	10:04	10:05	10:02	9:51	10:18	10:23	9:57	10:03	10:22	9:57	10:03				19

Gambar 6. Sebelum Penggabungan Record

Dari Gambar 6 dapat dilihat bahwa, disetiap *record* adalah *record* mahasiswa yang terdapat di 1 kelas mata kuliah. Sebagai contoh mahasiswa dengan nim 1302475733 memiliki 8 variabel y, dengan kata lain mahasiswa tersebut pada semester ini mengambil 8 mata kuliah. Sehingga setelah proses *grouping* maka hasilnya adalah setiap mahasiswa memiliki nilai rata-rata kehadiran / rata-rata variabel y ( ).

NIM	NAMA	y1	y2	y3	y4	y5	y6	y7	y8	y9	y10	y11	̄y
1322475733	A. Andri Setianto	9	13	13	8	9	13	12	11				11.00
1322477313	Aap Kurtubi	12	11	13	11	12	13	12	12				12.00
1322476303	Abdul Azis	10	10	11	13	9	10	12					10.71
1414481555	Abdul Azis	13	14	14	14	12	12	12	14				12.78
1411481283	Abdul Azis	11	10	10	10	11	9	13	13	12			11.00
1133168559	Abdul Fatah	10	11	11	13	10							11.00
1322477015	Abdul Haqy Aji Prastian	12	11	13	13	13	14	12	13	14			13.11
1222474392	Abdul Mujiid	12	10	13	12	12	11	11	12				11.63
1321476870	Abdul Mukarota	12	8	10	8	13	6	7					9.14
1433451330	Abdul Rizki Zarkasy	14	13	13	13	13	13	12					13.00
1011476941	Abdul Rosid	13	9	12	8	7							9.80
1413481646	Abdul Waci Rahmat	13	12	13	12	14	14	14					13.14
1212473717	Abdurrahman Muhammad Chusani	10	9	8	9	9	10	9	8				9.00
1311483314	Abdurrauf Rafli	13	11	9	9	12	9	11	8				10.25

Gambar 7. Hasil Dari Penggambaran dan Nilai

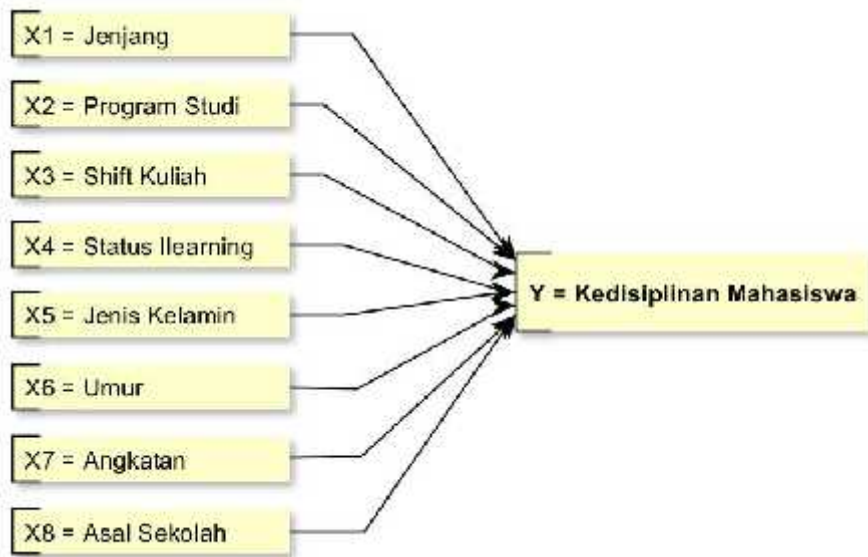
### C. Pemilihan Atribut

Pemilihan atribut atau variabel x harus sesuai dengan pertimbangan bahwa atribut tersebut terdapat pengaruh dan relevan terhadap variabel y. Dalam penelitian ini, telah dipilih beberapa atribut yang diasumsikan berpengaruh terhadap tingkat kedisiplinan mahasiswa (Tabel 1), diantaranya Jenjang kuliah, program studi, shift kuliah, status pekerjaan, status ilearning, jenis kelamin, umur, angkatan, dan asal sekolah. Namun saat penarikan data untuk atribut status pekerjaan, data yang didapat kurang mencukupi, sehingga atribut tersebut tidak dapat digunakan.

Tabel 1. Variabel X

X	Atribut
X1	Jenjang Kuliah
X2	Program Studi
X3	Shift Kuliah
X4	Status iLearning
X5	Jenis Kelamin
X6	Umur
X7	Angkatan
X8	Asal Sekolah

Dan jika dilihat dari hubungannya (Gambar 8), variabel y dengan variabel x memiliki hubungan asimetris lebih dari dua variabel atau dapat disebut hubungan variabel *multivariat*.



**Gambar 8.** Hubungan Antar Variabel Multivariat

#### D. Penarikan Data Untuk Variabel X

Penarikan data untuk variabel x bersumber data yang berbeda dengan set data AO. Sumber data tersebut didapat dari *web service* yang disediakan oleh Perguruan Tinggi Raharja untuk keperluan riset mahasiswa. Format pertukaran data yang disediakan berupa JSON (*JavaScript Object Notation*) sehingga perlu dilakukan ekstraksi data.

#### Ekstraksi Data

Dikarenakan format data yang digunakan berbeda dari sumber data maka harus ada perubahan format serta dilakukan penyimpanan sementara hasil ekstraksi untuk penggabungan dari sumber data lain (beda tabel).

```
[{"NIM": "1322476037", "NIPR": "", "NamaDepan": "Rizal", "NamaBelakang": "Lca Wanda", "NamaFanggilan": "Rizal", "TempatLahir": "Tangerang", "TanggalLahir": "1994-12-20", "KTP": "", "KodeNegara": "ID", "JenisKelamin": "L", "statusPerkawinan": "0"....}]
```

Json Tabel tbl\_TTMahasiswa

```
[{"NIM": "1322476037", "Jenjang": "400", "Jurusan": "420", "Konzentrasi": "422", "KodeKurikulum": "2245", "ShiftKuliah": "1", "Ilearning": "0", "Ilp": "0"}]
```

Json Tabel tbl\_TTMahasiswaJurusan

```
[{"KodePendFormal": "2234", "NIM": "1322476037", "JenjangPendidikan": "SMK", "NamaInstitusi": "SMK NEGERI 1 TANGERANG", "Jurusan": "MULTIMEDIA", "TanggalMasuk": "0000-00-0000", ...}]
```

Json Tabel tbl\_TTMahasiswaPerdidikanFormal

**Gambar 9.** Contoh JSON Dari Tabel-Tabel Mahasiswa



Gambar 9 adalah contoh dari format JSON yang dihasilkan dari web service yang datanya dapat diolah. Sebagai contoh, berikut adalah bagaimana cara mengambil data dari web service tersebut.

Contoh : pada tabel tbl\_TMMahasiswa, data diambil dari JSON pada url

[http://rapi.raharja.me/JSON/qTWfK1EAGJSbLKAc3qu/nim/NIM\\_MAHASISWA](http://rapi.raharja.me/JSON/qTWfK1EAGJSbLKAc3qu/nim/NIM_MAHASISWA)

Di ekstraksi menggunakan program sederhana dari bahasa pemrograman PHP. Berikut adalah *listing program*-nya :

```
<?php
$nim = "1322476037";
$url = "http://rapi.raharja.me/JSON/qTWfK1EAGJSbLKAc3qu/nim/" . $nim;
$file = file_get_contents($url);
$data = json_decode($file);
echo "<table border=1>
    <tr>
        <th>Nama</th>
        <th>Jenis Kelamin</th>
        <th>Tanggal Lahir</th>
    </tr>
    <tr>";
foreach($data as $key) {
    echo '<td>' . $key->NamaDepan . ' ' . $key->NamaBelakang . '</td>';
    echo '<td>' . $key->JenisKelamin . '</td>';
    echo '<td>' . $key->TanggalLahir . '</td>';
}
echo "</tr>
    </table>";
?>
```

Maka output dari *listing program* tersebut adalah sebagai berikut :

**Tabel 2.** Contoh Hasil Output Dari Penarikan Variabel X

Nama	Jenis Kelamin	Tanggal Lahir
Rizal Loa Wanda	L	1994-12-20

Kelebihan dari format pertukaran data JSON adalah data tersebut dapat disesuaikan dengan format yang diinginkan, dan hampir semua bahasa pemrograman dapat merepresentasikan data dari format JSON ini (tidak terbatas oleh *platform*).

NIM	NAMA	Jenjang	Program Studi	Shift Kuliah	iLearning	Jenis Kelamin	Umur	Angkatan	Asal Sekolah
1322476616	Riyan Nova Saputra	100	120	1	0	L	11/27/1995	13	SMAN 4 TANGERANG
1422381588	Riyan Juniarti	100	130	1	0	P	05/21/1995	14	
1422462912	Rizal Ika Wanda	400	430	1	0	L	12/20/1994	13	SMK NEGERI 1 TANGERANG
1412481217	Rizka Permata Dewi	400	410	1	1	P	12/21/1996	14	
1122468923	Rizka Sepriandy	100	120	2	0	L	9/20/1990	11	SMK Penerbangan Padang
1022464650	Rizki Adiguna	100	120	1	0	L	10/27/1991	10	SMA 1 Pasarkemis
1314476702	Rizki Atri Hani Firmansyah	400	410	1	0	P	4/20/1995	13	SMAN 11 TANGERANG
1431481280	Rizki Aprilia	400	450	1	0	L	4/17/1994	14	
1433481287	Rizki Aulia Ramdhani	400	450	1	1	L	2/20/1997	14	
1022476814	Rizki Inam Wahyudi	100	120	1	0	L	12/26/1995	13	SMAN 11 KABUPATEN TANGERANG
1411481101	Rizki Mandala Anugerah	100	110	2	0	L	5/25/1990	14	
1314476814	Rizki Regina Effendiah	400	410	2	0	P	10/22/1995	13	SMAN 17 KABUPATEN TANGERANG
1511476043	Rizki Setyawan	400	410	2	0	L	3/3/1994	13	SMAN 2
1414881193	Rizki Widi Darmawan	300	310	1	1	L	6/29/1996	14	
1414811001	Rizkiana Nurhidayah	100	110	2	0	P	2/5/1995	14	
1412477229	Rizky Agustian	100	110	1	1	L	8/17/1995	13	SMA Yuppentek 4
1333477047	Rizky Indra Kurniawan	400	450	1	1	L	03/21/1995	13	SMK Nusa Putra
1122467624	Rizky Martin	400	420	2	0	L	1/21/1994	11	SMAN 6 TANGERANG

Gambar 10. Hasil Ekstraksi Data Dari Web Service

Sehingga hasil ekstraksi dari beberapa tabel pada set data mahasiswa di *web service* menjadi seperti di Gambar 10, dimana dari format pertukaran data JSON menjadi format tabel yang mengisi setiap atribut x pada *record* mahasiswa.

## E. Transformasi Data

Transformasi data dilakukan agar nilai dari atribut-atribut yang sebelumnya masih berupa kode atau id, menjadi data kategorikal yang dapat dipahami (Gambar 11).

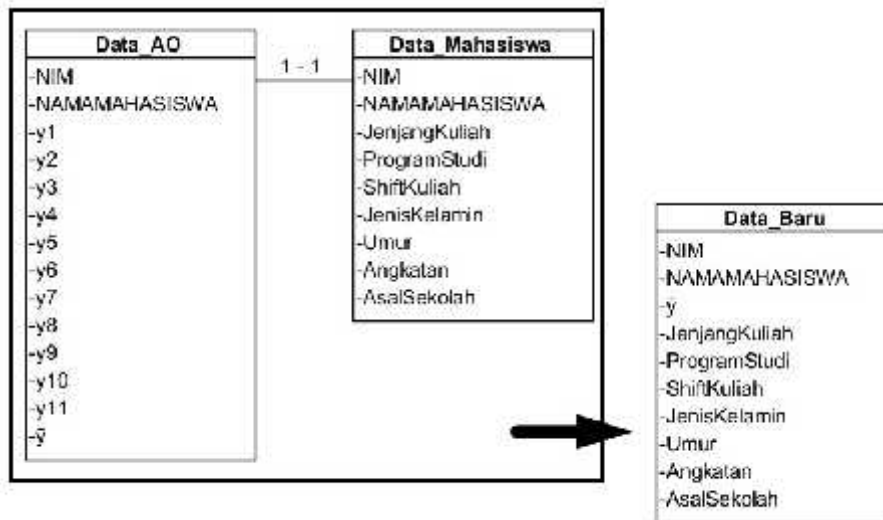
1. Pada atribut jenjang, dan program studi, perubahan data atribut dari kode menjadi data kategorikal didapat dari tabel yang terdapat pada *web service* Raharja.
2. Perubahan data pada atribut shift kuliah ialah dari data angka menjadi data kategorikal : 1 = siang, dan 2 = malam.
3. Perubahan data pada atribut iLearning ialah *discretization* dimana dari bentuk biner (0 dan 1) menjadi Tidak dan Ya.
4. Perubahan data pada atribut umur ialah dari yang sebelumnya tanggal lahir menjadi umur.

NIM	Nama	Jenjang	Program Studi	Shift Kuliah	iLearning	Jenis Kelamin	Umur	Angkatan	Asal Sekolah
1322476616	Riyan Nova Saputra	51	Teknik Informatika	Siang	Tidak	L	20	13	SMAN 4 TANGFRANG
1422381588	Riyan Juniarti	D9	Komputerisasi Akuntansi	Siang	Tidak	P	20	14	
1322476817	Rizal Ika Wanda	51	Teknik Informatika	Siang	Tidak	L	23	13	SMK NEGERI 1 TANGERANG
1412481217	Rizka Permata Dewi	51	Sistem Informasi	Siang	Ya	P	19	14	
1122468923	Rizka Sepriandy	51	Teknik Informatika	Malam	Tidak	L	25	11	SMK Penerbangan Padang
1022464650	Rizki Adiguna	51	Teknik Informatika	Siang	Tidak	L	24	10	SMA 1 Pasarkemis
1314476702	Rizki Atri Hani Firmansyah	51	Sistem Informasi	Siang	Tidak	P	23	13	SMAN 11 TANGERANG
1431481280	Rizki Aprilia	51	Sistem Komputer	Siang	Tidak	L	22	14	
1433481287	Rizki Aulia Ramdhani	51	Sistem Komputer	Siang	Ya	L	19	14	
1322476824	Rizki Inam Wahyudi	51	Teknik Informatika	Siang	Tidak	L	20	13	SMAN 11 KABUPATEN TANGERANG
1411481101	Rizki Mandala Anugerah	51	Sistem Informasi	Malam	Tidak	L	25	14	
1214476814	Rizki Regina Effendiah	51	Sistem Informasi	Malam	Tidak	P	22	12	SMAN 17 KABUPATEN TANGERANG
1311476043	Rizki Setyawan	51	Sistem Informasi	Malam	Tidak	L	23	13	SMAN 2
1414881193	Rizki Widi Darmawan	D9	Manajemen Informatika	Siang	Ya	L	19	14	
1414811001	Rizkiana Nurhidayah	51	Sistem Informasi	Malam	Tidak	P	21	14	
1212477229	Rizky Agustian	51	Sistem Informasi	Siang	Ya	L	21	12	SMA Yuppentek 4
1333477047	Rizky Indra Kurniawan	51	Sistem Komputer	Siang	Ya	L	22	13	SMK Nusa Putra
1122467624	Rizky Martin	51	Teknik Informatika	Malam	Tidak	L	22	11	SMAN 6 TANGERANG

Gambar 11. Hasil Dari Transformasi Data

## F. Konsolidasi Data

Konsolidasi data adalah replikasi data lebih dari satu set data untuk menjadi satu set data. Pada studi kasus ini, set data *absensi online* yang telah diketahui rata-rata dari variabel y yaitu kedisiplinan mahasiswa, dan set data mahasiswa yang berisi atribut-atribut atau variabel x akan digabung menjadi satu set data baru.



**Gambar 12.** Skema dari Konsolidasi Data

Gambar 12 adalah hubungan antara data *absensi online* dengan data mahasiswa yang telah diolah dan digabung menjadi satu. Penggabungan ini menghasilkan set data baru yang siap untuk diolah dalam data mining (Gambar 13).

NIM	Nama	Kata Rata Y	Jenjang	Program Studi	Shift Kuliah	Status iLearning	Jenis Kelamin	Umur	Angkatan	Asal Sekolah
1922475789	A. Andri Setiawan	11.00	SI	Teknik Informatika	Malam	Tidak	L	29	18	SMAN 1 Cemp
1872477818	Aap Kuntubi	12.00	SI	Teknik Informatika	Siang	Tidak	L	27	18	SMAN 18 KABUPATEN TANGERANG
1411481283	Abdul Aziz	11.00	SI	Sistem Informasi	Malam	Tidak	L	21	14	
1414101855	Abdul Aziz	12.70	SI	Sistem Informasi	Malam	Tidak	L	21	14	
1922476005	Abdul Aziz	10.71	SI	Teknik Informatika	Malam	Tidak	L	28	13	
1193488959	Abdul Fatah	11.00	SI	Sistem Komputer	Siang	Ya	L	21	11	MA Darussalam
1872477045	Abdul Haq Aj. Pramono	13.11	SI	Teknik Informatika	Siang	Tidak	L	20	18	SMK Azzena
122474397	Abdul Mujib	11.88	SI	Teknik Informatika	Siang	Tidak	L	23	12	SMK PARI 1 Balaraja
1921476870	Abdul Mukarobi	9.14	SI	Teknik Informatika	Siang	Ya	L	23	13	SMK PUSTEK SERPONG
1493481950	Abdul Ruki Zarkusy	13.00	SI	Sistem Komputer	Siang	Ya	L	20	14	
1311476941	Abdul Ruzid	9.80	SI	Sistem Informasi	Malam	Tidak	L	25	15	
1433481056	Abdul Waqi Rahmat	13.14	SI	Sistem Komputer	Siang	Ya	L	23	14	
1212473737	Abdurrahman Muhammad Ch	9.00	SI	Sistem Informasi	Siang	Ya	L	21	12	SMA Yuppentek 1
1501400014	Abdurrasul Rafli	10.25	SI	Sistem Komputer	Malam	Tidak	L	26	15	
1422481678	Adnan Sumadi	13.44	SI	Teknik Informatika	Siang	Tidak	L	19	14	
1472483006	Adin Marchelo Mardj	12.72	SI	Teknik Informatika	Siang	Tidak	L	20	14	
1221473800	Adipati Arni	11.57	SI	Teknik Informatika	Siang	Ya	L	21	12	SMAN 10 KABUPATEN TANGERANG
1100480076	Ahmad Dafriren A	11.75	SI	Sistem Komputer	Siang	Ya	L	20	11	SMAN 15 Tangerang
1391376143	Ahmad Daruljabr	10.78	SI	Teknik Informatika	Malam	Tidak	L	25	18	SMK YP Karya 2
1224743100	Ahmad Sholah Usdi	13.00	SI	Teknik Informatika	Malam	Tidak	L	22	12	SMKN 1 Kab. Tangerang
1221472305	Ahmad Kumlewan	9.71	SI	Teknik Informatika	Siang	Ya	L	20	12	SMA ISLAMIC CENTRE
1422482719	Ahmad Rizki Ansori	11.80	SI	Teknik Informatika	Siang	Tidak	L	20	14	
1922475440	Ahmad Ruby Farsa	9.63	SI	Teknik Informatika	Malam	Tidak	L	21	13	SMAN 17 KABUPATEN TANGERANG
1291371785	Ahmad Saadwan	10.50	SI	Teknik Informatika	Malam	Tidak	L	26	12	SMKN 3 Serpong
1221472946	Ahmad Subhan	11.25	SI	Teknik Informatika	Siang	Ya	L	21	12	SMA Nusantara 1
1422482714	Ahmad Wale Alqomri	11.50	SI	Teknik Informatika	Malam	Tidak	L	20	14	
1922476601	Ahmad Syahri Bayu Dwi Prabu	11.50	SI	Teknik Informatika	Malam	Tidak	L	22	13	SMKN 6 Kabupaten Tangerang
1221470903	Ahdu Priyandi	11.48	SI	Teknik Informatika	Siang	Ya	L	22	12	SMK WATI 109 Tangerang
1371476870	Ahli Setyoni	11.40	SI	Teknik Informatika	Siang	Ya	P	20	18	SMKN 4 TANGERANG
1121406981	Ahmad Maulana Bachman	7.45	SI	Teknik Informatika	Siang	Tidak	L	24	11	SMK Wachtech 1
1321476078	Ahmad Arif Runadi	11.40	SI	Teknik Informatika	Siang	Ya	L	21	13	PKOM Bina Warga
1301376614	Ade Ferdiansyah	13.38	SI	Teknik Informatika	Malam	Tidak	L	25	13	SMK Komputer
1311376249	Ade Hikmah Sunjam	9.50	SI	Manajemen Informatika	Malam	Tidak	P	20	13	SMK Al Rahman
1221472925	Ade Khasah	13.00	SI	Teknik Informatika	Siang	Ya	L	21	12	SMK AL FALAH
1121409073	Ade Maulana Khumaedi	9.75	SI	Teknik Informatika	Siang	Tidak	L	21	11	SMKN 3 PANGSERANG
1414100726	Ade Permata	12.20	SI	Sistem Informasi	Siang	Tidak	L	26	14	
1411482244	Ade Setiawan	13.22	SI	Sistem Informasi	Siang	Tidak	L	19	14	
1411483296	Ade Saadah Fauziah	12.88	SI	Sistem Informasi	Siang	Tidak	P	18	14	
1222473700	Ade Yohan Hayumi	10.88	SI	Teknik Informatika	Malam	Tidak	L	23	12	SMK Omotif Al-Husna
0021453731	Ahli Dimpatti	8.86	SI	Teknik Informatika	Siang	Tidak	L	25	02	SMAN 6 Tangerang
1011454611	Ahlihye Dwi Pericassaki	10.30	SI	Sistem Informasi	Siang	Tidak	L	24	10	SMA Islam Centre Tangerang
0814461654	Ahli Adhansari	8.50	SI	Sistem Informasi	Malam	Tidak	L	26	08	SMAN 1 Bulusjaya
1311483618	Ahli Alhan	11.48	SI	Sistem Informasi	Malam	Tidak	L	22	13	
1211473424	Ahli Gunawan	10.38	SI	Sistem Informasi	Siang	Tidak	L	23	12	SMKN 2 Kabupaten Tangerang
1321477148	Ahli Prasetyu	13.40	SI	Teknik Informatika	Siang	Ya	L	20	13	SMKN 5 KAB.TANGERANG

Gambar13. Hasil Dari Konsolidasi Data

## KESIMPULAN

*Preprocessing* data telah dilakukan sesuai dengan proses yang ada pada KDD, dengan menghasilkan set data yang baru hasil dari penggabungan dua sumber data yang berbeda yaitu data absensi online mahasiswa dan data mahasiswa. Proses yang dilakukan diantaranya penarikan data dari sumber data, pembersihan data, pemilihan atribut, transformasi data, dan konsolidasi data. *Preprocessing* data pada studi kasus ini menghasilkan 1.836 *record* mahasiswa yang siap untuk diproses dalam data mining. Penelitian ini juga telah menentukan atribut-atribut yang diasumsikan memiliki pengaruh terhadap kedisiplinan mahasiswa (variabel y), diantaranya ialah jenjang kuliah, program studi, shift kuliah, status iLearning, jenis kelamin, umur, angkatan dan asal sekolah.

**DAFTAR PUSTAKA**

- [1] D. Delen, *Real-World Data Mining: Applied Business Analytics and Decision Making*. New Jersey: Pearson Education, Inc, 2014.
- [2] E. Prasetyo, *DATA MINING - Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: C.V ANDI OFFSET, 2012.
- [3] G. Trayasiwi, "Penerapan Metode Klustering dengan Algoritma k-Means Untuk Prediksi Kelulusan Mahasiswa Pada Program Studi Teknik Informatika Strata Satu", Skripsi, Universitas Dian Nuswantoro, 2015.
- [4] H. Mustafidah, "Model Regresi Data Mining Motivasi Belajar Pengaruhnya Terhadap Tingkat Kedisiplinan Mahasiswa", *JUITA*, vol. 1, no. 1, 2010.
- [5] N. Astuti, K. Kurniasi and M. Arief, "Analisis Prediksi Tingkat Ketidaksiplinan Siswa Menggunakan Algoritma Naïve Bayes Classifier (Studi Kasus : SMK Negeri 1 Pacitan)", *Seminar Nasional Teknologi Informasi dan Multimedia*, vol. 3, no. 1, 2015.
- [6] P. Elisa, "Analisis Faktor-Faktor Yang Mempengaruhi Kedisiplinan Kerja Karyawan Pada PT. Suka Fajar Pekanbaru", Skripsi, Universitas Islam Negeri Sultan Syarif Kasim Riau, 2013.
- [7] S. Ginting, W. Zarman and I. Hamidah, "Analisis Dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik", *Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST)*, 2014.