Perbandingan Aplikasi Data Mining WEKA Dan SPSS Climentine Menggunakan Dataset Mahasiswa

Priati
Universitas Buana Perjuangan Karawang
Jl. H.S Ronggowaluyo, Telukjambe Timur, Karawang
e-mail: priati@ubpkarawang.ac.id

Abstrak

Data mining dapat dilakukan dengan aplikasi yang bersifat komersil seperti *SPSS Celemtine* maupun yang open source seperti *WEKA*. Aplikasi-aplikasi tersebut akan mempermudah dalam melakukan penggalian data. Penelitian data mining menggunakan aplikasi-aplikasi tersebut sudah banyak dilakukan akan tetapi jarang yang mencoba untuk membandingkan penggunaan aplikasinya. Hal ini mengakibatkan banyak pertanyaan mengenai aplikasi data mining apa yang baik untuk digunakan. Ilmu data mining hadir untuk memberikan solusi penanganan tumpukan data. Penggalian data mahasiswa yang dilakukan dengan cermat akan memberikan hasil yang memuaskan dan menghasilkan solusi yang bukan sesaat akan tetapi solusi yang bisa diterapkan secara berkesinambungan dan berkelanjutan. *WEKA* dan *SPSS Clementine* sangat mudah digunakan dengan dukungan algoritma yang tergolong banyak dan lengkap. Dalam penelitian ini digunakan metode klasifikasi serta membandingkan hasilnya. Perbandingan dilakukan pada 1160 dataset mahasiswa. Hasil penggalian data dengan kedua aplikasi ini mengindikasikan bahwa menggunakan WEKA maupun SPSS Clementine pada dasarnya adalah sama saja karena persentase akurasi hasil analisisnyapun hampir sama. Pada *WEKA* akurasi yang dihasilkan adalah 80,3446% dan pada *SPSS Celemtine* akurasi yang dihasilkan adalah 80,52%.

Kata kunci: *Aplikasi, Data mining, Klasifikasi, WEKA, Clementine, Mahasiswa*

1. Pendahuluan

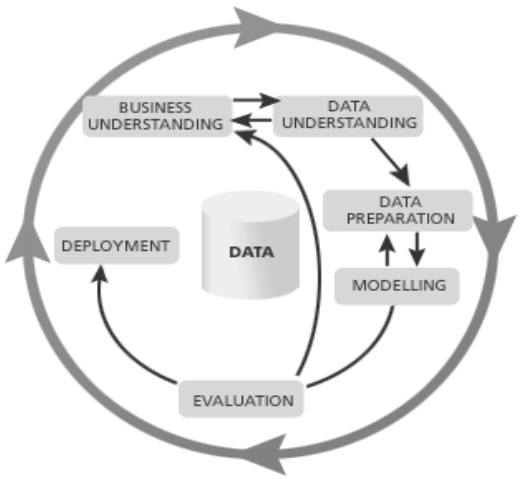
"Data Mining merupakan sebuah analisa dari observasi data dalam jumlah besar untuk menentukan hubungan yang tidak diketahui sebelumnya dan dua metode baru untuk meringkas data agar mudah dipahami serta kegunaannya untuk memilih data" [1]. Data mining hadir untuk dapat memberikan solusi dalam penanganan atau penggalian data yang tujuannya adalah menemukan solusi atau informasi yang tersembunyi dalam tumpukan data. Data mining dapat dilakukan dengan aplikasi seperti *Mathlab, Rapidminer, WEKA* maupun *SPSS Climentine*. Aplikasi-aplikasi tersebut akan mempermudah dalam melakukan penggalian data. Penelitian data mining menggunakan aplikasi-aplikasi tersebut sudah banyak tetapi jarang yang mencoba membandingkan penggunaan aplikasi-aplikasinya. *The Waikato Environment for Knowledge Analysis* (WEKA) adalah rangkaian lengkap perpustakaan kelas *Java* yang mengimplementasikan banyak *state-of-the-art* pembelajaran mesin dan algoritma *data mining* [2]. SPSS dipilih sebagai aplikasi yang digunakan untuk pengolahan data, pemilihan lebih karena fakta bahwa SPSS adalag aplikasi statistik terpopuler didunia, termasuk di Indonesia. SPSS sejak awal memang berkomitmen mengembangkan prosedur statistik yang dapat digunakan pada bidang bisnis mulai dari yang sederhana, cukup kompleks seperti multivariate, metode SEM (dengan mengakuisisi AMOS), sampai aplikasi data mining lewat aplikasi Clementine [3]. SPSS Clementine merupakan aplikasi data mining yang bersifat komersial. Aplikasi ini mendukung berbagai format data seperti Excel, SAS dan lain sebagainya. Aplikasi ini menyediakan menu bantuan serta mendukung berbagai algoritma data mining dengan jumlah data yang sangat besar tanpa memenuhi memori [4]. C5.0 adalah versi komersial dari C4.5 yang secara luas digunakan dibanyak pemaketan *data mining* seperti Clementine dan RuleQuest. Tidak seperti C4.5, penggunaan algoritma yang tepat untuk C5.0 belum terungkap. Hasil menunjukkan bahwa C5.0 meningkatkan pada penggunaaan memori sekitar 90%, lebih cepat daripada C4.5 [5]. Selama 10 tahun terakhir, sudah banyak aplikasi data mining yang dikembangkan baik yang bersifat komersial maupun open source. Aplikasi-aplikasi tersebut antara lain *SAS Enterprise Miner, SPSS Clementine, Weka, RapidMiner, R, STATISTICA, Data Miner, Orange Canvas, DataEngine, DBMiner, WebMiner, MARS, Datamite, GainSmart, XLMiner, IntelligentMiner, Darwin, AI Trilogy, Alice, AnswerTree, BrainMaker, JDBCMiner, Braincel, DecisionTime* [6].

Penelitian ini mencoba hadir sebagai jawaban atas banyaknya kebingungan pemilihan antara menggunakan aplikasi *data mining* WEKA atau SPSS Clementine.

2. Metode Penelitian

Penelitian ini menggunakan metode CRISP-DM [7]. Tahapan-tahapannya adalah:

1. Fase Pemahaman Bisnis
   Pada fase ini akan ditentukan tujuan proyek dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan. Selanjutnya akan dilakukan penerjemahan tujuan dan batasan menjadi formula dari permasalahan *data mining* serta menyiapkan strategi awal untuk mencapai tujuan. Adapun tujuan dari penelitian ini adalah membandingkan dua hasil pengolahan dataset mahasiswa dengan menggunakan WEKA dan SPSS Clementine.
2. Fase Pemahaman Data
   Mengumpulkan data mahasiswa dengan atribut yang telah ditentukan yaitu JNJNG, PRODI, TPTLHR, JNSKLMN, JMLSKS, dan IPK. Menggunakan analisis penyelidikan data untuk mengenali lebih lanjut data dan pencarian pengetahuan awal serta mengevaluasi kualitas data.
3. Fase Pengolahan Data
   Pada fase pengolahan data akan disiapkan dari data awal. Kumpulan data ini akan digunakan untuk keseluruhan fase berikutnya. Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif. Pilih kasus dan variabel yang ingin dianalisis dan yang sesuai analisis yang akan dilakukan. Lakukan perubahan pada beberapa variabel jika dibutuhkan. Serta siapkan data awal sehingga siap untuk perangkat pemodelan.
4. Fase Pemodelan
   Dalam fase pemodelan akan dipilih dan diaplikasikan teknik pemodelan yang sesuai. Melakukan kalibrasi aturan model untuk mengoptimalkan hasil. Yang perlu diperhatikan adalah bahwa beberapa teknik mungkin untuk digunakan pada permasalahan data mining yang sama. Jika diperlukan, proses dapat kembali ke fase pengolahan data untuk menjadikan data kedalam bentuk yang sesuai dengan spesifikasi kebutuhan teknik *data mining* tertentu.
5. Fase Evaluasi
   Mengevaluasi satu atau lebih model yang digunakan dalam fase pemodelan untuk mendapatkan kualitas dan efektivitas sebelum disebarkan untuk digunakan. Melakukan penetapan apakah terdapat model yang memenuhi tujuan pada fase awal dan menentukan apakah terdapat permasalahan penting dari bisnis atau penelitian yang tidak tertangani dengan baik serta mengambil keputusan berkaitan dengan penggunaan hasil dari *data mining*.
6. Fase Penyebaran
   Fase terakhir ini adalah fase dimana semua model yang dihasilkan akan digunakan. Yang perlu diingat adalah terbentuknya model tidak menandakan telah terselesaikannya proyek. Contoh sederhana penyebaran adalah pembuatan laporan. Sedangkan yang lebih kompleks adalah penerapan proses data mining secara parallel pada departemen lain .



Gambar 1. Proses CRISP-DM [7]

2.1. Fase Pemahaman Bisnis

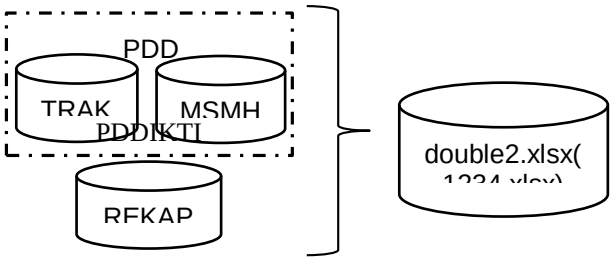
Tujuan proyek dalam penelitian ini adalah mengkaji komparasi hasil analisis menggunakan WEKA dan SPSS Clementine dengan algoritma C4.5. Pada tahap ini dilakukan pemahaman terhadap tujuan proyek

dan kebutuhan secara detail dalam lingkup bisnis atau unit penelitian secara keseluruhan dan menerjemahkannya kedalam tujuan *data mining*.

Strategi awal untuk mencapai tujuan adalah melakukan permintaan data mahasiswa kepada Bagian Akademik (BAAK) STMIK Rosma Karawang.

2.2. Fase Pemahaman Data

Data mahasiswa yang didapat dari Bagian Akademik (BAAK) STMIK Rosma Karawang berupa dokumen *spreadsheet* dan *DBF*. Sumber data utama yang digunakan dalam penelitian ini adalah *Data* mahasiswa STMIK Rosma Karawang jenjang DIII dan S1 pada tahun 2000 sampai dengan tahun 2011 dengan format xlsx dan DBF. Data dalam tabel 3.1 adalah gabungan dari data yang ada di *folder* REKAP_NILAI, *Data* PDDIKTI yaitu master mahasiswa (MSMHS.DBF) dan transkrip nilai mahasiswa (TRAKM.DBF). Hasil penggabungan dari file di *folder* REKAP_NILAI, *file* MSMHS.DBF dan *file* TRAKM.DBF adalah *file* dengan nama 1234.xlsx.



Gambar 2. Penggabungan data

2.3. Fase Pengolahan Data

Fase ini merupakan pekerjaan berat yang perlu dilaksanakan secara intensif. Persiapan data mencakup semua kegiatan untuk membangun Data mahasiswa yang akan diterapkan kedalam alat pemodelan dari data mentah awal berupa *Data* mahasiswa dan selanjutnya akan melakukan proses *data mining*. Data awal mahasiswa tersebut ditransformasi menjadi data kategori. Hal ini bertujuan agar dapat mempermudah dalam penggalian data dan mudah diproses oleh alat bantu *data mining*

Hasil evaluasi terhadap kualitas data adalah masih terdapat data yang rangkap atau *double.* Memilih dan menciptakan satu data untuk mendukung proses penemuan knowledge akan dilakukan. Mencari dari data yang tersedia, memperoleh data tambahan yang dibutuhkan, mengintegrasikan seluruh data untuk menemukan pengetahuan dalam sebuah *Data*. Peneliti melakukan *preprocessing* dan *cleansing* seperti menangani data yang tidak lengkap dan menghilangkan gangguan atau *outlier*.

Fase ini adalah fase pengolahan yang menggunakan data dari Fase Pemahaman Data . Variabel yang digunakan antara lain Tempat Lahir (TPTLHR), Jenis Kelamin (JNSKLMN), Jenjang (JNJNG), Program Studi (PRODI), Jumlah SKS (JMLSKS), IPK, Tanggal Kelulusan (KELULUSAN).

Data awal mahasiswa tersebut ditransformasi menjadi data kategori. Hal ini bertujuan agar dapat mempermudah dalam penggalian data dan mudah diproses oleh alat bantu *data mining*. Adapun pengkategorian data sebagai berikut:

(1) Variable TPTLHR (Tempat Lahir)
Jenis datanya dikategorikan Karawang dan Non-Karawang (tempat lahir diluar Karawang)

(2) Variable JNSKLMN (Jenis Kelamin)
Jenis datanya dikategorikan Laki-laki dengan inisial L dan Perempuan dengan inisial P.

(3) Variabel JNJNG (Jenjang)
Jenis datanya dikategorikan seperti pada tabel 1 berikut:

Tabel 1 Kode Jenjang

Kode	Jenjang
C	Strata Satu (S1)
E	Diploma Tiga (DIII)

(4) Variabel PRODI (Program Studi)
Kategori Prodi dapat dilihat pada tabel 2 berikut:

Tabel 2 Kode Program Studi

Kode	Prodi	Kategori
55201	Teknik Informatika (S1)	TIC
56201	Sistem Komputer (S1)	SKC
57201	Sistem Informasi (S1)	SIC
57401	Manajemen Informatika (D3)	MIE

57402	Komputerisasi Akuntansi (D3)	KAE

(5) Variabel JMLSKS (Jumlah SKS)
Jenis data JMLSKS merupakan data real yang dikategorikan menjadi 4 seperti terlihat pada tabel 3.

Tabel 3 Kategori Jumlah SKS

Kategori	Jumlah SKS
SATU	C≥144
DUA	C<144
TIGA	E≥110
EMPAT	E<110

(6) Variabel IPK (Indeks Prestasi Komulatif)
Jenis data IPK dikategorikan menjadi 3 seperti ditampilkan pada tabel 4 berikut,

Tabel 4 Kategori IPK

Kategori	Nilai IPK
BESAR	IPK≥3,5
SEDANG	2,75<IPK<3,5
KECIL	IPK≤2,75

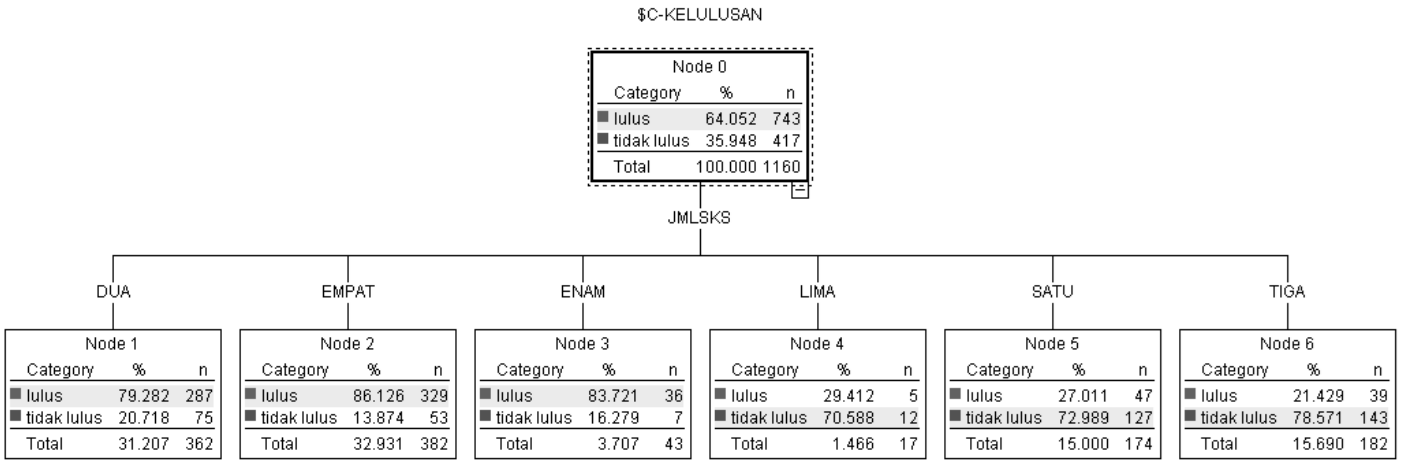
(7) Variabel KELULUSAN (Tanggal Kelulusan)
Variabel ini adalah data yang berjenis numerikal yang harus dilakukan proses inisiasi data terlebih dahulu kedalam bentuk nominal. Inisiasi tahun lulus dilakukan dengan: Mahasiswa dari setiap angkatan yang sudah terdapat tahun kelulusan dinyatakan "lulus". Mahasiswa dari setiap angkatan yang belum terdapat tahun kelulusan dinyatakan "tidak lulus".

3. Hasil dan Pembahasan

Untuk dapat mengolah data pada aplkasi data mining, file data yang ada harus disesuaikan dengan aplkasi tersesbut. Pada SPSS Clementine 10.1 peneliti menggunakan format file "sav", sedangkan pada WEKA peneliti menggunakan format file "csv".

Berikut disajikan hasil penelitian dan pembahasan dengan menggunakan 1160 data menggunakan dua aplikasi *data mining* diatas.

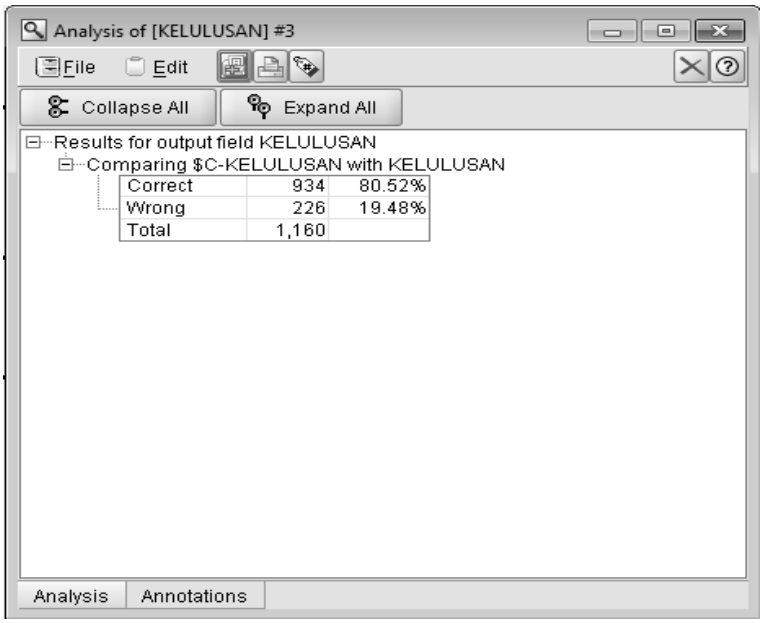
1. SPSS Clementine 10.1



Gambar 3 Pohon Keputusan dari 1160 data

Berdasarkan pohon keputusan yang terbentuk dari 1160 data didapatkan hasil bahwa jika JMLSKS masuk dalam kategori DUA maka sebanyak 287 data "lulus" dan 75 data "tidak lulus". Jka JMLSKS masuk dalam kategori EMPAT maka sebanyak 329 data "lulus" dan 53 data "tidak lulus". Jika JMLSKS masuk dalam kategori ENAM maka sebanyak 36 data "lulus" dan 7 data "tidak lulus". Jika JMLSKS masuk dalam kategori LIMA maka sebanyak 5 data "lulus" dan 12 data "tidak lulus". Jika JMLSKS masuk dalam kategori SATU maka sebanyak 47 data "lulus" dan 127 "tidak lulus". Jika JMLSKS masuk dalam kategori TIGA maka sebanyak 39 data "lulus" dan 143 data "tidak lulus".

Berikut adalah persentase akurasi hasil analisis.



Gambar 4. Persentase Akurasi hasil Analisis

Berdasarkan analisis dengan menggunkan Clementine, akurasi yang diperoleh sebesar 80,52 %.

2. WEKA

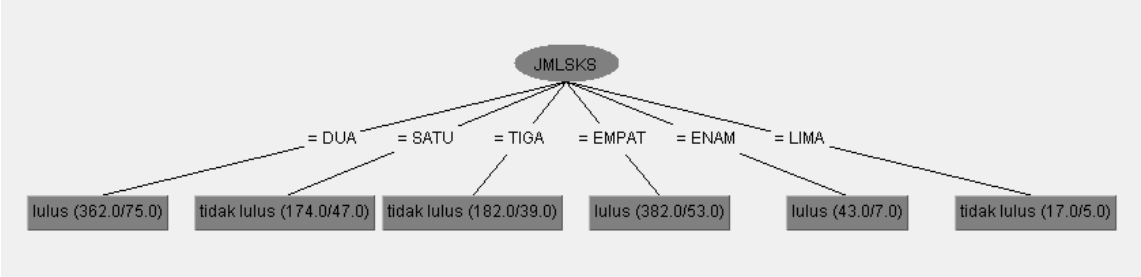
```
=== Run information ===
Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     1160_Data
Instances:    1160
Attributes:   7
              TPTLHR
              JNSKLMN
              JNJNG
              PRODI
              JMLSKS
              IPK
              KELULUSAN
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===
J48 pruned tree
------------------
JMLSKS = DUA: lulus (362.0/75.0)
JMLSKS = SATU: tidak lulus (174.0/47.0)
JMLSKS = TIGA: tidak lulus (182.0/39.0)
JMLSKS = EMPAT: lulus (382.0/53.0)
JMLSKS = ENAM: lulus (43.0/7.0)
JMLSKS = LIMA: tidak lulus (17.0/5.0)
Number of Leaves  :  6
Size of the tree :  7
Time taken to build model: 0.06 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         932               80.3448 %
Incorrectly Classified Instances       228               19.6552 %
Kappa statistic                          0.5626
Mean absolute error                      0.3114
Root mean squared error                  0.3965
Relative absolute error                 67.6013 %
Root relative squared error             82.6321 %
```

```
Total Number of Instances          1160

=== Detailed Accuracy By Class ===
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.878    0.329     0.826     0.878     0.851      0.776    lulus
                 0.671    0.122     0.755     0.671     0.711      0.776    tidak lulus
Weighted Avg.    0.803    0.254     0.801     0.803     0.801      0.776

=== Confusion Matrix ===
   a   b   <-- classified as
 652  91 |   a = lulus
 137 280 |   b = tidak lulus
```

Berdasarkan hasil analisis tersebut, *confution matrix* yang diperoleh adalah 80.3448 %.



Gambar 5. Pohon Keputusan

Secara keseluruhan persentase hasil analisis menggunakan SPSS Clementine dan WEKA adalah 80,52% dan 80,3448%, seperti disajikan dalam tabel 5.

Tabel 5. Presentase Hasil Analisis

Algoritma	Clementine	WEKA
C4.5	80,52%	80,3448%

4. Simpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa menggunakan WEKA maupun SPSS Clementine pada dasarnya adalah sama saja karena persentase akurasi hasil analisisnyapun hampir sama. Penelitian ini tidak berusaha untuk melakukan perbandingan terkontrol algoritma mana yang terkuat pada masing-masing aplikasi data mining. Harapan kedepannya adalah penelitian ini bisa dijadikan sebagai gambaran dan acuan tentang pendekatan metode klasifikasi yang digunakan pada masing-masing aplikasi. Kami menyarankan kepada para peneliti lain untuk menggunakan berbagai metode klasifikasi yang berbeda menggunakan aplikasi data mining yang disebutkan untuk membandingkan dan membuat keputusan sendiri tentang aplikasi semacam itu.

Daftar Pustaka

[1] Jefri. Implementasi Algoritma C4.5 Dalam Aplikasi Untuk Memprediksi Jumlah Mahasiswa Yang Mengulang Mata Kuliah Di STMIK AMIKOM Yogyakarta, Yogyakarta. 2013
[2] Witten, Ian H, dkk. Data Mining: Practical Machine Learning Aplkation and Techniques. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA. 2011
[3] Santoso, Singgih. Statistik Multivariat, Jakarta : PT Gramedia. 2010
[4] SPSS Inc. Clementine 11.1 User's Guide. SPSS Incorporated. 2007
[5] Larose, Daniel. T. Discovering Knowledge in Data: An Introduction to Data Mining. John Willey & Sons. Inc. 2005
[6] Satyanarayana, A. Software Tools For Teaching Undergraduate Data Mining Course. New York City College of Technology. 2013.
[7] Dunham, M.H. Data Mining Introductory And Advanced Topic. New Jersey: Prentice Hall. 2003