IJCCS, Vol. 1 No. 2, June, 2007: 115 - 122

Fuzzy C-Means Clustering Model Data Mining For Recognizing Stock Data Sampling Pattern

Sylvia Jane Annatje Sumarauw[1], Subanar[2]
[1]Faculty of Mathematics and Natural Science, Manado State University
[2]Faculty of Mathematics and Natural Science, Gadjah Mada University

Abstract

Capital market has been beneficial to companies and investor. For investors, the capital market provides two economical advantages, namely deviden and capital gain, and a non-economical one, that is a voting share in Shareholders General Meeting. But, it can also penalize the shareowners. In order to prevent them from the risk, the investors should predict the prospect of their companies. As a consequence of having an abstract commodity, the share quality will be determined by the validity of their company profile information. Any information of stock value fluctuation from Jakarta Stock Exchange can be a useful consideration and a good measurement for data analysis. In the context of preventing the shareholders from the risk, this research focuses on stock data sample category or stock data sample pattern by using Fuzzy c-Means Clustering Model which providing any useful information for the investors. The research analyses stock data such as Individual Index, Volume and Amount on Property and Real Estate Emitter Group at Jakarta Stock Exchange from January 1 till December 31 of 2004. The mining process follows Cross Industry Standard Process model for Data Mining (CRISP-DM) in the form of circle with these steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation and Deployment. At this modelling process, the Fuzzy c-Means Clustering Model will be applied. Data Mining Fuzzy c-Means Clustering Model can analyze stock data in a big database with many complex variables especially for finding the data sample pattern, and then building Fuzzy Inference System for stimulating inputs to be outputs that based on Fuzzy Logic by recognising the pattern.

Keywords: Data Mining, Fuzzy c-Means Clustering Model, Pattern Recognition

1. Pengantar

Perkembangan teknologi informasi dalam beberapa dekade terakhir, memungkinkan semakin mudahnya memperoleh data dan informasi dalam jumlah yang besar. Pertumbuhan data yang tersimpan dalam suatu database yang besar, telah jauh melebihi kemampuan manusia untuk bisa memahami sehingga diperlukan alat dan metode tepat yang mampu mentransformasikan sejumlah besar data kedalam informasi yang berguna yang menopang keakuratan informasi itu sendiri.

Data mining adalah salah satu metode yang digunakan untuk menganalisis data karena banyak memberikan ide untuk menghasilkan hal-hal yang berbeda dari sebelumnya. Para ahli data mining menggunakan teknik statistika, dan analisis clustering adalah salah satu teknik baru dalam statistika yang dikembangkan untuk menyelesaikan masalah analisis data multivariat. Teknik ini mencari pola sampel data berdasarkan proses pertumbuhan kelompok data homogen yang disebut cluster.

Menurut Han dan Kamber [1], dalam praktek data tidak terdefinisi dengan baik dan pengaturannya tidak terdefinisi dengan tepat, batasan antara dua atau lebih cluster menjadi kurang jelas yaitu; titik data bisa menjadi milik dua cluster atau lebih. Fuzzy Clustering mengusulkan model yang lebih mendekati pada persoalan dunia nyata, yaitu bagaimana data dihasilkan dari pola yang teridentifikasi. Fuzzy c-Means Clustering adalah pengenalan pola dengan pemodelan yang lebih fleksibel dan memudahkan pemecahan perhitungan dari masalah yang dirumuskan.

Dari segi perusahaan pasar modal adalah salah satu alternatif yang dapat dimanfaatkan untuk memenuhi kebutuhan dananya, sedangkan bagi investor atau pemegang saham ada beberapa manfaat yang

dapat diperoleh antara lain; manfaat ekonomis yang bisa diperoleh meliputi perolehan dividen yaitu sebagian keuntungan perusahaan yang dibagikan kepada investor dan perolehan capital gain yaitu keuntungan yang diperoleh investor dari hasil jual beli saham, berupa selisih antara nilai jual yang lebih tinggi dibandingkan nilai beli yang lebih rendah. Selain itu pula manfaat non ekonomis yang bisa diperoleh oleh pemegang saham adalah kepemilikan hak suara dalam Rapat Umum Pemegang Saham (RUPS) untuk menentukan jalannya perusahaan.

Selain manfaat diatas, investor juga bisa mengalami kerugian sebagai resiko yang harus ditanggungnya. Untuk menghindari resiko kerugian maka selain investor dapat menghubungi penasehat investasi, investor sebaiknya memprediksikan apakah perusahaan emiten mempunyai prospek yang bagus atau tidak atau dengan kata lain apakah perusahaan beresiko tinggi atau tidak.

Karena masyarakat pemodal atau investor membeli suatu komoditi yang sangat abstrak maka kualitas dari komoditi ini yaitu saham ditentukan oleh kualitas informasi yang tersedia dari perusahaan emiten yang bersangkutan.

Bursa Efek Jakarta adalah gudang data perdagangan saham yang bisa dijadikan informasi perusahaan emiten dan bisa dijadikan tolok ukur penilaian perusahaan emiten. Fluktuasi nilai indeks harga saham individual sangat penting khususnya bagi calon investor dalam penentuan jenis saham yang akan dibelinya [2]. Selain itu banyaknya transaksi jual beli saham perhari dan total nilai transaksi dalam rupiah perhari bisa juga memberikan gambaran situasi perusahaan emiten. Oleh sebab itu pola sampel data nilai perdagangan saham bisa memberikan informasi kualitas dari perusahaan emiten. Untuk menjawab kebutuhan masyarakat khususnya para calon investor dalam melihat pola sampel data fluktuasi nilai perdagangan saham dan untuk penentuan kualitas perusahaan emiten, penulis melakukan penelitian dengan memilih metode fuzzy c-means clustering yakni algoritma pengclusteran terawasi untuk pengenalan pola sebagai informasi kepada para investor bagaimana pola sampel data saham dan kualitas perusahaan emiten pada bursa efek jakarta.

2. Cara Penelitian

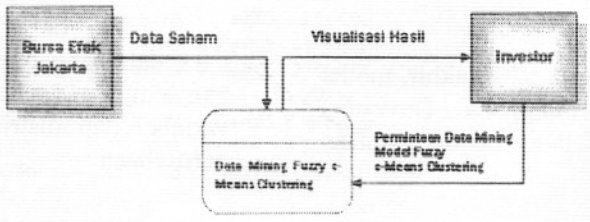
Penelitian ini mengikuti sistem format proses data mining dari Cross Industry Standard Process Model for Data mining (Crisp-DM) dengan tahapan-tahapan; Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation dan Deployment.

2.1 Business Understanding

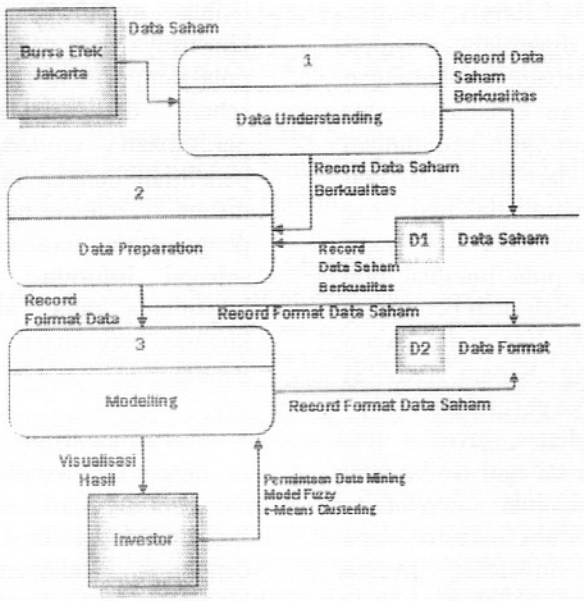
Pada fase awal proses data mining yakni business understanding terfokus pada bagaimana memahami tujuan dan kebutuhan proyek dari sudut pandang bisnis, kemudian merubah pengetahuan menjadi sebuah definisi permasalahan data mining dan sebuah rencana awal yang dirancang untuk mencapai tujuan.

2.2 Data Understanding

Pada fase ini dilakukan beberapa tahapan yaitu; pengumpulan data, deskripsi data, eksplorasi data dan verifikasi kualitas. Data yang digunakan adalah data saham pada Bursa Efek Jakarta periode 1 Januari sampai 31 Desember 2004. Tahap pertama yaitu menyusun informasi berbentuk angka dari data yang merupakan sari numerik statistik kemudian untuk menentukan distribusi data secara visual data dibuat dalam bentuk diagram. Digunakan diagram kotak garis (boxplot)



Gambar 1. Diagram Konteks



Gambar 2. Diagram Arus Data Proses Data Mining

2.3 Data Preparation

Selanjutnya tahap data preparation, dilakukan select data dan clean data. Penulis menggunakan variabel; Date, Day, Code, Company, Volume, Amount dan Individual Index. Data asli tersimpan pada format tabel data Excel, di export ke MS Access kemudian dibaca oleh Matlab melalui ODBC.

2.4 Modelling

Model yang digunakan adalah Fuzzy c-Means Clustering Kemudian setelah didapat pola maka dilanjutkan dengan membangun fuzzy inference system (FIS).

2.5 Evaluation

Setelah terbentuk model data mining fuzzy c-Means Clustering dan fuzzy inference system maka penulis melakukan evaluasi untuk memastikan bahwa model yang dihasilkan mencapai tujuan bisnis.

2.6 Deployment

Pada tahap ini yaitu tahap akhir, model ditawarkan kepada pengguna (investor) berupa visualisasi dalam bentuk tulisan (Thesis).

3 Hasil dan Pembahasan

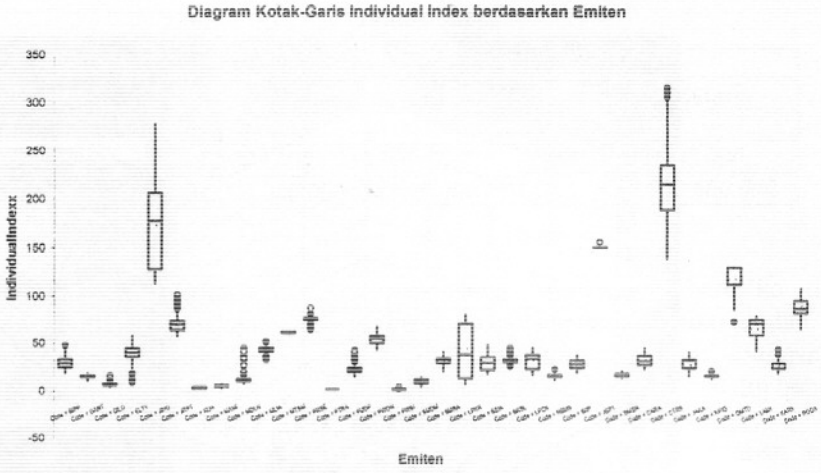
3.1 Deskripsi dan Eksplorasi Data

Pada proses data understanding dilakukan deskripsi dan eksplorasi data, Berikut diberikan contoh hasil deskripsi dan eksplorasi data (gambar 3 dan 4).

IndividualIndex	
Mean	47,25970027
Standard Erro	0,589236739
Median	30,489
Mode	1,48
Standard Devi	50,9613139
Sample Variar	2597,055514
Kurtosis	4,720926631
Skewness	2,075157133
Range	314,107
Minimum	0,893
Maximum	315
Sum	353502,558
Count	7480

Gambar 3. Deskripsi IndIndex

Dari hasil deskripsi dan eksplorasi data, data sangat bervariasi terutama tentang banyaknya data untuk masing-masing variabel dan distribusi data masing-masing perusahaan.



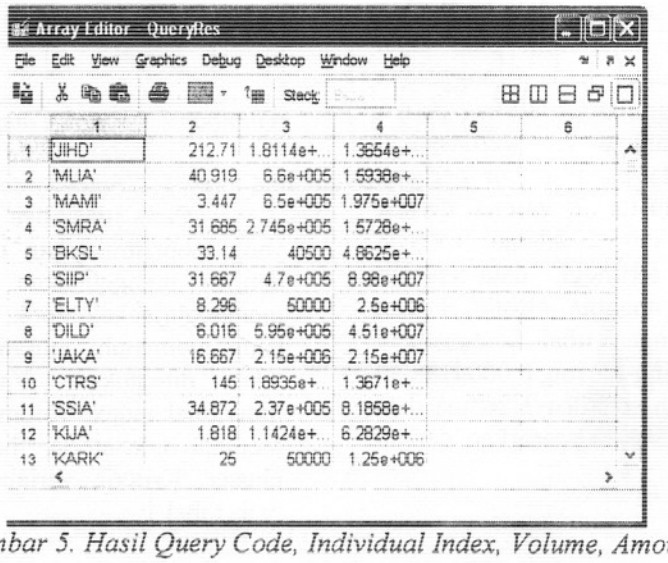
Gambar 4. Diagram Kotak Garis Individual Index

3.2 Data Cleaning

Pada proses data preparation ada beberapa tahapan yang pada dasarnya bertujuan untuk cleaning data, Hasil data cleaning akan tampil sesuai dengan kebutuhan melalui perintah query. Pada gambar 5 ditunjukkan contoh hasil query yaitu kode perusahaan, individual index, volume dan amount.

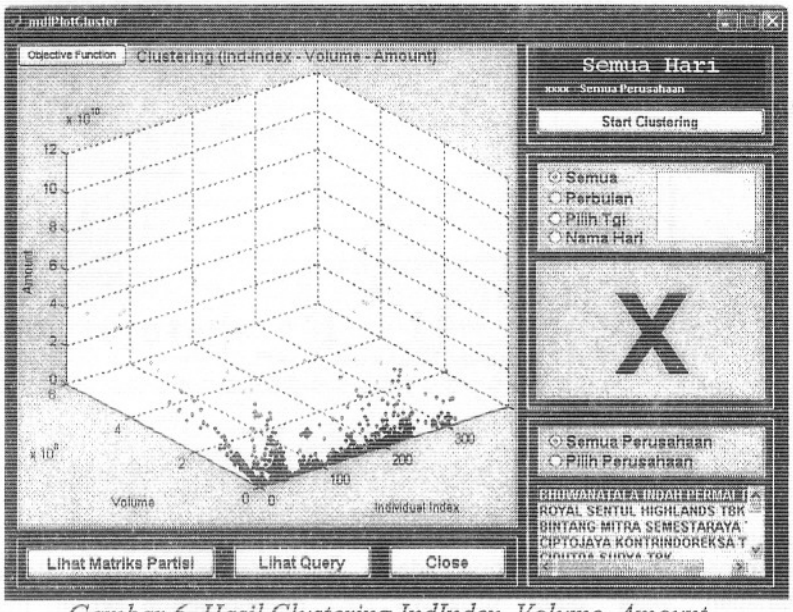
3.3 Fuzzy c-Means Clustering

Setelah tahapan preparation maka dilanjutkan dengan tahapan modelling. Pada tahap modelling digunakan algoritma clustering yaitu algoritma Fuzzy c-Means Clustering. Algoritma ini menentukan pusat cluster yang optimal. Pada kondisi awal pusat cluster belum akurat, iterasi akan berulang sampai konvergen. Ditentukan 3 cluster. Untuk semua data, clustering optimal terbentuk pada iterasi ke 193 dengan nilai fungsi objektif 14596082325051882000000, dan nilai masing-masing pusat cluster: Pusat-1 = (104.28, 53526714.45, 12179722883.41), pusat-2 = (41.95, 2146034.19, 414195692.41) dan pusat-3 = (97.34, 222336444.45, 47729411989.94)



	1	2	3	4	5	6
1	'JIHD'	212.71	1.8114e+...	1.3654e+...		
2	'MLIA'	40.919	6.6e+005	1.5938e+...		
3	'MAMI'	3.447	6.5e+005	1.975e+007		
4	'SMRA'	31.685	2.745e+005	1.5728e+...		
5	'BKSL'	33.14	40500	4.8625e+...		
6	'SIIP'	31.667	4.7e+005	8.98e+007		
7	'ELTY'	8.296	50000	2.5e+006		
8	'DILD'	6.016	5.95e+005	4.51e+007		
9	'JAKA'	16.667	2.15e+006	2.15e+007		
10	'CTRS'	145	1.8935e+...	1.3671e+...		
11	'SSIA'	34.872	2.37e+005	8.1858e+...		
12	'KIJA'	1.818	1.1424e+...	6.2829e+...		
13	'KARK'	25	50000	1.25e+006		

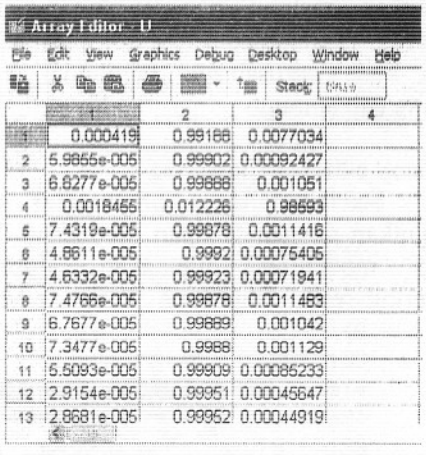
Gambar 5. Hasil Query Code, Individual Index, Volume, Amount



Gambar 6. Hasil Clustering IndIndex, Volume, Amount

Gambar titik-titik data dan hasil clustering terlihat pada gambar 6 dengan masing-masing titik pusat dan keanggotaan pada cluster. Pada gambar plot cluster terlihat data mengelompok pada cluster bawah artinya banyak data yang dibawah rata-rata.

Untuk mengetahui keanggotaan tiap titik data pada cluster, dilihat dari matriks partisi untuk cluster optimal yaitu derajat keanggotaan masing-masing titik data ke cluster. Banyaknya kolom menunjukkan banyaknya cluster dan banyaknya baris menunjukkan banyaknya data .Jumlah keanggotaan adalah satu, titik data akan menjadi anggota cluster pada keanggotaan terbesar. Hasil tampilan matriks partisi ditunjukkan pada gambar 7. Berdasarkan gambar tersebut data pada urutan pertama menjadi anggota cluster 2 karena derajat keanggotaan di cluster 2 adalah 0,99188 derajat keanggotaan tertinggi, data tersebut juga bisa menjadi anggota cluster 1 dan cluster 3 tetapi dengan nilai keanggotaan yang kecil yaitu masing-masing 0,000419 dan 0,0077034. Demikian seterusnya semua data menjadi anggota cluster berdasarkan derajat keanggotaan.



	1	2	3	4
1	0.000419	0.99188	0.0077034	
2	5.9855e-005	0.99902	0.00092427	
3	6.8277e-005	0.99888	0.001051	
4	0.0018465	0.012226	0.98593	
5	7.4319e-005	0.99878	0.0011416	
6	4.8611e-005	0.9992	0.00075405	
7	4.6332e-005	0.99923	0.00071941	
8	7.4766e-005	0.99878	0.0011483	
9	6.7677e-005	0.99889	0.001042	
10	7.3477e-005	0.9988	0.001129	
11	5.5093e-005	0.99909	0.00085233	
12	2.9154e-005	0.99951	0.00045647	
13	2.8681e-005	0.99952	0.00044919	

Gambar 7. Contoh Tampilan Matriks Partisi

3.4 Fuzzy Inference System

Pada dasarnya FIS adalah proses yang mensimulasikan input yang diberikan dengan menggunakan logika fuzzy yaitu mengelola sekelompok input menjadi output yang memenuhi aturan-aturan (rules) yang telah ditentukan. Variabel input yaitu individual index, volume dan amount dibagi menjadi 3 kriateria berdasarkan hasil plot cluster. Kriteria variabel input ditampilkan pada tabel 1.

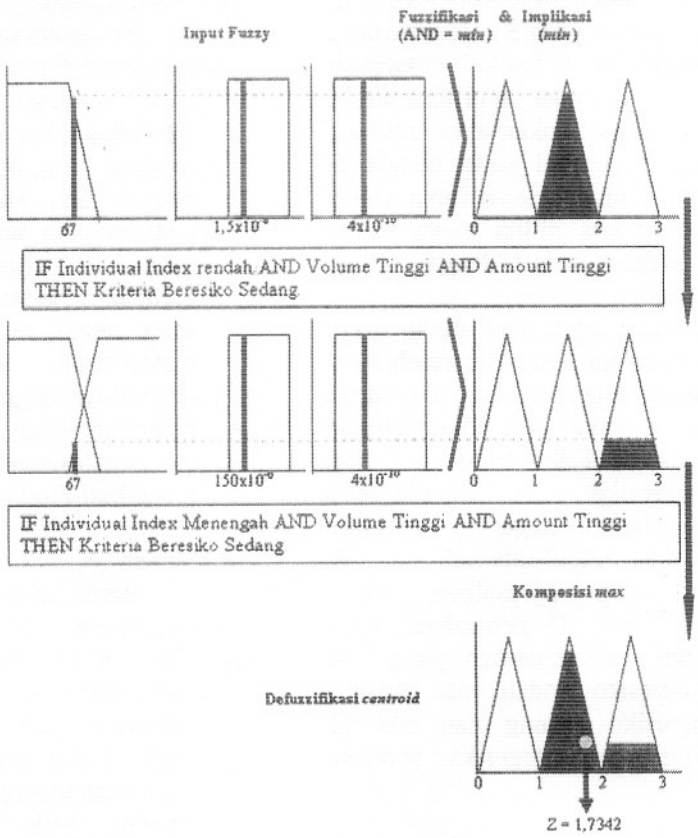
Tabel 1. Kriteria Variabel Input

Variabel	Kriteria		
	Rendah/Kecil	Menengah/Sedang	Tinggi/Besar
IndIndex	< 69	69 - 100	> 100
Volume	< 112.000.000	112.000.000 - 138.000.000	> 138.000.000.
Amount	< 24.000.000.000	24.000.000.000-.30.000.000.000	> 30.000.000.000

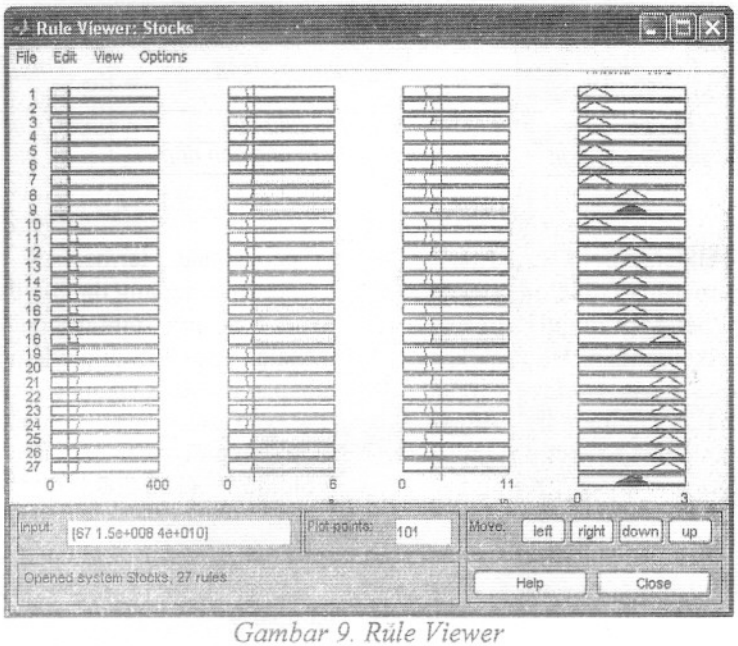
Selain pengelompokkan kriteria variabel input, output juga dikelompokkan menurut kriteria; beresiko tinggi, beresiko sedang dan beresiko rendah. Berdasarkan pengelompokkan kriteria variabel input dan output diatas, dibentuk fungsi keanggotaan (membership function) untuk masing-masing variabel. Kemudian dibentuk aturan fuzzy untuk mendapatkan fuzzy decision yang pada akhirnya akan menghasilkan output keputusan apakah perusahaan termasuk beresiko tinggi, beresiko sedang dan beresiko rendah.

Secara umum dalam menganalisis masalah pada FIS dilakukan melalui beberapa tahapan proses analisis sistem fuzzy.

Contoh hasil proses analisis sistem fuzzy untuk data Individual Index (67), Volume (150.000.000) dan Amount (40.000.000.000) seperti pada gambar 8. Sedangkan hasil implementasi Rule Viewer Fuzzy Matlab Toolbox ditunjukkan pada gambar 9.



Gambar 8. Diagram Proses Analisis Sistem



Gambar 9. Rule Viewer

Pada gambar tampilan 9 di atas, terlihat bahwa gambar pada 3 kolom pertama merupakan variabel input dan kolom terakhir adalah output. Pada pilihan input bisa ditulis nilai input yang kita inginkan atau langsung menggeser garis vertikal yang berada di masing-masing kolom variabel input sesuai dengan data yang kita miliki maka kolom output akan bergeser sesuai dengan aturan yang sudah dibuat.

Berdasarkan input, dan aturan fuzzy yang sudah ditentukan maka diperoleh hasil dari 4360 transaksi selama tahun 2004, hanya ada 22 transaksi yang bisa dikategorikan beresiko rendah, 478 transaksi bisa dikategorikan beresiko sedang dan ada 3860 transaksi yang bisa dikategorikan beresiko tinggi. Berdasarkan output rata-rata transaksi dari masing-masing perusahaan maka diperoleh hasil, dari 35 perusahaan yang diteliti tidak terdapat perusahan yang bisa dikategorikan beresiko rendah, dan hanya 3 perusahaan beresiko sedang dan ada 32 perusahaan yang bisa dikategorikan beresiko tinggi.

4 Kesimpulan dan Saran

4.1 Kesimpulan

Berdasarkan hasil dan pembahasan maka dapat disimpulkan sebagai berikut:

1. Data mining model fuzzy c-means clustering bisa menganalisis data dalam database yang besar dengan variabel yang banyak dan kompleks terutama untuk mencari pola sampel data, jumlah iterasi pada hasil clustering setiap percobaan berubah tergantung matriks partisi awal, akan tetapi nilai pusat masing-masing cluster tetap.
2. Kemudian dengan mengenal pola dari data dibangun Fuzzy Inference System untuk mensimulasikan input menjadi output mengikuti Fuzzy Logic.

4.2 Saran

Berdasarkan hasil dan pembahasan maka disarankan hal-hal sebagai berikut;

1. Data mining sangat menarik dan sangat bermanfaat terutama untuk data dalam database yang besar, variabel yang banyak dan kompleks maka disarankan agar dilakukan penelitian lebih lanjut data mining baik model clustering yang berbeda atau dengan model bukan

clustering dengan macam-macam aplikasi.

2. Resiko pembelian saham biasanya terkait dengan variabel return. Untuk pengembangan penelitian sebaiknya menggunakan data variabel return yang dihitung dari rata-rata selisih individual index antara dua hari transaksi yang berurutan.

Daftar Pustaka

[1] J. Han, M. Kamber, *Data Mining Concepts and Techniques*, Simon Fraser University, Morgan Kaufmann Publishers, 2001.

[2] P. Anoraga, P. Pakarti, *Pengantar Pasar Modal*, Edisi Revisi, Rineka Cipta, Semarang, 2001.