Penerapan C4.5 Berbasis *Particle Swarm Optimization* (PSO) dalam Memprediksi Siswa Lolos Seleksi Perguruan Tinggi

Sulistiyanto[1)]
Magister Teknik Informatiaka Institute Informatika dan Bisnis Darmajaya
Jl.Zainal Aabidin Pagar Alam No 93, Gedong Meneng Bandar Lampung 35141
e-mail: Sulistiyanto9@gmail.com[1)]

Abstrak

Semakin banyaknya sekolah terutama sekolah menengah atas membuat banyak calon siswa mempunyai banyak pilihan untuk menentukan di sekolah mana mereka akan melanjutkan pendidikan menengah atasnya. Salah satu faktor dalam memilih sekolah adalah kualitas lulusanya yang lolos seleksi perguruan tinggi. Namun pihak sekolah jarang melakukan evaluasi terhadap alumninya yang telah lolos seleksi, sehingga tidak mendapatkan pengetahuan tentang faktor yang mendukung lolos seleksi dari alumninya. Oleh sebab itu diperlukan kajian prediksi menggunakan data mining dengan teknik klasifikasi. salah satu teknik klasifikasi yang dinilai tepat digunakan yakni C4.5 yang selanjutnya dioptimasi dengan algoritma Particle Swarm Optimization (PSO). Dari hasil pengujian algoritma C4.5 menghasilkan akurasi sebesar 89,89% sedangkan akurasi C4.5 dengan PSO sebesar 94,56%. Hasil percobaan terbukti bahwa algoritma PSO dapat meningkatkan akurasi C45.

Kata kunci: *Data Mining, Prediksi, Klasifikasi, C4.5, PSO, Optimasi, Lolos Seleksi*

1. Pendahuluan

Semakin banyaknya sekolah menengah atas yang bermunculan membuat calon siswa dihadapkan dengan banyak pilihan untuk melanjutkan sekolah ke jenjang selanjutnya. Jenjang menengah atas merupakan jenjang awal untuk memasuki dunia pendidikan yang lebih tinggi yakni tingkat perguruan tinggi. Dengan banyaknya sekolah menengah atas yang ada, membuat calon siswa harus selektif dalam memilih dan menentukan sekolah mana yang akan menjadi tujuan mereka selanjutnya. Banyak faktor yang harus di pertimbangkan oleh calon siswa dalam memilih sekolah. Selain karena akreditasi sekolah, peluang lolos seleksi masuk perguruan tinggi pun harus menjadi salah satu pertimbangan dalam memilih sekolah, baik melalui seleksi tanpa tes maupun dengan tes. Karena jika sekolah tersebut lulusannya banyak yang lolos seleksi masuk perguruan tinggi, berarti bisa dikatakan sekolah tersebut memiliki kualitas yang bagus.

Namun banyak sekolah yang tidak memperhatikan hal tersebut dikarenakan tidak pernah melakukan pemantauan dan evaluasi terhadap alumni yang telah lolos seleksi masuk perguruan tinggi. Akibatnya sekolah tidak memiliki pengetahuan tentang faktor apa saja yang menjadikan siswa lolos seleksi. Berdasarkan hal tersebut, data mining dinilai mampu untuk melakukan prediksi dengan teknik klasifikasi. Salah satu teknik klasifikasi yang dinilai tepat yakni *Decision Tree* khususnya C4.5. Pada penelitian sebelumnya menyatakan bahwa algoritma C4.5 termasuk algoritma yang lemah [1], maka diperlukanlah optimasi atau peningkatan akurasi. *Particle Swarm Optimization* (PSO) merupakan algoritma optimasi yang dapat digunakan untuk meningkatkan akurasi algoritma lain.

2. Metode Penelitian

2.1 Data Mining

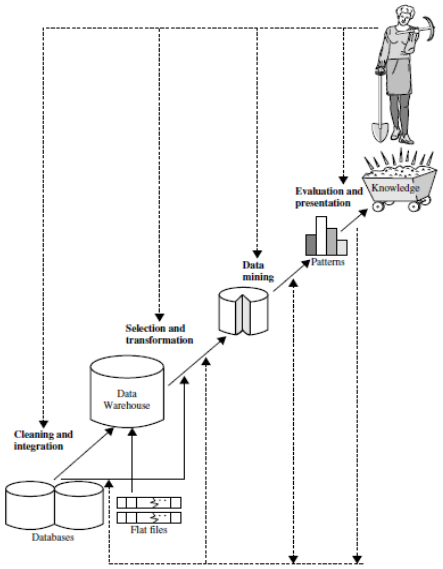
Data mining merupakan sebuah istilah yang digunakan untuk menggambarkan sebuah kegiatan penemuan pengetahuan didalam sebuah kumpulan data. Data mining merupakan proses

yang menggunakan teknik statistik, matematika, kecerdasan buatan dan pembelajaran berbasis mesin untuk mengekstrak dan mengidentifikasi informasi yang berguna beserta pengetahuan dari sebuah bank data yang besar [2].

Setiap aplikasi kelas data mining dibantu oleh pendekatan sebuah algoritma untuk mengekstrak suatu hubungan dalam sebuah data. Perbedaan kelas pendekatan algoritma tersebut dapat membantu dalam menyelesaikan masalah yang ditemui. Beberapa kelas tersebut adalah:

1. Clasifikasi
   Mendefinisikan karakteristik dari sebuah grub atau membagi kedalam beberapa jenis yang diketahui.
2. Clustering
   Mengelompokkan data yang tidak diketahui label kelasnya kedalam sejumlah kelompok tertentu sesuai ukuran kemiripannya.
3. Association
   Mengidentifikasi hubungan antara kejadian yang terjadi dalam satu waktu. Pendekatan ini didasarkan pada analisis keranjang belanja.
4. Sequencing
   Sama dengan pendekatan association, namun yang membedakan hanya di waktu terjadinya. Sequence mengukur waktu berdasarkan periode tertentu.
5. Regression
   Menemukan suatu fungsi yang memodelkan data dengan kesalahana prediksi (galat) seminimal mungkin.

Gambar 1 berikut adalah proses *mining data* untuk mendapatkan pengetahuan:



Gambar 1. Proses *Mining Data* untuk Mendapatkan Pengetahuan [3]

2.2 Algoritma C4.5

Algoritma C4.5 atau sering disebut juga *Decision Tree* merupakan metode klasifikasi dan prediksi yang sangat terkenal. Metode ini mengubah fakta yang sangat besar menjadi pohon keputusan yang representasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami

[4]. Proses pada *decision tree* adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule* dan menyederhanakannya.

Saat menyusun sebuah *decision tree* pertama yang harus dilakukan adalah menentukan atribut mana yang akan menjadi simpul akar dan atribut mana yang akan menjadi simpul selanjutnya. Pemilihan atribut yang baik adalah atribut yang memungkinkan untuk mendapatkan *decision tree* yang paling kecil ukurannya atau atribut yang bisa memisahkan objek menurut kelasnya. Untuk dapat membuat sebuah pohon keputusan, maka nilai entropynya harus di cari dahulu menggunakan persamaan:

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

keterangan:
S : himpunan kasus
A : atribut
N : jumlah partisi S
Pi : proporsi dari Si terhadap

Setelah itu mencari nilai *information gain. Information Gain* merupakan kriteria yang paling populer untuk pemilihan atribut. Information gain dapat dihitung dari output data atau variabel dependen *y* yang dikelompokkan berdasarkan atribut *A*, dinotasikan dengan gain (*y,A)*. Gain *(y,A)* dari atribut *A* relatif terhadap output data *y* adalah:

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

2.3 Algoritma *Particle Swarm Optimization (PSO)*

Algoritma ini merupakan algoritma berbasis populasi yang mengeksploitasi individu dalam pencarian. Dalam PSO populasi disebut swarm dan individu disebut particle. Tiap particle berpindah dengan kecepatan yang diadaptasi dari daerah pencarian dan menyimpan sebagai posisi terbaik yang pernah dicapai.

Setiap partikel dalam PSO juga dikaitkan dengan kecepatan partikel terbang melalui ruang pencarian dengan kecepatan yang dinamis disesuaikan untuk perilaku historis mereka. Oleh karena itu, partikel memiliki kecenderungan untuk terbang menuju daerah pencarian yang lebih baik selama proses pencarian (Nuswantoro, 2013) [5]. Persamaan untuk menghitung perpindahan posisi dan kecepatan partikel yaitu:

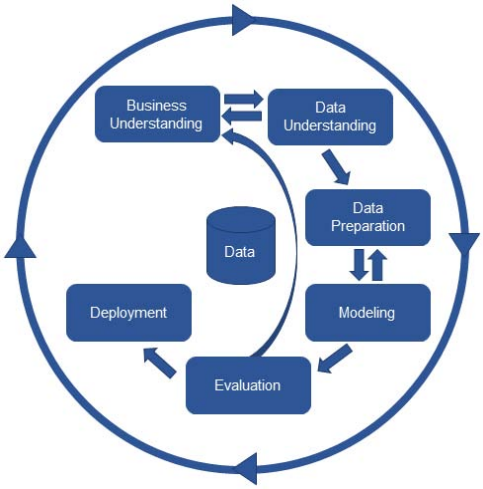
$$V_i(t) = V_i\ (t\text{-}1) + c_1 r_1 [X_{pbest\ i} - X_i\ (t)] + c_2 r_2 [X_{Gbest} - X_i\ (t)]$$

$$X_i\ (t) = X_i\ (t\text{-}1) + V_i\ (t)$$

Dimana:

$V_i(t)$	:kecepatan partikel *i* saat iterasi *t*
$X_i\ (t)$	: posisi partikel *i* saat iterasi *t*
c_1 *dan* c_2	: *learning rate* untuk kemampuan individu (cognitive) dan pengaruh sosial
r_1 *dan* r_2	: bilangan random yang berdistribudi uniformal dalam interval 0 dan 1
$X_{pbest\ i}$	: posisi terbaik partikel *i*
X_{Gbest}	: posisi terbaik global

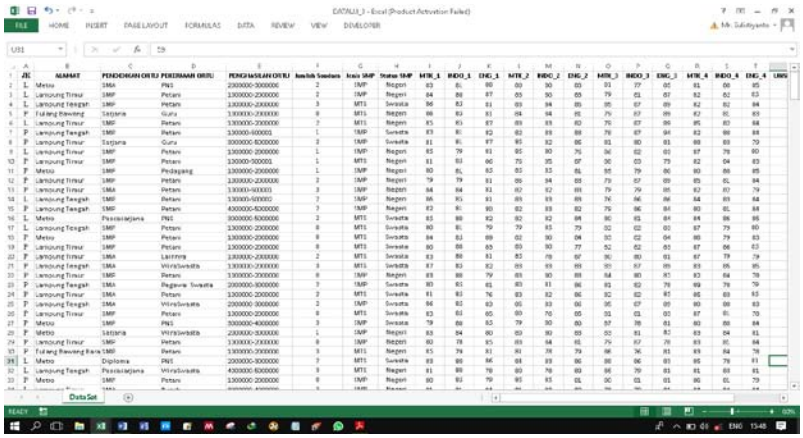
Model yang digunakan dalam penelitian ini adalah model CRISP-DM seperti pada gambar 2 di bawah ini:



Gambar 2. Siklus CRISP-DM [6]

Adapun penjelasan dari setiap langkah CRISP-DM adalah seperti berikut:

1. *Pemahaman Bisnis*. Kualitas suatu sekolah salah satunya diukur dari rasio lolosnya siswa lulusanya ke perguruan tinggi dari jalur bebas tes maupun melalui jalur tes. Perlu diadakan evaluasi dan pemantauan terhadap alumni yang telah melanjutkan ke perguruan tinggi, faktor apa saja yang dapat menunjang lulus tes seleksi.
2. *Pemahaman data*. Data yang digunakan adalah data siswa tahun 2015 - 2017 dan data nilai dari semester 1 – 4.
3. *Persiapan data*. Data yang terkumpul akan disatukan menjadi *dataset* dan membersihkan data dari *missing value* dan diambil data siswa yang lolos dan tidak lolos.Atribut yang digunakan yakni, jenis kelamin, alamat, pendidikan orang tua, pekerjaan orang tua, penghasilan orang tua, jumlah saudara, jenis SMP, status SMP, nilai mapel UN SMP dan rerata UN SMP. Terdapat pada gambar 3 berikut ini:



Gambar 3. *Dataset*

4. *Pemodelan.*Menggunakan metode C4.5 yang dioptimasi dengan algoritma *Particle Swarm Optimization.*
5. *Validasi dan evaluasi.* Menggunakan framework RapidMiner yang selanjutnya akan dilakukan pengujian keakurasian dengan *confusion matrix* dengan persamaan

$$Accuracy = \frac{a+b}{a+b+c+d} \text{ atau } \frac{TP+TN}{TP+FP+TN+FN}$$

3. Hasil dan Pembahasan

Algoritma C4.5 merupakan sebuah model dengan membentuk pohon keputusan. Untuk membuat pohon keputusan, berikut langkah-langkah yang harus dilakukan:

1. Menentukan akar dari pohon. Akar diambil dari atribut terpilih dengan cara menghitung nilai entropy masing-masing atribut dengan persamaan sebagai berikut:

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

Tabel 1 di bawah ini merupakan tabel jumlah kasus:

Tabel 1. Jumlah Kasus

	NILAI	JUMLAH KASUS	LULUS	TIDAK LULUS	Entropy
Total Kasus		386	78	308	0.726053857
Jenis Kelamin					
	Laki-Laki	106	23	83	0.754616702
	Perempuan	280	55	225	0.714727473
Alamat					
	Metro	72	10	62	0.581321499
	Lampung Tengah	87	21	66	0.79732651
	Lampung Timur	170	33	137	0.710006223
	Tulang Bawang	10	3	7	0.881290899
	Tubabar	4	4	0	0
	Pesawaran	8	2	6	0.811278124
	Lampung Barat	2	1	1	1
	Mesuji	9	5	4	0.99107606
	Tanggamus	1	1	0	1
	Way Kanan	5	1	4	0.721928095
	Pringsewu	1	0	1	0
	Lainnya	5	1	4	0.721928095
Dst..					

2. Setelah mendapat nilai entropy dari masing-masing atribut, selanjutnya mencari nilai gain dengan persamaan terdapat pada gambar 2 sebagai berikut ini:

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

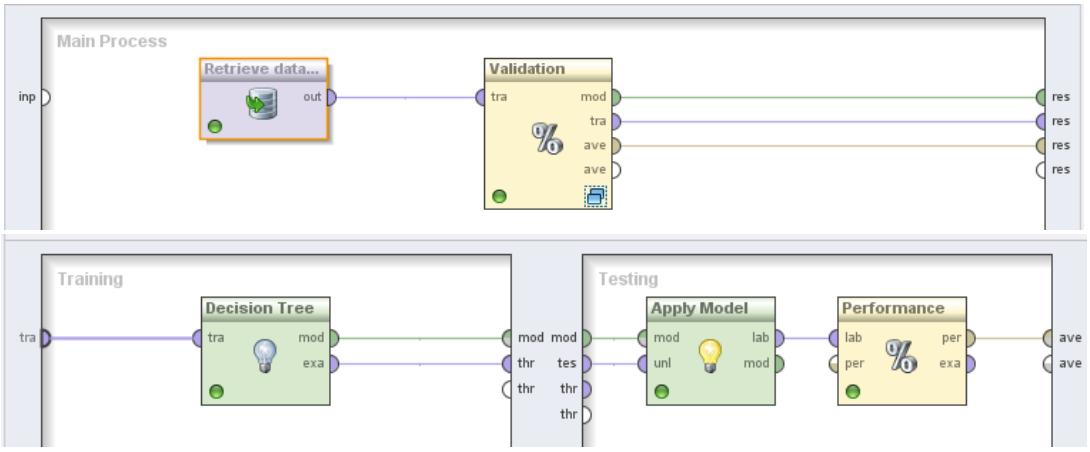
Tabel 2 berikut merupakan tabel nilai gain:

Tabel 2. Tabel Nilai Gain

	NILAI	JUMLAH KASUS	LULUS	TIDAK LULUS	Entropy	Gain
Total Kasus		386	78	308	0.726053857	
Jenis Kelamin						0.000372347
	Laki-Laki	106	23	83	0.754616702	
	Perempuan	280	55	225	0.714727473	
Alamat						0.035987247
	Metro	72	10	62	0.581321499	
	Lampung Tengah	87	21	66	0.79732651	
	Lampung Timur	170	33	137	0.710006223	
	Tulang Bawang	10	3	7	0.881290899	
	Tubabar	4	4	0	0	
	Pesawaran	8	2	6	0.811278124	
	Lampung Barat	2	1	1	1	
	Mesuji	9	5	4	0.99107606	
	Tanggamus	1	1	0	1	
	Way Kanan	5	1	4	0.721928095	
	Pringsewu	1	0	1	0	
	Lainnya	5	1	4	0.721928095	
Dst..						

3. Ulangi langkah 1 dan 2 untuk mengisi entropy dan nilai gain sampai semua atribut terisi
4. Proses partisi pohon keputusan akan berhenti saat:
   a. Semua tupel dalam node N mendapat kelas yang sama
   b. Tidak ada atribut di dalam tupel yang dipartisi lagi
   c. Tidak ada tupel di dalam cabang yang kosong

5. Pembentukan rule C4.5 dengan menggunakan RapidMiner terdapat pada gambar 4 di bawah ini:



Gambar 4. Pemodelan Decision Tree C4.5 dengan RapidMiner

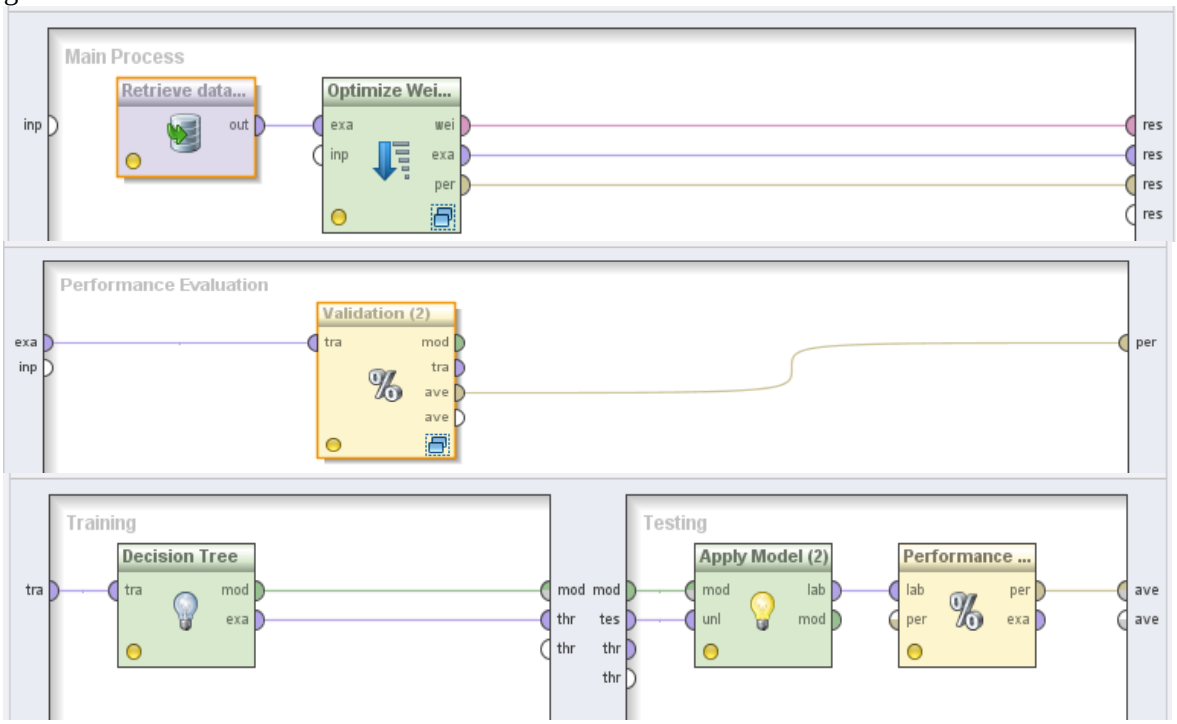
Dari pemodelan diatas, maka didapat hasil sebagai berikut dengan nilai akurasi sebesar 89,89%, terdapat pada gambar 5 di bawah ini:

accuracy: 89.89% +/- 4.71% (mikro: 89.90%)

	true Lulus	true Tidak Lulus	class precision
pred. Lulus	61	22	73.49%
pred. Tidak Lulus	17	286	94.39%
class recall	78.21%	92.86%	

Gambar 5. Hasil dari Pemodelan C4.5 dengan RapidMiner

Agar nilai akurasi tersebut dapat meningkat, maka diperlukanlah algoritma optimasi yang salah satunya algoritma PSO. Pemodelan menggunakan algoritma optimasi PSO terdapat pada gambar 6 berikut ini:



Gambar 6. Pemodelan Decision Tree C4.5 dengan PSO dengan RapidMiner

Dari model optimasi dengan PSO diatas, hasil yang didapat akan meningkat dari 89,89% menjadi 94,57%. Mengalami peningkatan 3,68%. Terdapat pada gambar 7 berikut ini:

accuracy: 94.57% +/- 2.92% (mikro: 94.56%)

	true Lulus	true Tidak Lulus	class precision
pred. Lulus	63	6	91.30%
pred. Tidak Lulus	15	302	95.27%
class recall	80.77%	98.05%	

Gambar 7. Hasil pemodelan C4.5 dengan PSO di RapidMiner

3.1 Validasi dan evaluasi

Dalam tahap ini dilakukan validasi dan pengukuran keakuratan hasil yang dicapai oleh model menggunakan beberapa teknik yang terdapat di framework RapidMiner yaitu *confusion matrix*. Tabel 3 berikut merupakan *Confusion Matrix* Algoritma C4.5:

Tabel 3. *Confusion Matrix* Algoritma C4.5

	True Lulus	True Tidak Lulus
Prediksi Lulus	61	22
Prediksi Tidak Lulus	17	286

Dari tabel 3 tersebut dapat dihitung akurasi sebagai berikut:

$$\text{Akurasi} = \frac{61+286}{61+22+17+286} * 100\% = 89{,}89\%$$

Tabel 4 berikut merupakan *Confusion Matrix* algoritma C4.5 dengan PSO

Tabel 4. *Confusion Matrix* Algoritma C4.5 dengan PSO

	True Lulus	True Tidak Lulus
Prediksi Lulus	63	6
Prediksi Tidak Lulus	15	302

Dari tabel 4 tersebut dapat dihitung akurasi sebagai berikut:

$$\text{Akurasi} = \frac{63+302}{63+6+15+302} * 100\% = 94{,}57\%$$

4. Simpulan

Berdasarkan hasil pengujian diatas, dapat disimpulkan bahwa Algoritma C4.5 termasuk algoritma yang *fair clasification* sehingga perlu di optimasi agar tingkat akurasinya menjadi meningkat dan algoritma PSO terbukti dapat meningkatkan tingkat akurasi algoritma C4.5. Sebelum PSO dilibatkan, akurasi yang dihasilkan C4.5 yakni 84,84%, namun setelah PSO dilibatkan dalam pemodelan, maka akurasi naik menjadi 94,57.

Hal ini menunjukkan bahwa algoritma PSO terbukti dapat meningkatkan tingkat akurasi dari Algoritma C4.5 dengan kenaikan sebesar 3,68%. Diharapakan dengan adanya pengujian ini, akan bermanfaat bagi pihak sekolah agar bisa mempersiapkan segala sesuatunya berkaitan dengan siswa untuk dapat lulus seleksi masuk perguruan tinggi.

Daftar Pustaka

[1] Nurul, Oktariani. Analisa perbandingan algoritma K-Means, Decision tree dan naive bayes untuk sistem pengelompokan siswa otomatis. *Jurnal Ilmian Tenologi Informasi Terapan (JITTER).* 2016; Volume II(no 2):10.

[2] Turban, et al. Decision Support System and Intellegent System.*7th.*New Delhi: Prentice-Hall. 2007: 263.

[3] Han, et al. Data Mining Concept and techniques. 3rd. USA: Elsevier Inc. 2006:7.

[4] Fiastantyo, Gian. Perbandingan kinerja metode klasifikasi data mining menggunakan Naïve Bayes dan Algoritma C4.5 untuk prediksi ketepatan waktu kelulusan mahasiswa. 2009.
[5] Muarif, Khoirul. Komparasi pemodelan data menggunakan C4.5 dan C4.5 berbasis Particle Swarm optimization untuk memprediksi kelulusan mahasiswa. Semarang Universitas Dian Nuswantoro; 2013.
[6] Watson, Hugh J, et al. The CRISP-DM: nwe blueprint of data mining. *Jurnal of Data Mining.* 2000; vol 5(no 4).