Jurnal Teknovasi
Volume 02, Nomor 01, 2015, 33 – 60
ISSN : 2355-701X

ANALISIS SUBSPACE CLUSTERING MENGGUNAKAN DBSCAN DAN SUBCLU UNTUK PROYEKSI PEKERJAAN ALUMNI PERGURUAN TINGGI

Anni Rotua Aritonang[1*], Sutarman[1] & Poltak Sihombing[1]
[1]Program S2 Teknik Informatika Universitas Sumatera Utara, Medan, Indonesia
*E-Mail : anniaritonang@yahoo.co.id

ABSTRAK

Subspace clustering diproyeksikan sebagai teknik pencarian untuk mengelompokkan data atau atribut pada klaster yang berbeda, Pengelompokan dilakukan dengan menentukan tingkat kerapatan data dan juga mengidentifikasi *outlier* atau data yang tidak relevan, sehingga masing-masing *cluster* ada dalam subset tersendiri. Tesis ini mengusulkan inovasi algoritma *subspace clustering based on density connection*. Pada tahap awal akan dihitung kerapatan dimensi, hasil kerapatan dimensi akan dijadikan data masukan untuk menentukan klaster awal yang berdasarkan kerapatan dimensi, yakni dengan menggunakan Algoritma DBSCAN. Data pada setiap klaster kemudian akan diuji apakah memiliki hubungan dengan data pada klaster yang lain, yakni dengan menggunakan Algoritma SUBCLU. Hasil dari penelitian ini ditemukan bahawa SUBCLU tidak memiliki *un-cluster* dataset nyata, sehingga persepsi hasil *cluster* akan menghasilkan informasi yang lebih akurat sedangkan untuk kepuasan kerja dataset DBSCAN membutuhkan waktu lebih lama daripada metode SUBCLU. Untuk lebih besar dan lebih kompleks data, kinerja SUBCLU terlihat lebih efisien daripada DBSCAN.

Kata Kunci : *Subspace clustering*, DBSCAN, SUBCLU.

PENDAHULUAN

Seiring dengan perkembangan teknologi, khususnya didalam teknologi informasi dan komunikasi, menyulut kebutuhan tenaga kerja yang berlatar belakang pendidikan dalam bidang ilmu komputer dan teknologi informasi. Sebagai konsekuensinya pertumbuhan lembaga pendidikan tinggi yang berbasis pada ilmu komputer dan teknologi informasi (Diploma dan Sarjana) juga mengalami pertumbuhan yang sangat pesat. Situasi ini memicu persaingan yang ketat antar alumni perguruan tinggi untuk merebut pasar kerja. Persaingan yang ketat bagi alumni perguruan tinggi untuk meraih pasar kerja menjadi salah satu isu yang sering menjadi topik-topik pembahasan di kalangan pimpinan perguruan tinggi. Sehinggan memunculkan pertanyaan bagaimana rancangan program belajar mengajar bagi mahasiswa untuk bekal ilmu kepada mahasiswa sehingga menghasilkan alumni dengan kompetensi sesuai dengan pasar kerja yang tersedia.

Subspace clustering diproyeksikan sebagai teknik pencarian untuk mengelompokkan data atau atribut pada klaster yang berbeda, Pengelompokan dilakukan dengan menentukan tingkat kerapatan data dan juga mengidentifikasi outlier atau data yang tidak relevan, sehingga masing-masing cluster ada dalam subset tersendiri. Tesis ini mengusulkan inovasi algoritma subspace clustering based on density connection. Pada tahap awal akan dihitung kerapatan dimensi, hasil kerapatan dimensi akan dijadikan data masukan untuk menentukan klaster awal yang berdasarkan kerapatan dimensi, yakni dengan menggunakan Algoritma DBSCAN. Data pada setiap klaster kemudian akan diuji apakah memiliki hubungan dengan data pada klaster yang lain, yakni dengan menggunakan Algoritma SUBCLU. Jika data memiliki hubungan dengan data di klaster yang lain maka akan dikelompokkan sebagai sebuah subspace.

DBSCAN memiliki cara kerja clustering yang hampir mirip dengan DENCLUE. Secara signifikan, DBSCAN bekerja dengan efisien dalam membentuk arbitrary-shaped cluster. Pengelompokan dilakukan terhadap titik dengan ketetanggaannya yang berada di dalam jarak (ε) tertentu yang harus memenuhi jumlah titik minimum (minPts). Pembentukan ketetanggaan dapat ditentukan melalui pemilihan fungsi jarak antara dua buah titik. DBSCAN menggunakan konsep titik pusat (core point), titik batas (border point), dan noise. Titik yang memiliki sejumlah titik tetangga dan memenuhi jumlah titik minimum, serta berada dalam jarak tertentu disebut sebagai titik pusat, sedangkan titik batas memiliki jumlah titik tetangga namun tidak memenuhi jumlah titik minimum. Titik batas tersebut biasanya merupakan titik di dalam ketetanggaan dari titik pusat. Kriteria suatu titik dikatakan sebagai noise yaitu pada saat titik tersebut tidak termasuk titik pusat maupun titik batas, selain itu titik tersebut tidak memenuhi konsep directly density-reachable dari suatu titik pusat (Ester et al. 1996).

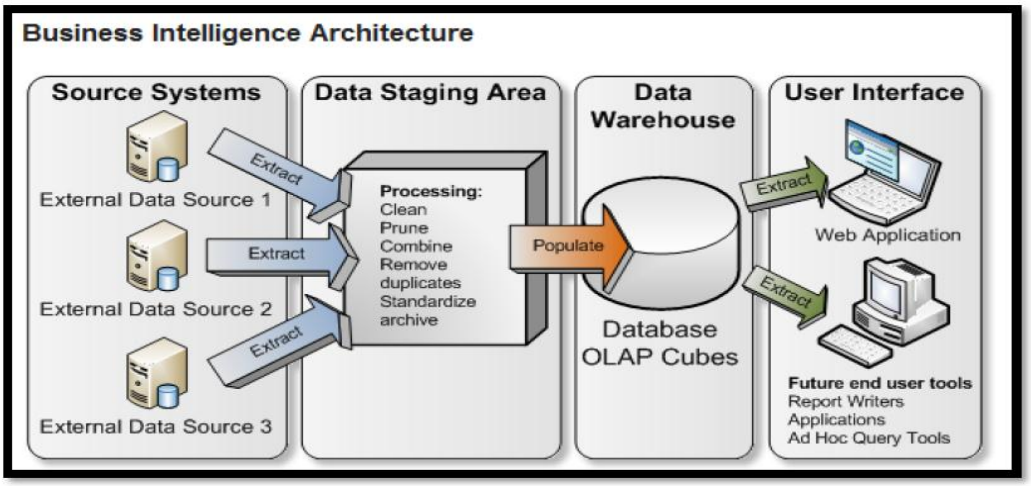
Clustering adalah salah satu alat data mining, pengelompokan penggunaan untuk membagi data ke dalam cluster yang bermakna atau berguna. Sebagian besar algoritma umum gagal untuk menghasilkan hasil yang berarti karena sparsity yang melekat pada benda. Dengan dimensi data yang tinggi, dan distribusi bola super, cenderung gagal dalam data terorganisir properti karakteristik pengelompokan data. Kesamaan, ketidaksamaan ukuran dan masalah kepadatan harus mendefinisikan. Algoritma klasterisasi tradisional sering gagal untuk mendeteksi cluster bermakna dan kesulitan menemukan kelompok bentuk sewenang-wenang dan penanganan outlier. Berkaitan dengan dataset nyata, penelitian ini akan mengeksplorasi cluster subspace yang ideal untuk menentukan prediksi pekerjaan untuk HLI.

Tulisan ini akan membahas data dimensi tinggi pengelompokan berdasarkan pengelompokan subspace. Metode seperti PCA, ICA, MDS, dan peta difusi linear atau nonlinearly menciptakan dimensi baru berdasarkan kombinasi informasi dalam ukuran aslinya. Dimensi ini baru dapat membuktikan sulit untuk menafsirkan, membuat hal yang sulit untuk dimengerti. Proyeksi clustering, atau subspace clustering, tantangan dimensi tinggi dengan membatasi pencarian dalam domain dari ruang data asli (Xu, p.253, 2009). Dimana disetiap metode mempunyai kelebihan yang berbeda-beda.

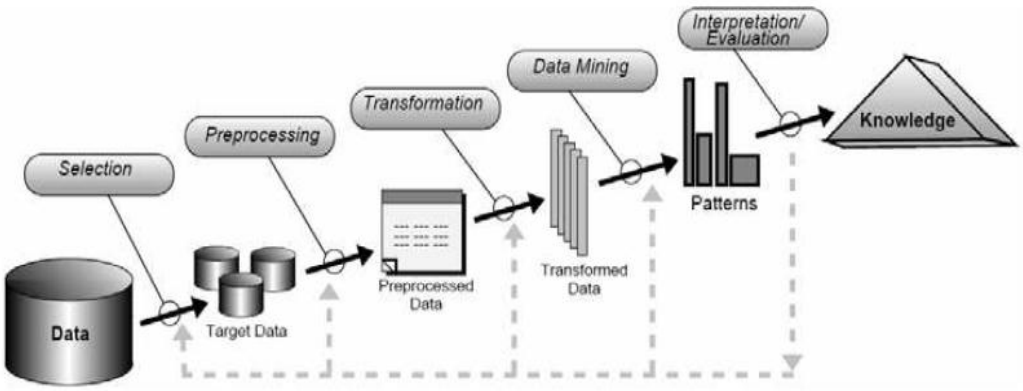
Pada bab ini membahas tentang sistem pendataan di segala sistem dan data–data yang digunakan dalam teknik data mining dan aplikasinya. Tesis ini fokus pada pengelompokan data multidimensi dalam deteksi cluster, dan analisis klaster. Untuk klaster subspace sangat penting untuk dibahas terutama untuk menentukan data berdasarkan klaster. Ada dua macam metode pengelompokan subspace yang akan dibahas dan membandingkan analisis subspace clustering menggunakan DBSCAN dan SUBCLU untuk proyeksi pekerjaan alumni perguruan tinggi dengan tujuan penulis pada data mining di bidang pendidikan, klasifikasi mahasiswa, dan aturan prediksi yang dapat diklasifikasikan di perusahaan.

Pesatnya perkembangan teknologi informasi dalam peningkatan kapasitas media penyimpanan, koneksi jaringan komputer besar akan mengakibatkan peningkatan penggunaan pengolahan data digital. Jumlah data yang lebih besar dan beragam jenis lainnya tentu sulit diolah menjadi informasi yang berguna. Struktur informasi umum dipengaruhi oleh beberapa faktor, termasuk penyimpanan, pemrosesan dan transmisi, seperti yang ditunjukkan pada gambar 1 (Berka, 2009). Untuk menghasilkan informasi yang berguna diperlukan pengolahan data yang baik. Pengolahan data dipengaruhi oleh teori komputasi, pemrograman, database, dan basis pengetahuan yang di identifikasi sebagai data yang kaya tapi miskin informasi, data yang hanya menghasilkan informasi yang sangat sedikit. Dalam kondisi pengetahuan yang minim suatu penemuan terbaru sangat perlu untuk diterapkan.

Jurnal Teknovasi
Volume 02, Nomor 01, 2015, 33 – 60
ISSN : 2355-701X



Gambar 1. Struktur sistem datamining



Gambar 2. Tahapan proses sistem

Data mining adalah metodologi analisis data yang telah sukses dalam banyak bidang. Biasanya *data mining* digunakan dalam dunia bisnis, dan masih sangat jarang digunakan dalam dunia pendidikan, oleh demikian menjadi sebuah tantangan bagi penulis untuk menggunakan *data mining* dalam bidang pendidikan. Perkembangan *world wide web (www)* yang menghasilkan banyak data dengan banyak topik, sehingga menjadikan *data mining* berkembang dengan pesat dan sangat relevan digunakan untuk mengekplorasi informasi dari dunia maya.

Beberapa penerapan *data mining* dalam pendidikan antara lain adalah untuk memprediksi tipologi *learning outcome*, dan memprediksi alumni yang memberikan komitmen paling menjanjikan. Penerapan lainnya adalah untuk menganalisa aktivitas dalam dunia pendidikan misalnya untuk mengevaluasi aktivitas, sistem pembelajaran, dan membentuk model interaksi antara mahasiswa dengan sistem. *Data mining* juga telah digunakan sebagai metode untuk mengekstraksi penilaian (*assessment*) jangka pendek sehingga dapat diketahui rata-rata waktu mengerjakan tugas *Data mining* telah digunakan untuk mengeksplorasi data alumni, menemukan alumni manakah yang berpotensi untuk memberikan donasi yang besar bagi almamater. Selain itu Merceron dan Yacef juga mengevaluasi model pembelajaran yang menarik bagi mahasiswa dan dosen dalam perspektif pedagogi Tabel 1 menunjukkan beberapa pertanyaan kritis dalam dunia perguruan tinggi yang ekuivalen dengan pertanyaan dalam lingkup bisnis pribadi, yang dapat dijawab oleh data mining

Jurnal Teknovasi
Volume 02, Nomor 01, 2015, 33 – 60
ISSN : 2355-701X

Tabel 1. Beberapa permasalahan dalam lingkup pribadi yang umum dengan masalah dalam perguruan tinggi.

Lingkup Pribadi	Perguruan Tinggi
Siapakah pelanggan yang paling memberikan keuntungan	Siapakah mahasiswa yang paling banyak mengambil sks
Siapakah pengunjung website yang paling banyak berkunjung?	Mahasiswa manakah yang paling sering mengulang matakuliah
Siapakah pelanggan yang paling setia	Siapakah mahasiswa paling gigih di universitas
Pelanggan manakah yang cenderung menaikkan pesananya	Siapakah alumni yang paling banyak memberi donasi paling besar

Luan, J berpendapat bahwa perguruan tinggi akan lebih menemukan aplikasi yang cukup besar dan luas bagi *data mining* daripada penerapannya dalam dunia bisnis *(www.cabrillo.edu/services/pro/oir_reports/ UCSFpaper.pdf.).*

Hal tersebut dikarenakan perguruan tinggi mengemban 3 tugas utama yang secara intensif dapat menjadi wahana yang tepat bagi penerapan *data mining.* Ketiga hal tersebut adalah penelitian yang berhubungan dengan penemuan ilmu pengetahuan ilmu pengetahuan baru, pembelajaran yang berkaitan dengan proses transfer pengetahuan, dan penelitian institusi dalam kaitannya dengan penerapan pengatahuan untuk pengambilan keputusan.

Banyak hal yang dapat diprediksikan informasi masa depannya dengan menggunakan *data mining*. Beberapa hal berikut merupakan garapan yang menarik dikaji dengan menggunakan *data mining*.

a. Pelacakan alumni. Beberapa penelitian menunjukkan bahwa *data mining* memberikan hasil yang memuaskan dari data pelacakan alumni. Hasil tersebut akan memberikan kontribusi positif bagi pengembangan institusi di masa mendatang.
b. Memprediksi kebutuhan *stakeholder*. Lulusan sebuah perguruan tinggi akan cepat diserap oleh pasar kerja jika kemampuannya sesuai dengan kebutuhan *stakeholder*. Data mining dapat menjawab tantangan prediksi kebutuhan *stakeholder* berdasarkan basis data yang dimiliki perguruan tinggi.
c. Memprediksi tingkat kualitas calon mahasiswa baru. Proses penjaringan mahasiswa baru dari tahun ke tahun akan meninggalkan sejumlah data data calon mahasiswa, yang dapat digunakan untuk melihat seperti apakah kualitas calon mahasiswa baru di sebuah perguruan tinggi di masa mendatang.
d. Memprediksi tingkat kualitas lulusan. Proses pembelajaran yang terjadi di perguruan tinggi membuat terkumpulnya data-data akademik dari mahasiswa, yang jika dikaji lebih mendalam dapat dimanfaatkan untuk mengetahui pola kualitas lulusan perguruan tinggi.
e. Tingkat serapan pasar kerja. Data alumni perguruan tinggi yang telah bekerja dari tahun ke tahun, selayaknya menjadi perhatian bagi pengambil kebijakan di perguruan tinggi sehingga dapat ditentukan tingkat serapan pasar kerja terhadap lulusan perguruan tinggi tersebut.

Memperhatikan luasnya cakupan yang diemban oleh perguruan tinggi, tentu masih banyak potensi informasi yang dapat digali di sebuah perguruan tinggi. Dengan begitu sudah saatnya perguruan tinggi memanfaaatkan teknik pengambilan keputusan yang lebih akurat guna menjawab tantangan derasnya arus informasi di abad ini.

Fungsi dan Tugas Data Mining

Data mining menganalisis data menggunakan tool untuk menemukan pola dan aturan dalam himpunan data. Perangkat lunak bertugas untuk menemukan pola dengan mengidentifikasi aturan dan fitur pada data. Tool Data mining diharapkan mampu mengenal pola ini dalam data dengan input minimal dari user. Dalam penelitian ini pembahasan Data Mining diklasifikasikan dalam fungsi Association.

Ada dua aturan pengukuran untuk '*association rule*' :

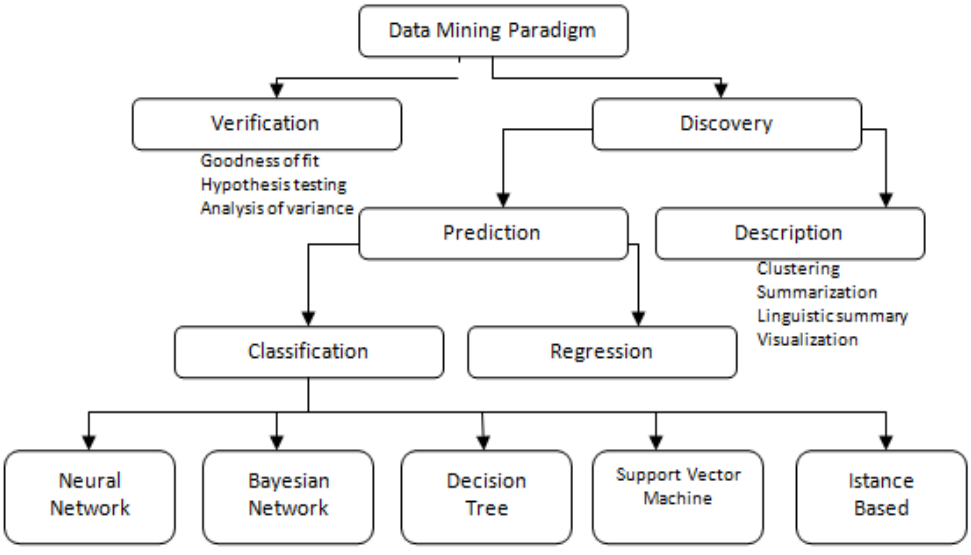
1. Support
   Support untuk himpunan item adalah prosentase transaksi yang berisi semua item-item ini. Support untuk aturan LHS RHS di-support untuk himpunan item-item LHS RHS.
2. Confidence
   Pertimbangkan transaksiyang berisi semua item dalam LHS. Confidence untuk rule : LHS RHS adalah prosentasi transaksi yang juga terdiri semua item-item dalam RHS.

Lebih tepatnya, misalkan sup (LHS) adalah prosentase transaksi yang berisi LHS dan sup (LHS RHS) adalah prosentase transaksi yang berisi LHS dan RHS, maka confidence rule LHS RHS adalah sup (LHS RHS)/sup (LHS). Permasalahan Association Rule dapat dikomposisikan menjadi dua sub masalah, yaitu:

1. Penemuan semua kombinasi item-item, yang disebut frequent-item set, yang support-nya lebih besar daripada minimum support.
2. Gunakan frequent-item set untuk membangkitkan aturan yang diinginkan. Idenya adalah, katakan, ABCD dan AB sering muncul dalam transaksi, maka aturan AB CD akan dipenuhi jika perbandingan antara support (ABCD) terhadap support (AB) minimum sebesar minimum confidence . Semua rule akan mempunyai minimum support karena ABCD sering muncul dalam transaksi

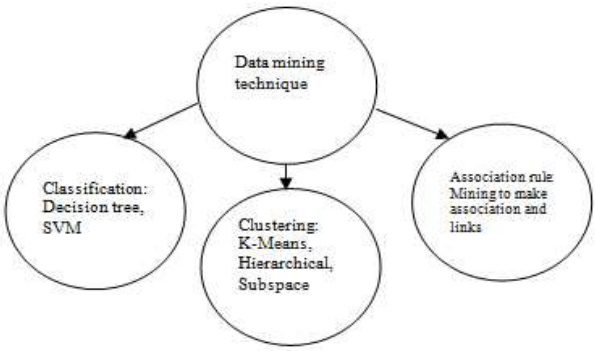
Teknik Pertambangan Data dan Aplikasinya

Teknik Data mining terdiri dari enam kelas umum kegiatan: deteksi anomali, aturan asosiasi belajar, clustering, klasifikasi, regresi, dan summarization. Data mining adalah persimpangan interdisipliner kecerdasan buatan, pembelajaran mesin, statistik, dan sistem database. Beberapa tahun terakhir tren data mining meliputi distribusi data mining, hypertext/hypermedia pertambangan data mining, serta multimedia, spasial, time series, dan data mining sekuensial (Hsu, 2002). Upaya utama dalam data mining adalah untuk mengekstrak pengetahuan dari data. sedangkan data mining taksonomi ditunjukkan pada gambar 3 (Maimon 2005).



Gambar 3. Struktur data mining taksonomi

Umumnya, teknik data mining gambar 4 didasarkan pada logika induktif, penalaran statistik, pemrograman, fuzzy set, pembelajaran mesin dan teknik jaringan syaraf. Berdasarkan hipotesis informasi dari dataset akan mengekstrak dan diamati. Pola yang muncul akan mengamati untuk menjawab atau studi tentang aturan penemuan untuk partisi data ke dalam kelompok tertentu dan membuat asosiasi antara data, atau menemukan aturan data yang disesuaikan.



Gambar 4. Teknik dalam data mining

Di masa depan penggunaan data mining akan semakinluas dari internet, nirkabel gadget, dan akan memanfaatkan sejumlah besar data. Pra-pengolahan akan menjadi bagian penting dari data mining, cepat dan transparan (Kriegel, 2007).

Clustering

Clustering adalah sebuah metode untuk mengelompokkan beberapa macam obyek yang serupa (*similar*) kedalam *class-class*. Sebuah *cluster* adalah sekumpulan data yang mirip satu sama lain dan tidak mirip dengan data-data pada *cluster* lain. *Clustering* berbeda dengan klasifikasi karena pada *clustering* tidak ada *class-class* target yang telah diset sebelumnya. *Clustering* algoritma akan berusaha membagi data yang ada menjadi kelompok-kelompok data dimana data pada kelompok (*cluster*) yang sama relatif lebih homogen bila dibandingkan dengan data-data pada kelompok lain. *Clustering* berusaha memaksimalkan kesamaan (*similarity*) dari data-data pada *cluster* yang sama dan meminimalkan kesamaannya dengan data-data pada cluster lainnya (Larose, 2005).

Analisa *cluster* adalah suatu teknik analisa *multivariate* (banyak variabel) untuk mencari dan mengorganisir informasi tentang variabel tersebut sehingga secara relatif dapat dikelompokkan dalam bentuk yang homogen dalam sebuah *cluster*. Secara umum, bisa dikatakan sebagai proses menganalisa baik tidaknya suatu proses pembentukan *cluster*. Analisa *cluster* bisa diperoleh dari kepadatan *cluster* yang dibentuk (*cluster density*). Kepadatan suatu *cluster* bisa ditentukan dengan *variance within cluster* (Vw) dan *variance between cluster* (Vb). Varian tiap tahap pembentukan *cluster* bisa dihitung dengan rumus:

$$Vc^2 = \frac{1}{n_c - 1}\sum_{i=1}^{n}\left(y_i - \bar{y}_c\right)^2$$

Dimana:

Vc^2 = varian pada *cluster c*
c = *1..k, dimana k = jumlah cluster*
n_c = jumlah data pada *cluster c*
yi = data ke-*i* pada suatu cluster
yi = rata-rata dari data pada suatu *cluster*

Selanjutnya dari nilai varian diatas, kita bisa menghitung nilai *variance within cluster* (Vw) dengan rumus:

$$Vw = \frac{1}{N - c}\sum_{i=1}^{c}\left(n_i - 1\right)\cdot Vi^2$$

Dimana, N = Jumlah semua data
n_i = Jumlah data *cluster* i
Vi = Varian pada *cluster* i

Dan nilai *variance between cluster* (Vb) dengan rumus:

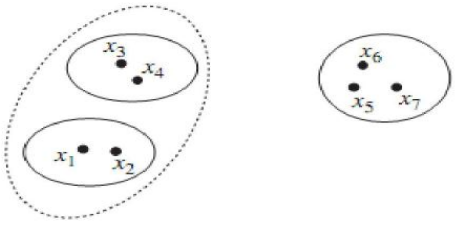
$$Vb = \frac{1}{c-1}\sum_{i=1}^{c} n_i \left(\overline{y}_i - \overline{y}\right)^2$$

Dimana, y = rata-rata dari yi

Salah satu metode yang digunakan untuk menentukan *cluster* yang ideal adalah batasan *variance*, yaitu dengan menghitung kepadatan *cluster* berupa *variance within cluster* (Vw) dan *variance between cluster* (Vb). *Cluster* yang ideal mempunyai Vw minimum yang merepresentasikan *internal homogenity* dan maksimum Vb yang menyatakan *external homogenity.*

$$V = \frac{Vw}{Vb}$$

Clustering adalah metode data mining yang Unsupervised, karena tidak ada satu atributpun yang digunakan untuk memandu proses pembelajaran, jadi seluruh atribut input diperlakukan sama. Kebanyakan Algoritma Clustering membangun sebuah model melalui serangkaian pengulangan dan berhenti ketika model tersebut telah memusat atau berkumpul (batasan dari segmentasi ini telah stabil).



Gambar 5. Contoh dalam clustering

Algoritma Clustering

Pendekatan alternative untuk menentukan clustering yang paling sesuai dengan seperangkat data x adalah dengan mempertimbangkan semua clustering yang mungkin dan pilih salah satu yang paling masuk akal sesuai dengan kriteria dan rasionalitas. Sebagai contoh, seseorang dapat memilih clustering yang mengoptimasi kriteria yang terpilih, mengkuantisasi vektor-vektor yang lebih mirip kedalam satu kelas yang sama dan vektor-vektor yang kurang mirip kedalam kelas yang berbeda. Namun, jumlah semua clustering yang mungkin terjadi adalah besar, bahkan untuk sejumlah pola N yang tidak terlalu banyak. Cara untuk mengatasi masalah ini adalah dengan mengembangkan algoritma clustering, yang hanya mempertimbangkan sebagian kecil dari clustering yang mungkin terjadi. Pertimbangan clustering tergantung pada prosedur algoritma yang spesifik.

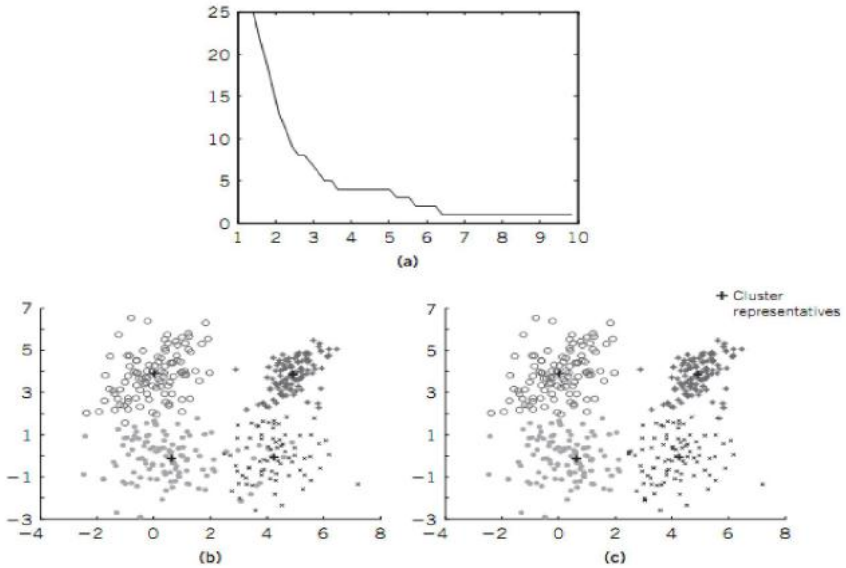
Beberapa algoritma clustering telah dikembangkan, beberapa diantaranya merupakan clustering tunggal, dan yang lainnya adalah clustering hierarki. Klasifikasi berikut berisi sebagian besar algoritma clustering yang terkenal.

Algoritma clustering tunggal meliputi :

a. Sequential algorithms, dengan konsep sederhana, bekerja pada seperangkat data tunggal atau data yang sangat sedikit.
b. Cost function optimization algorithms, yang mengadopsi fungsi biaya J dengan mengkuantisasi istilah masuk akal (sensible) dan menghasilkan clustering dengan optimasi J. Yang termasuk dari kategori ini adalah hard clustering algoritms seperti k-means, fuzzy clustering algoritms seperti fuzzy c-means (FCM), probabilistic clustering algoritms seperti EM dan probabilistic algoritm.
c. Miscellaneous algorithms, yang tidak sesuai dengan kategori sebelumnya, sebagai contoh competitive learning algorithms, valley-seeking algorithms, density-based algorithms, and subspace-clustering algorithms.

Algoritma clustering hierarki meliputi :

a. Agglomerative algorithms, yang menghasilkan clustering sekuensial dari pengurangan sejumlah kelas, m. Pada setiap tahap, pasangan kelas terdekat pada clustering saat ini diidentifikasi dan digabung menjadi satu dalam rangka untuk membangkitkan clustering berikutnya.
b. Divisive algorithms, yang berbeda dengan agglomerative algorithms, menghasilkan clustering sekuensial dari penambahan sejumlah kelas. Pada setiap tahap, sebuah kelas yang telah dipilih dibagi menjadi dua kelas yang lebih kecil.



Sumber : Theo Chapter.

Gambar 6. Proses clustering data

Subspace Clustering

Bottom up subspace clustering yang dimulai dari semua subruang satu dimensi yang mengakomodasi setidaknya satu cluster dengan menggunakan strategi pencarian yang mirip dengan algoritma pertambangan set item yang sering. CLIQUE merupakan perwakilan dari bottom up subspace clustering.

CLIQUE (Kailing, 2009) mengidentifikasi kelompok padat dalam domain dari dimensi maksimum. Setelah subruang yang tepat ditemukan, tugas ini adalah untuk menemukan cluster dalam proyeksi yang sesuai. Titik data dipisahkan sesuai dengan lembah fungsi kepadatan. Cluster adalah serikat unit kepadatan tinggi yang terhubung dalam subruang, kemudian akan menghasilkan deskripsi klaster dalam bentuk ekspresi DNF yang diminimalkan untuk kemudahan pemahaman. Ini menghasilkan hasil identik terlepas dari urutan catatan masukan disajikan dan tidak menganggap bentuk matematika tertentu untuk distribusi data.

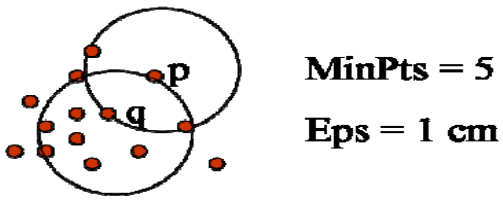
CLIQUE mulai dari mengidentifikasi subruang yang mengandung cluster. Pada fase ini dapat menemukan unit yang padat, dengan menentukan unit pertama padat 1dimensi dengan membuat lulus atas data. Setelah menetapkan (k-1)-dimensi unit padat, calon unit k-dimensi ditentukan dengan menggunakan prosedur generasi calon diberikan di bawah ini. Sementara prosedur saja dijelaskan secara dramatis mengurangi jumlah unit yang diuji untuk menjadi padat, kita mungkin masih memiliki tugas komputasi tidak layak di tangan untuk data dimensi tinggi. Sebagai dimensi dari subruang dianggap meningkat, ada ledakan dalam jumlah unit yang padat, dan jadi kita perlu memangkas set unit padat ini kemudian digunakan untuk membentuk unit calon di tingkat berikutnya dari algoritma generasi satuan padat. Setelah mengidentifikasi subruang mengandung klaster, diikuti dengan mengidentifikasi cluster dan generasi deskripsi minimal untuk cluster.

Jurnal Teknovasi
Volume 02, Nomor 01, 2015, 33 – 60
ISSN : 2355-701X

DBSCAN

DBSCAN adalah salah satu algoritma *clustering density-based*. Algoritma memperluas wilayah dengan kepadatan yang tinggi ke dalam *cluster* dan menempatkan *cluster* irregular pada database spasial dengan *noise*. Metode ini mendefiniskan *cluster* sebagai *maximal set* dari titik-titik yang *density-connected*. DBSCAN memiliki 2 parameter yaitu *Eps* (radius maksimum dari *neighborhood*) dan *MinPts* (jumlah minimum titik dalam *Eps-neighborhood* dari suatu titik). Ide dasar dari *density-based clustering* berkaitan dengan beberapa definisi baru:

1. *Neighborhood* dengan radius *Eps* dari suatu obyek disebut *Epsneighborhood* dari suatu obyek tersebut
2. Jika *Eps-neighborhood* dari suatu obyek mengandung titik sekurangkurangnya jumlah minimum, *MinPts*, maka suatu obyek tersebut dinamakan *core object*
3. Diberikan set obyek *D*, obyek *p* dikatakan *directly density-reachable* dari obyek *q* jika p termasuk dalam *Eps-neighborhood* dari *q* dan *q* adalah *core objek*.



Sumber : Arthur (2010)

Gambar 7. *Eps-neighborhood*

SUBCLU

SUBCLU (density terhubung subspace pengelompokan) menggunakan konsep-density konektivitas yang mendasari algoritma DBSCAN, SUBCLU didasarkan pada gagasan pengelompokan formal. Berbeda dengan pendekatan berbasis grid yang ada, SUBCLU mampu mendeteksi cluster yang tumpang tindih dibentuk dan diposisikan dalam domain. Monotonisitas of-density konektivitas digunakan untuk efisien memangkas ruang bagian dalam proses menghasilkan semua cluster dalam cara bottom up.

Top-down metode pengelompokan subspace menganalisis ruang dimensi penuh untuk menemukan pola bercak cluster, dimana setiap objek database beberapa pengelompokan bermakna mungkin ada. Subruang dimana cluster eksis diidentifikasi berdasarkan distribusi data seputar pola. Multi-resolusi Korelasi deteksi Cluster, sebagai metode scalable untuk mendeteksi cluster korelasi dalam kisaran sekitar 5 sampai 30 sumbu (Cordeiro, 2010), sedangkan mendeteksi cluster subruang alternatif yang didasarkan pada yang sudah dikenal subspace pengelompokan bisa deteksi cluster subspace alternatif , klaster berlebihan non dan memiliki klaster alternatif (Gunnemanns, 2010). Gunnemanns diusulkan sebagai alternatif ASCLU subspace clustering, idenya berdasarkan cluster subruang C = (O, S) adalah seperangkat benda O ⊆ DB dan satu set dimensi S ⊆ Dim. Obyek O serupa dalam dimensi yang relevan S sedangkan dimensi Dim\S tidak relevan untuk cluster. k-berarti algoritma mungkin untuk menggeneralisasi pengelompokan data dimensi tinggi, seperti yang diusulkan dalam GKM (Generalized k-mean). GKM menggunakan keuntungan dari k-berarti sewenang-wenang, memilih titik data k di X sebagai pusat klaster awal, masing-masing pusat cluster i C dikaitkan dengan vektor i W komponen yang sama satu, kemudian mengulangi langkah-langkah untuk mengoptimalkan tujuan fungsi E (W, C).

Algoritma SUBCLU (Kailing, 2004) didasarkan pada bottom-up, algoritma serakah untuk mendeteksi cluster kepadatan yang terhubung dalam semua subruang data dimensi tinggi. Algoritma dimulai dengan menghasilkan semua cluster 1dimensi dengan menerapkan DBSCAN kepada setiap subruang 1-dimensi. Untuk setiap terdeteksi klaster kita harus memeriksa, apakah klaster ini masih ada dalam domain dimensi yang lebih tinggi. Tidak ada kelompok lain yang bisa eksis dalam domain dimensi yang lebih tinggi. Untuk setiap subruang k-dimensi, mencari semua subruang k-dimensi lain yang memiliki (k-1) atribut yang sama dan bergabung dengan mereka untuk menghasilkan (k+1)-dimensi subruang calon. Himpunan subruang calon (k+1)-dimensi

Jurnal Teknovasi
Volume 02, Nomor 01, 2015, 33 – 60
ISSN : 2355-701X

dinotasikan dengan. Untuk setiap subruang kandidat mengandung setiap k-dimensi subruang T ∈ S (JTJ = k), kemudian memangkas kandidat ini memiliki subspace setidaknya satu k-dimensi tidak termasuk dalam Sk. Hal ini akan mengurangi jumlah (k+1)-dimensi subruang calon.

METODE PENELITIAN

Subjek Penelitian

Subjek penelitian ini adalah para alumni Jurusan teknik komputer Medicom kota medan yang telah lulus pada tahun 2005-2010, baik yang berada di medan maupun yang berada di luar kota medan, yang tersebar di seluruh Indonesia. Subyek penelitian adalah khusus bagi mahasiswa reguler.

Rancangan Penelitian

Dalam metodologi penelitian ini menjelaskan kajian proses pembuatan dan desain dari berbagai teknik digunakan untuk melakukan penelitian, algoritma, tes, percobaan, survei, dan studi kasus.

Penulis membuat studi utama dari penelitian ini untuk menemukan pokok permasalah penelitian didalam tesis ini. Tinjauan kasus dan analisis diambil dari literatur sebagai titik awal untuk menyorot desain dan pengembangan subspace clustering yang berdasarkan koneksi kepadatan. Studi kasus dan percobaan akan mengidentifikasi dan mengevaluasi, untuk diterapkan dalam data mining pendidikan, terutama dalam kesempatan kerja. Pokok permasalahan terhadap studi kasus ini juga menyediakan dasar literatur untuk menentukan kriteria untuk mengembangkan instrumen penelitian. Pengumpulan data dan analisis instrumen yang digunakan untuk pengujian dan sesi pelatihan dalam tahap percobaan.

Alur Kerja Aplikasi

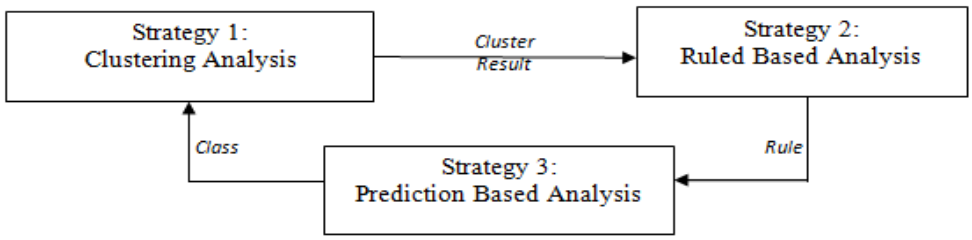
Aplikasi berbasis mempunyai alur kerja secara garis besar sebagai berikut :

1. Alumni mahasiswa dan perusahaan memasukkan NIM atau NIK dan password agar dapat menggunakan aplikasi ini
2. Program akan memeriksa NIM atau NIK dan password tersebut ke dalam *database*
3. Apabila NIM atau NIK dan password yang dimasukkan benar, maka dapat menggunakan aplikasi ini

Multidimensi dan Analisis strategi data

Data mining multidimensi melibatkan lima langkah proses: memutuskan antara sensus dan data sampel, mengidentifikasi hubungan dalam data, memodifikasi atau mengubah data, mengembangkan sebuah model yang menjelaskan hubungan data, dan pengujian akurasi model. Untuk menawarkan data yang lebih rinci untuk melengkapi dan terbaik memahami masalah penelitian, penelitian ini menggunakan tiga strategi analisis data mining multidimensi, termasuk analisis clustering, analisis berbasis peraturan, dan analisis berbasis prediksi gambar 8.

Clustering penelitian dengan menggunakan teknik pengelompokan diaplikasikan untuk menganalisis kemungkinan klaster antara keterampilan kompetensi pengetahuan dan kompetensi soft skill antara pelatihan industri siswa dan industri. Pendekatan berbasis diperintah ditindaklanjuti dengan beberapa kasus dari hasil clustering untuk menyelidiki hasil lebih. Akhirnya, analisis predictionbased digunakan untuk menjelaskan hasil dari pendekatan berbasis memerintah dan untuk membangun gambaran yang lengkap tentang penelitian, dan menggambarkan kesempatan kerja bagi Alumni Perguruan Tinggi.



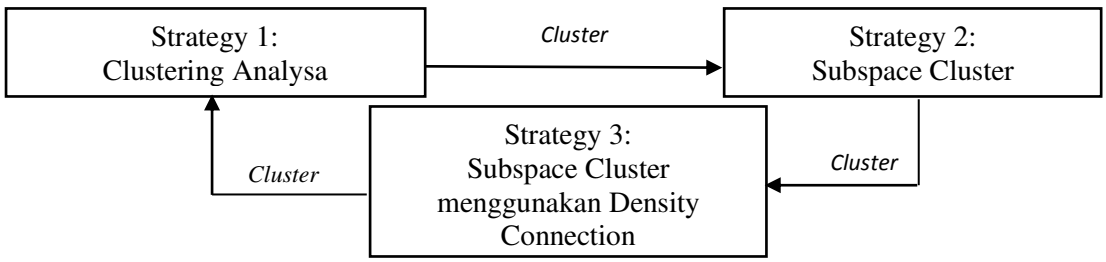
Gambar 8. Strategi analisis data mining multidimensi Penelitian

Analisis Klustering data

Manusia yang terampil membagi objek ke dalam kelompok-kelompok (dikenal sebagai pengelompokan), menempatkan objek tertentu kepada kelompok-kelompok (dikenal sebagai klasifikasi), dan kemudian melakukan prediksi untuk objek tertentu. Penggunaan pengelompokan dalam penelitian ini untuk memahami anggota objek untuk setiap set data. Ini prototipe cluster dapat digunakan sebagai dasar untuk sejumlah analisis data atau pengolahan data kelompok analisis data yang technique. Clustering benda hanya berdasarkan informasi yang ditemukan dalam data set yang menggambarkan objek dan hubungan mereka. Grup obyek harus sama satu sama lain Namun pada kenyataannya, ada banyak situasi di mana obyek cukup daripada yang bisa ditempatkan di lebih dari satu cluster. Dalam kasus ini, subruang klaster harus digunakan untuk memecahkan masalah.

Penelitian ini untuk mengidentifikasi objek tertentu dan tempat dalam cluster terkait, untuk memastikan hubungan metode yang digunakan DBSCAN sebagai teknik pengelompokan dasar. Ada beberapa teknik subspace klaster, seperti SUBCLU dan DBSCAN. Pengelompokan ini akan digunakan untuk melatih set data, penulis akan melatih setiap set data yang sampai hasil klaster konvergen. Sebagai hal baru dari penelitian ini kami mengusulkan sebuah algoritma clustering subspace ditingkatkan berdasarkan koneksi berbasis kepadatan, dan menggunakannya sebagai teknik pengelompokan utama prediksi pekerjaan Alumni Perguruan Tinggi. Untuk menawarkan data yang lebih rinci untuk melengkapi dan terbaik memahami masalah penelitian, penelitian ini menggunakan tiga strategi analisis data mining multidimensi, termasuk analisis clustering, analisis berbasis peraturan, dan analisis berbasis prediksi gambar 9.

Clustering penelitian dengan menggunakan teknik pengelompokan diaplikasikan untuk menganalisis kemungkinan cluster antara keterampilan kompetensi pengetahuan dan kompetensi soft skill antara pelatihan industri siswa dan industri. Pendekatan berbasis Diperintah ditindaklanjuti dengan beberapa kasus dari hasil clustering untuk menyelidiki hasil lebih. Akhirnya, analisis berbasis prediksi digunakan untuk menjelaskan hasil dari pendekatan berbasis memerintah



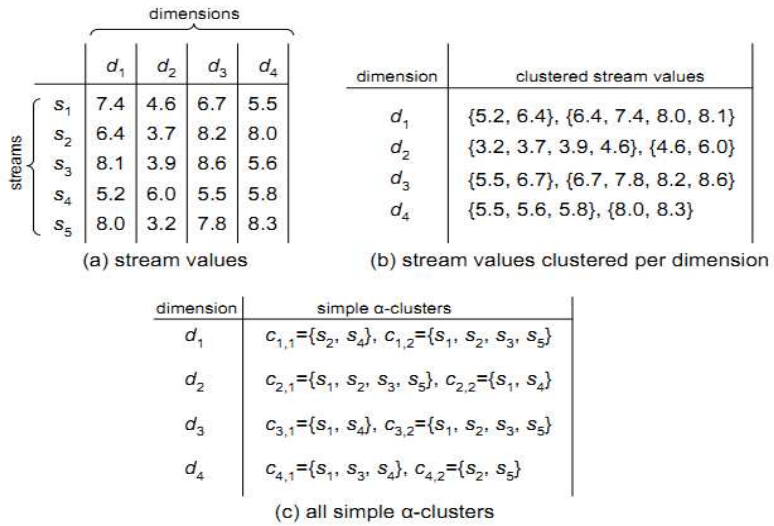
Gambar 9. Analisa strategi data mining.

Analisis Clustering

Manusia yang terampil membagi objek ke dalam kelompok-kelompok (dikenal sebagai pengelompokan), menempatkan objek tertentu kepada kelompok-kelompok (dikenal sebagai klasifikasi), dan kemudian melakukan prediksi untuk objek tertentu. Penggunaan DBSCAN dalam penelitian ini untuk memahami anggota objek untuk setiap set data. Ini prototipe cluster dapat digunakan sebagai dasar untuk sejumlah analisis data atau teknik pengolahan data.

Tujuan dari inisialisasi klaster adalah untuk menen-tukan satu set awal subruang maksimal α-cluster, berdasarkan pada nilai-nilai W terakhir setiap kali seri streaming. Proses CI terdiri dari serangkaian langkah-langkah. Pada langkah pertama, setiap kali instance (dimensi) diperiksa secara terpisah untuk mencegah-tambang sederhana α-cluster (yang didefinisikan dalam satu dimensi saja). Selanjutnya, semua cluster yang mengandung m = 2 aliran dalam jumlah maksimum yang mungkin dari dimensi yang dihasilkan.

Dalam setiap langkah berikutnya algoritma mencoba untuk meningkatkan jumlah aliran per cluster (m = m +1), sampai semua kemungkinan subruang maksimal α-cluster yang dihasilkan, sesuai dengan nilai-nilai α, minRows dan minCols. Cluster yang mengandung kurang dari dimensi minCols dibuang secara permanen di setiap langkah algoritma, karena mereka tidak dapat berkontribusi pada jawaban akhir.



	d_1	d_2	d_3	d_4
s_1	7.4	4.6	6.7	5.5
s_2	6.4	3.7	8.2	8.0
s_3	8.1	3.9	8.6	5.6
s_4	5.2	6.0	5.5	5.8
s_5	8.0	3.2	7.8	8.3

dimension	clustered stream values
d_1	{5.2, 6.4}, {6.4, 7.4, 8.0, 8.1}
d_2	{3.2, 3.7, 3.9, 4.6}, {4.6, 6.0}
d_3	{5.5, 6.7}, {6.7, 7.8, 8.2, 8.6}
d_4	{5.5, 5.6, 5.8}, {8.0, 8.3}

dimension	simple α-clusters
d_1	$c_{1,1}=\{s_2, s_4\}$, $c_{1,2}=\{s_1, s_2, s_3, s_5\}$
d_2	$c_{2,1}=\{s_1, s_2, s_3, s_5\}$, $c_{2,2}=\{s_1, s_4\}$
d_3	$c_{3,1}=\{s_1, s_4\}$, $c_{3,2}=\{s_1, s_2, s_3, s_5\}$
d_4	$c_{4,1}=\{s_1, s_3, s_4\}$, $c_{4,2}=\{s_2, s_5\}$

Gambar 10. Inisialisasi Kluster

Data benda kelompok analisis pengelompokan hanya berdasarkan informasi yang ditemukan dalam set data yang menggambarkan objek dan hubungan mereka. Grup obyek harus sama satu sama lain. Namun pada kenyataannya, ada banyak situasi di mana obyek cukup daripada yang bisa ditempatkan di lebih dari satu cluster. Dalam kasus ini, subruang klaster harus digunakan untuk memecahkan masalah.

Karena masalah penelitian, dalam cluster subspace penelitian untuk mengidentifikasi objek tertentu dan tempat dalam cluster terkait, untuk memastikan hubungan yang kita gunakan DBSCAN sebagai teknik pengelompokan dasar. Ada beberapa teknik subspace klaster, seperti SUBCLU. Pengelompokan ini akan digunakan untuk melatih set data. Kami akan melatih setiap set data yang sampai hasil klaster konvergen. Sebagai hal baru dari penelitian ini kami mengusulkan sebuah algoritma clustering subspace ditingkatkan berdasarkan koneksi berbasis density (disebut sebagai Damira), dan menggunakannya sebagai teknik pengelompokan utama prediksi pekerjaan Lembaga Higher Learning.

Analisis Subspace Clustering

Hasil pengelompokan pada satu strategi sehingga di atas akan diproses oleh pertambangan aturan asosiasi, untuk menemukan aturan asosiatif antara kombinasi tertentu. Sebuah jenis tertentu penalaran yang menggunakan "if-then-else" pernyataan aturan akan dilaksanakan, hanya sebagai pola dan pencarian mesin inferensi untuk pola dalam aturan yang sesuai pola dalam data. Penelitian ini akan menerapkan subspace klaster, sebagai operator belajar dari kedua data nominal dan numerik.

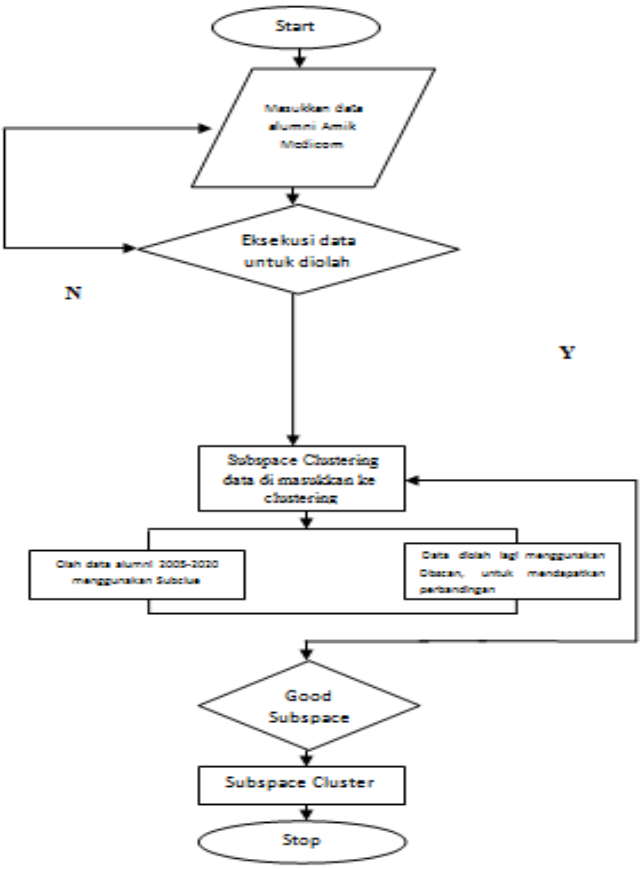
Subspace Cluster Berbasiskan Density Connection.

Saat aturan telah ditemukan, penelitian ini akan menggunakan koneksi berbasis density untuk mengidentifikasi dan mencari peluang kerja. Hasil ini digunakan sebagai pengetahuan Model

basis untuk mewakili objek dalam informasi. Model menangkap hubungan antara banyak faktor untuk memungkinkan penilaian potensi terkait dengan set kondisi tertentu.

RancanganPenelitian

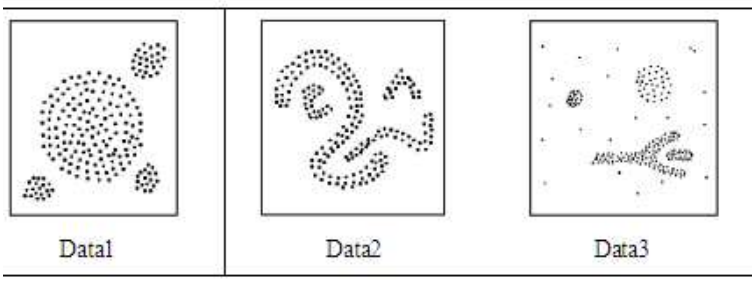
Langkah - langkah penelitian dapat dilihat pada gambar 11.



Gambar 11. Langka-langkah Penelitian

Kepadatan Data

Titik utama dari tesis ini adalah pendekatan pengelompokan kepadatan berbasis, khususnya konsep cluster kepadatan terhubung mendasari algoritma DBSCAN (Density-Based Clustering Spasial Aplikasi dengan Noise). Kami mengusulkan teknik ditingkatkan untuk mengatasi tantangan clustering di data mining pendidikan, yaitu sebagai Damira (multidimensi data mining subruang Pendekatan Clustering). Penelitian data mining dimulai dari kepadatan berbasis. Seperti yang terlihat pada gambar di gambar 12, dengan mudah mengidentifikasi sekelompok poin dan juga mengidentifikasi outlier yang terbentuk (Ester et al, 1996). Sebagai cluster titik-titik memiliki kepadatan mendekati poin dari yang lain. Sementara titik di luar kelompok disebut sebagai kebisingan.

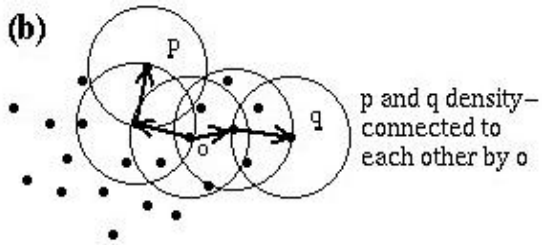


Sumber : Ester, (1996)

Gambar 12. Density Based Cluster

Alasan utama mengapa *cluster-cluster* pada gambar 12 dapat dibentuk adalah karena kepadatan *point-point* data pada sebuah *cluster* relatif lebih padat bila dibandingkan dengan *point-point* data diluar *cluster* (Ester, 1996).

Gagasan utama kepadatan di setiap cluster adalah bahwa data memiliki jumlah minimum data tetangga, di mana kerapatan data harus lebih dari batas tertentu. Bentuk tetangga akan ditentukan berdasarkan fungsi jarak dari titik p dan q, dinotasikan sebagai dist (p, q). Ada sejumlah minimum poin (MinPts) dalam cluster dalam Eps-tetangga, maka akan ada dua jenis poin dalam cluster, poin inti dan titik batas seperti gambar 13. Jumlah titik di perbatasan akan kurang dari pada intinya, tetapi harus menetapkan jumlah minimum poin termasuk dalam cluster yang sama. Nilai ini, bagaimanapun, tidak akan membentuk karakteristik yang spesifik untuk setiap cluster, terutama jika ada kebisingan. Oleh karena itu, diperlukan untuk setiap titik p dalam gugus C dan q di C sehingga p adalah di Eps-tetangga q dan N (q) mengandung setidaknya MinPts jumlah poin.



Gambar 13. Kepadatan data yang saling terhubung

HASIL DAN PEMBAHASAN

Implementasi Data

Data yang diambil adalah data alumni kampus teknik informatikan MEDICOM medan, dimana data alumni yang diambil dibatasin hanya 5 tahun yaitu stambuk 2005-2010, dan data yang didapat sekitar 10.000 alumni, dimana tujuan dari thesis yaitu memproyeksi alumni yang terserap pekerjaan di perusahaan sesuai bidang, adapun tampilan data yang diolah di tampilkan di tabel 2.

Tabel 2. Contoh Data daripada mahasiswa 2005-2010

No	Nama Alumni	Asal Sekolah	Tamat Medicom	Tempat Bekerja	I.P	Keputusan
1	Abdi Sahputra, A.Md	SMAN 18 Medan	2009	Dinas Perhubungan Laut	3.00	Sesuai
2	Afni, A.Md	SMA Medan Putri	2005	Dinas Pertamanan Tebing Tinggi	2.8	Sesuai
3	Afri Simaremare, A.Md	SMK Abdi Sejati	2008	Korem 002/Bb Pematang Siantar	3.00	Sesuai
4	Asrina Pandiangan, A.Md	SMA Budi Murni-2 Medan	2008	Sekretariat Komisi Pemilihan Umum	2.7	Sesuai
5	Bella Rina Barus, A.Md	SMAN 17 Medan	2009	PT.PLN Persero Samosir	3.2	Sesuai
6	Cepilian Erina Sijabat, A.Md	SMKN-3 Medan	2009	Kementrian Perhubungan Laut	3.00	Sesuai
7	Daniel Ginting, A.Md	SMAN-7 Medan	2005	Kejaksaan Negeri Pematang Siantar	2.78	T. Sesuai
8	Daniel Sihombing, A.Md	SMA HKBP Sidorame Medan	2005	Dinas Tata Kota Binjai	3.00	T. Sesuai

9	Elisabeth Sianipar, A.Md	SMAN Medan	2007	Biro Keuangan Dan Perbendaharaan Pemprovsu	2.98	Sesuai
10	Evi Jayati, A.Md	SMAN 7 Medan	2008	Kantor Imigrasi Medan	3.22	T. Sesuai
11	Fitri Simanjuntak, A.Md	SMA HKBP Sidorame Medan	2006	MABES POLRI Jakarta	3.00	Sesuai
12	Henny Yunita Gultom, A.Md	SMAN 1 Medan	2006	Departemen Kejaksaan Pemkot Jambi	3.00	Sesuai
13	Ita Sasalina Purba, A.Md	SMA Al Azhar Medan	2007	Satuan Kesdam I BB	2.8	Sesuai
14	Jaka Priatama Ginting, A.Md	SMAN-15 Medan	2010	Administrasi Pembangunan Setda Kota Medan	3.00	Sesuai
15	Julfan Saragih, A.Md	SMA Dwi Warna Medan	2005	Departemen Kejaksaan Negeri Medan	2.7	Sesuai
16	Marnaek Sirait, A.Md	SMAN 18 Medan	2009	Badan Kesatuan Bangsa Kab Deli Serdang	3.2	Sesuai
17	Nancy Silitonga, A.Md	SMA Trisakti Medan	2008	Kepegawaian Pemerintahan Dan Pendidikan Siantar	3.00	Sesuai
18	Putra Sarialim, A.Md	SMAN 7 Medan	2009	Dinas Pendidikan Sman 7 Medan	2.78	Sesuai
19	Putri Widyasti, A.Md	SMA Kartika	2009	Pemerintahan Kabupaten Langkat	3.00	T. Sesuai
20	Rama Fitria, A.Md	SMA Sampali Medan	2005	Dinas Kehutanan	2.98	T. Sesuai
21	Rensus Simanjuntak, A.Md	SMK Putra Anda Binjai	2006	Dinas Pendidikan Smpn 4 Tebing Tinggi	3.22	Sesuai
22	Roy Berutu, A.Md	SMA Dharma Bakti Medan	2008	Dinas Kebudayaan Pakpak Barat	3.00	T. Sesuai
23	Sofian Winata, A.Md	SMK Telkom Sandy Putra	2007	Kejaksaan Negeri Pematang Siantar	2.78	Sesuai
24	Sri Melina, A.Md	SMKN 10 Medan	2006	Pemkab Serdang Bedagai	3.00	Sesuai
25	Tiki Berutu, A.Md	SMK Teladan Tembung	2009	Kantor Camat Pakpak Barat	2.98	Sesuai
26	Tugas Fernando, A.Md	SMA N 1 Binjai	2005	Departeman Agama Sumatera Utara	3.22	Sesuai
27	Veritawati Munthe, A.Md	SMA N 1 Kabanjahe	2007	Pemkab Kabanjahe	3.00	Sesuai
28	Wenni Indriani, A.Md	SMK Eka Prasetya Medan	2010	Dinas Perindustiran Dan Perdagangan Asahan	3.00	Sesuai
29	Wintar Limbong, A.Md	SMU N-1 Sianjur Mula-Mula	2005	Departeman Hukum Dan Ham Republik Indonesia	2.98	T. Sesuai
30	Yoyon Haryono, A.Md	SMKN-2 Medan	2010	BPPT-SU	3.22	T. Sesuai
31	Sumardi Tumanggor, A.Md	SMA Raksana Medan	2007	PT. PLN Sidikalang Ophar Gi	3.00	Sesuai
32	Eka Chandra Purba, A.Md	SMAN1 Palu	2008	Kementrian Hukum Dan Ham	2.78	T.Sesuai

33	Ernodi, A.Md	SMKN-2 Balikpapan	2008	Dinas Pendidikan Pemkot Medan	3.00	Sesuai
34	Roy Sium Panjaitan, A.Md	SMAN 3 Pematang Siantar	2009	Pemkot Pematang Siantar	2.98	Sesuai
35	Sahat, A.Md	SMAN-5 Tebing	2008	Dinas Pendapatan Batubara	3.22	Sesuai
36	Saurma Siregar, A.Md	SMAN-1 Laguboti	2008	Kementerian Kehutanan	3.22	Sesuai
37	Veritawati Munthe, A.Md	SMAN1 Kabanjahe	2007	Pemkab Kabanjahe	3.00	Sesuai
38	Verry Sihombing, A.Md	SMA Sudiro Husodo	2009	Departemen Perhubungan Darat Deli Serdang	2.78	sesuai
39	Rensus Simanjuntak, A.Md	SMK Putra Anda Binjai	2006	Dinas Pendidikan Smpn 4 Tebing Tinggi	3.00	T. Sesuai
40	Berton Tarigan, A.Md	SMA Katolik Kabanjahe	2009	Kantor Camat Pemerintah Kabupaten Tanah Karo	2.98	T. Sesuai
41	Dippos Silitnga, A.Md	SMAN1 Lubuk Pakam	2005	Dinas Pengerjaan Umum Kab. Deli Serdang	3.22	Sesuai
42	Evi Sembiring, A.Md	SMA Katolik-1 Kabanjahe	2008	Pemerintah Kabupaten Karo	3.00	T. Sesuai
43	Patar Julianto, A.Md	SMU HKBP Pematang Siantar	2006	Dinas Pendidikan Kota Pematang Siantar	3.00	Sesuai
44	Azis Simbolon, A.Md	SMAN1 Pangururan	2006	Departemen Komunikasi Dan Informatika Pemprrovsu	2.98	Sesuai
45	Bastian Simatupang, A.Md	SMAN-1 Tarutung	2009	Dinas Tata Ruang Dan Tata Bangunan Pemko Medan	3.22	Sesuai
46	Berta Oktavia Silalahi, A.Md	SMUN-1 Tarutung	2011	Dinas Pendapatan Pengelolaan Keuangan	3.00	Sesuai
47	Casfarof Tampubolon, A.Md	SMUN-1 Balige	2006	Kantor Kesatuan Bangsadan Politik Labuhan Batu	3.22	Sesuai
48	Chanra Leo Saragih, A.Md	SMK Nusantara Dlksanggul	2010	Dinas Perkebunan Provinsi Sumatera Utara	3.00	Sesuai
49	Dayanti Malau, A.Md	SMAN1 Pangururan	2008	Kantor Bupati Samosir	2.78	T. Sesuai
50	Debora Hutabarat, A.Md	SMAN-2 Sidikalang	2006	Dinas Pendidikan Sidikalang	3.00	T. Sesuai
51	Dina Rosti Simarmata, A.Md	SMAN1 Pangururan	2008	Dinas Pendidikan Kabupaten Samosir	2.98	Sesuai
52	Duma Marpaung, A.Md	SMAN-1 Doloksanggul	2008	Dinas Kehutanan Provsu	3.22	T. Sesuai
53	Gita Pardede, A.Md	SMKN-2 Balige	2006	Dinas Pertanahan Aek Kanopan-Labuhan Batu	3.00	Sesuai
54	Hetty, A.Md	SMAN 1 Tarutung	2009	Dinas Pertanahan Pakpak Bharat	3.00	Sesuai
55	Kristina Lumban Tobing, A.Md	SMAN1 Pangururan	2008	Dinas Pendidikan Samosir	2.98	Sesuai
56	Lamria Simatupang, A.Md	SMKN-2 Balige	2009	Dinas Kominfo Pemprov DKI Jakarta	3.22	Sesuai
57	Lince Sianturi, A.Md	SMA ST. Petrus Sidikalang	2008	Satuan Kesdam I BB	3.00	Sesuai

58	Masrina Malau, A.Md	SMAN 2 Pangururan	2009	Kantor Bupati Samosir	3.22	Sesuai
59	Masrina Pandiangan, A.Md	SMAN 2 Pangururan	2006	Dinas Pertanian Samosir	3.00	T. Sesuai
60	Modesta Silalahi, A.Md	SMAN-1 Sidikalang	2009	Kantor Pemko Sidikalang	2.78	T. Sesuai
61	Murni Bowman, A.Md	SMAN 1 Salak	2009	Pemkab Pakpak Bharat	3.00	Sesuai
62	Riamsauli Sagala, A.Md	SMAN1 Sianjurmula-Mula	2009	Dinas Pendidikan Samosir	2.98	T. Sesuai
63	Rikardo Simbolon, A.Md	SMA N 1 ST. MICHAEL Pangururan	2006	Dinas Kependudukan Pemkab Samosir	3.22	Sesuai
64	Silviza, A.Md	SMA Al Azhar Medan	2007	Dinas Bkd Pemprov SU	3.00	Sesuai
65	Parlin Simatupang, A.Md	SMAN1 Pangaribuan	2010	PT. Gas Negara Republik Indonesia	3.00	Sesuai

Analisis *Subspace Clustering*

Hasil pengelompokan pada satu strategi sehingga di atas akan diproses oleh pertambangan aturan asosiasi, untuk menemukan aturan asosiatif antara kombinasi tertentu. Sebuah jenis tertentu penalaran yang menggunakan "if-then-else" pernyataan aturan akan dilaksanakan, hanya sebagai pola dan pencarian mesin inferensi untuk pola dalam aturan yang sesuai pola dalam data. Penelitian ini akan menerapkan subspace klaster, sebagai operator belajar dari kedua data nominal dan numerik.

Subspace Cluster Berbasiskan Density Connection.

Saat aturan telah ditemukan, penelitian ini akan menggunakan koneksi berbasis density untuk mengidentifikasi dan mencari peluang kerja. Hasil ini digunakan sebagai pengetahuan Model basis untuk mewakili objek dalam informasi. Model menangkap hubungan antara banyak faktor untuk memungkinkan penilaian potensi terkait dengan set kondisi tertentu.

Pengelompokan satu set data dengan subspace clustering yang menghadirkan tantangan khusus karena dua titik data milik cluster tampilan yang sama seperti yang berbeda sebagai sepasang sewenang-wenang titik data. Gagasan utama subspace pengelompokan berdasarkan kepadatan di setiap cluster adalah bahwa data memiliki jumlah minimum data tetangga, di mana kerapatan data harus lebih dari batas tertentu.

Usulan proses klasterisasi subclue melibatkan lima tahap:

a. Format data 1 dimensi menggunakan DBSCAN misalnya, ada data matriks sebagai berikut:

Table 3. Example of Initial Data

data1	data2	data3
0	0	120
0	0	98
0	275	150
100	150	100
200	100	125

Dengan menggunakan fungsi berikut:

```
public function format1D()
{
$this->_headerData=array('x','y','mark','cid','rid','kom');
```

```
$dataPoint =array();
$dmAll = new DataOperation;
#print_r($this->_data);
foreach($this->_header as $k)
{
   foreach($this->_header as $j)
   {
      if($k==$j) break;
         $dm = new DataOperation;
         $dm->attribute(array("x","y"));
         $i=0;
         foreach($this->_data->data as $e)
               {
               $data=array($e->{$j},$e->{$k},0,0,$i,($j.'-U-'.$k));
               $dm->addData(array_combine($this->_headerData, $data));
               $i++;
               }
              $dmAll->addData(array($j.'_u_'.$k=>$dm));
         }
   }
$this->_data1D= $dmAll;
}
```

Setelah di data di format maka akan menghasilkan tabel matriks 1-dimensi sebagai berikut:

Tabel 4. Hasil terpisah multidimensi menjadi 1-dimensi

Dimension-1		Dimension-2		Dimension-3	
data1	data2	data1	data3	data2	data3
0	0	0	120	0	120
0	0	0	98	0	98
0	275	0	150	275	150
100	150	100	100	150	100
200	100	200	125	100	125
250	200	250	122	200	122

b. Menghasilkan cluster berdasarkan koneksi kerapatan Clustering menggunakan DBSCAN untuk setiap 1-Dimensi, dan memiliki hasil sebagai berikut:

Tabel 5. Hasil pengelompokan berdasarkan DBSCAN

0	1	0	0	75625	0	32500	0	50000	0	102500	0
0	1	0	0	75625	0	32500	0	50000	0	102500	0
75625	0	75625	0	0	3	25625	0	70625	0	68125	0
32500	0	32500	0	25625	0	0	4	12500	0	25000	0
50000	0	50000	0	70625	0	12500	0	0	5	12500	0
102500	0	102500	0	68125	0	25000	0	12500	0	0	6

c. Menghasilkan subspace klaster Kemudian fungsi berikut akan penciptaan subruang clusters Test calon dan menghasilkan klaster dimensi *$subspace[1]->addData(array('subspace'=>$d->perRegion()));* Yang akan dihasilkan pengelompokan subspace cluster sebagai berikut:

Tabel 6. Nilai yang dihasilkan oleh Subspace Cluster

S-1	s-1	C-0
		C-1
	s-2	C-0
		C-1
	s-3	C-0
		C-1

Setelah itu setiap cluster dikelompokkan bersama membentuk klaster yang terpisah, sehingga hasilnya adalah sebagai berikut.

Table 7. Result of Group of Subspace Cluster

CandSk+1	S1 + S2
	S1 + S3
	S2 + S3

d. Uji kandidat dan menghasilkan klaster dimensi Langkah selanjutnya adalah uji dan menghasilkan klaster dimensi dengan fungsi sebagai berikut:

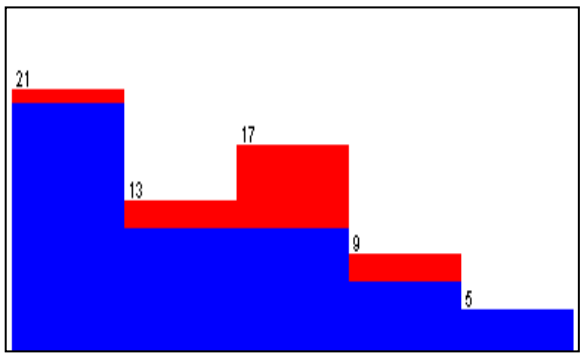
```
for($i=0;$i<count($data);$i++)
{
$s1=$data[$i];
#print_r($s1);
#echo "<hr />";
for($j=$i+1;$j<count($data);$j++)
        {
        $s2 = $data[$j];
                $candidate = array($s1->subspace,$s2->subspace);
                $candidates[] = $candidate;
                $x++;
                #print_r($candidate);
                break;
        }
}
```

Hasil Penelitian

Untuk memverifikasi kualitas clustering diperoleh melalui teknik kami (Damira) dan untuk mempercepat tahap pertama, kita jalankan DBSCAN, FIRES, INSCY, SUBCLU, dan Damira. Parameter Pengaturan dilakukan pada subspace klaster bracketing, dan dimensi rata-rata dan jumlah klaster didefinisikan. Tabel 8 menunjukkan milik dataset, kami diimplementasikan dalam 3 dataset nyata, dan dataset pendidikan tinggi 6. Untuk setiap penelitian uji menggunakan MinPoints = 6 dan Epsilon = 0,9, berdasarkan percobaan sebelumnya kriteria ini membuat cluster terbaik.

Tabel 8. Properti dari dataset

Dataset	Attributes	No of data	M	E
Glass	10	214	6	0.9
Liver-dis	7	345	6	0.9
Job satisfaction	8	288	6	0.9
Mahasiswa medicom tahun 2005	44	100	6	0.9
Mahasiswa medicom tahun 2006	67	100	6	0.9
Mahasiswa medicom tahun 2007	62	100	6	0.9
Mahasiswa medicom tahun 2008	8	71	6	0.9
Mahasiswa medicom tahun 2009	60	71	6	0.9
Mahasiswa medicom tahun 2010	61	71	6	0.9



Gambar 14. Distribusi Data Dataset

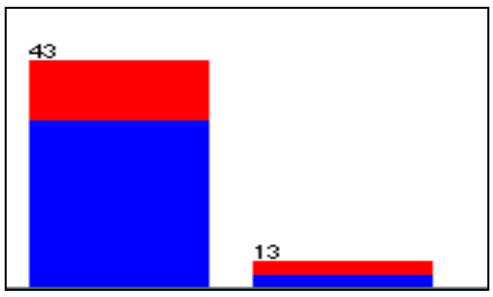
Berdasarkan grafik yang dihasilkan, diperoleh bahwa data diklasifikasikan dalam 5 kelompok, yaitu :

a. Kelompok tahun 2002–2003 : diketahui bahwa data alumni ada 21 orang. Jumlah yang sesuai bidang pekerjaan diberi warna biru, dan yang tidak sesuai diberi warna merah.
b. Kelompok tahun 2004 : diketahui terdapat 13 data.
c. Kelompok tahun 2005 : terdapat 17 data
d. Kelompok tahun 2006 : terdapat 9 data.
e. Kelompok tahun 2007 – 2010 : terdapat 5 data.

Untuk data berdasarkan tahun tamat SMA/SMK diketahui bahwa :

a. Nilai maximun : 2008
b. Nilai Minimum : 2002
c. Mean : 2004, 304
d. Standard Deviasi : 1,53

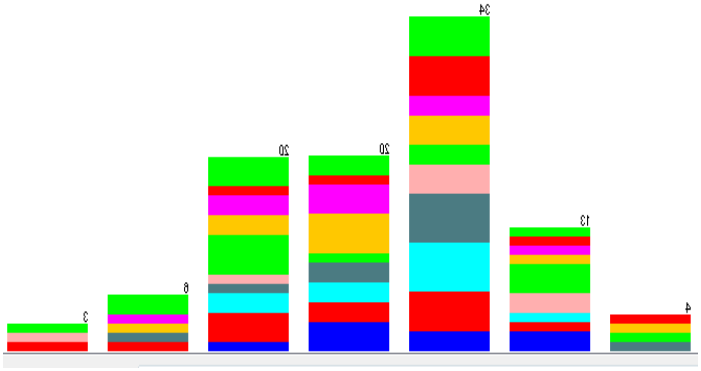
Berdasarkan tahun tamat Medicom :



Gambar 15. Distribusi Data dataset kerja yang memenuhi sesuai bidang kompetensi

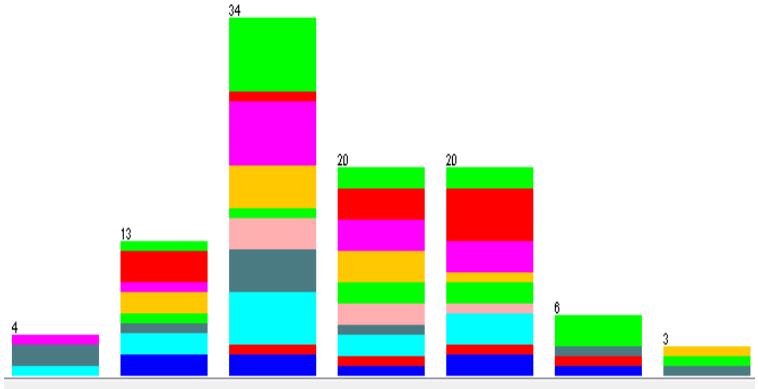
Berdasarkan grafik yang diperoleh maka diketahui :

a. Data alumni yang tamat Medicom tahun 2010 : 43 orang
b. Data alumni yang tamat Medicom tahun 2009 : 13 orang



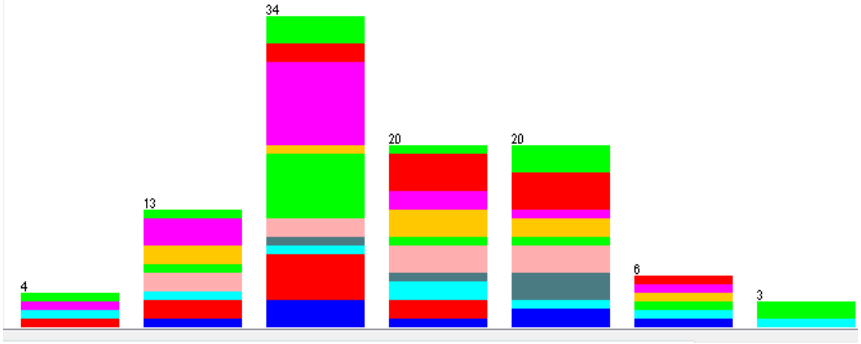
Gambar 16. Distribusi Data Mahasiswa Medicom tahun 2005

Berdasarkan grafik yang dihasilkan, diperoleh bahwa data diklasifikasikan dalam 7 kelompok, yaitu Kelompok tahun 2005 diketahui bahwa data alumni ada 34 orang Jumlah yang sesuai bidang pekerjaan.



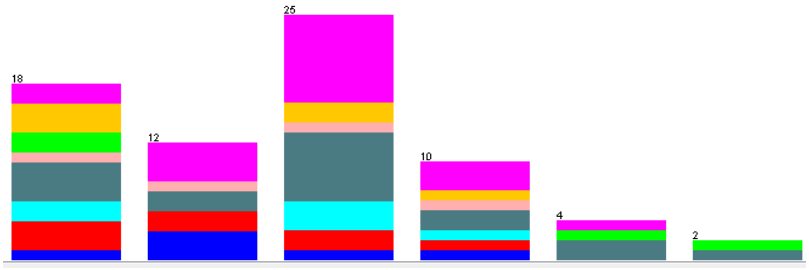
Gambar 17. Distribusi Data Mahasiswa Medicom tahun 2006

Berdasarkan grafik yang dihasilkan, diperoleh bahwa data diklasifikasikan dalam kelompok, yaitu Kelompok tahun 2006 diketahui bahwa data alumni ada 34 orang Jumlah yang sesuai bidang pekerjaan.



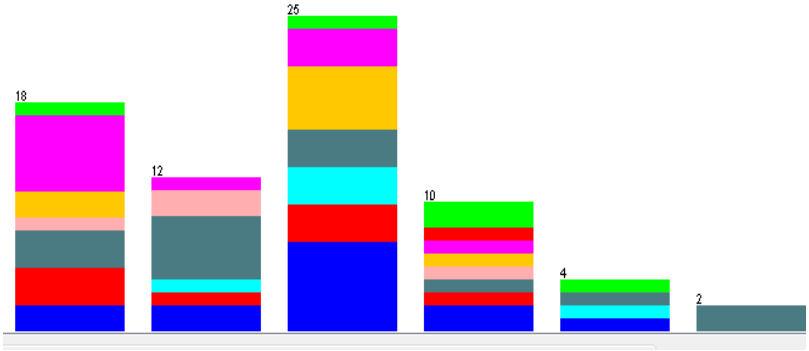
Gambar 18. Distribusi Data Mahasiswa Medicom tahun 2007

Berdasarkan grafik yang dihasilkan, diperoleh bahwa data diklasifikasikan dalam 5 -kelompok, yaitu Kelompok tahun 2007 diketahui bahwa data alumni ada 34 orang. Jumlah yang sesuai bidang pekerjaan



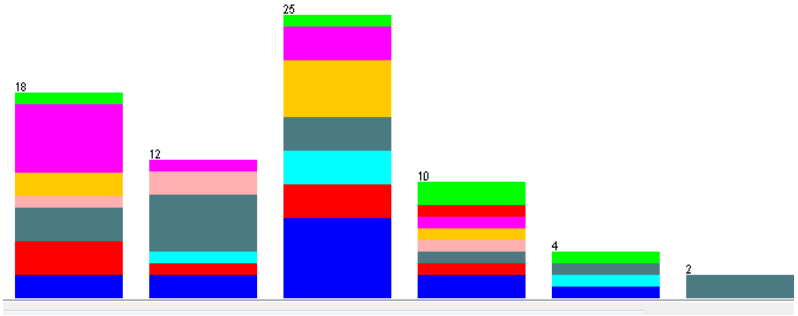
Gambar 19. Distribusi Data Mahasiswa Medicom tahun 2008

Berdasarkan grafik yang dihasilkan, diperoleh bahwa data diklasifikasikan dalam 5 kelompok, yaitu Kelompok tahun 2008 diketahui bahwa data alumni ada 25 orang Jumlah yang sesuai bidang pekerjaan.



Gambar 20. Distribusi Data Mahasiswa Medicom tahun 2009

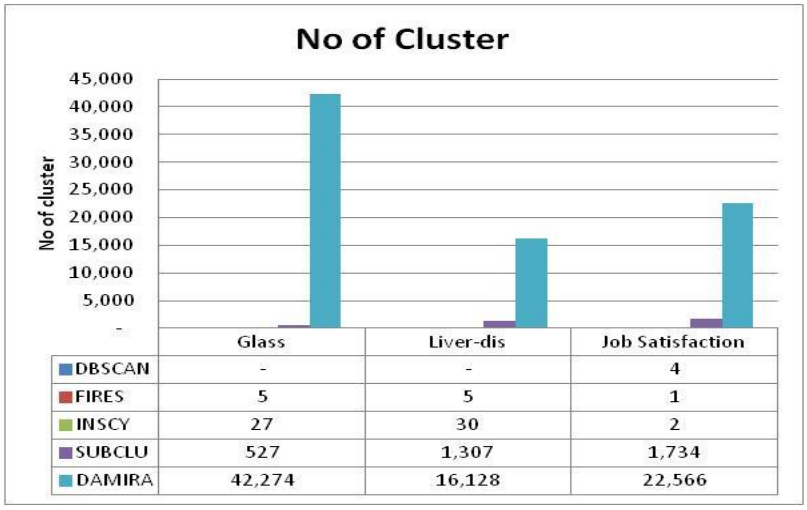
Berdasarkan grafik yang dihasilkan, diperoleh bahwa data diklasifikasikan dalam 5 kelompok, yaitu Kelompok tahun 2007 diketahui bahwa data alumni ada 25 orang.



Gambar 21. Distribusi Data Mahasiswa Medicom tahun 2010

Berdasarkan grafik yang dihasilkan, diperoleh bahwa data diklasifikasikan dalam 5 kelompok, yaitu Kelompok tahun 2010 diketahui bahwa data alumni ada 25 orang. Jumlah yang sesuai bidang pekerjaan.

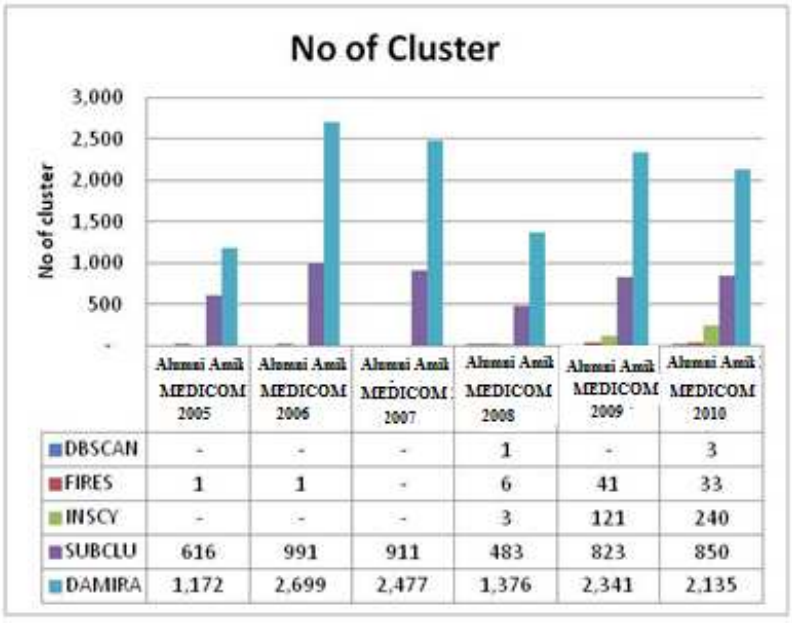
Sebuah aspek penting dari metode yang diusulkan adalah jumlah cluster yang dihasilkan. Berdasarkan hasil tes, DBSCAN cluster yang dihasilkan hanya 4 cluster diperoleh pada dataset alumni yang sesuai kerja dengan bidangnya juga, jumlah cluster diidentifikasi sangat rendah, mulai dari 5 cluster. Sementara itu Damira Damira berhasil membangun sangat besar jumlah cluster untuk setiap dataset.



	Glass	Liver-dis	Job Satisfaction
DBSCAN	-	-	4
FIRES	5	5	1
INSCY	27	30	2
SUBCLU	527	1,307	1,734
DAMIRA	42,274	16,128	22,566

Gambar 22. Jumlah cluster dataset secara realtime

Demikian pula, jumlah kelompok dataset institusi pendidikan tinggi, membentuk lain metode subspace klaster, seperti fire cluster dan INSCY, kecenderungan untuk gagal untuk membentuk kelompok di setiap ruang bagian, seperti yang ditunjukkan pada gambar 23.

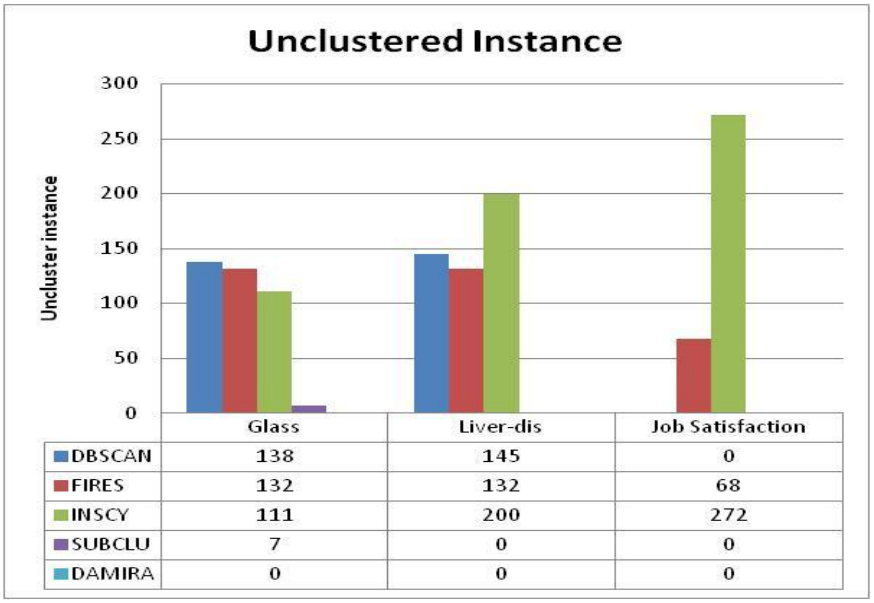


	Alumni Amik MEDICOM 2005	Alumni Amik MEDICOM 2006	Alumni Amik MEDICOM 2007	Alumni Amik MEDICOM 2008	Alumni Amik MEDICOM 2009	Alumni Amik MEDICOM 2010
DBSCAN	-	-	-	1	-	3
FIRES	1	1	-	6	41	33
INSCY	-	-	-	3	121	240
SUBCLU	616	991	911	483	823	850
DAMIRA	1,172	2,699	2,477	1,376	2,341	2,135

Gambar 23. Jumlah cluster Tertinggi dataset menurut alumni

Kekurangan lain dari pengelompokan adalah berapa banyak data yang hilang atau un-cluster, Seperti ditunjukkan dalam Gambar 24, SUBCLU dan Damira tidak memiliki un-cluster dataset nyata, sehingga persepsi hasil cluster akan menghasilkan informasi yang lebih akurat.

Jurnal Teknovasi
Volume 02, Nomor 01, 2015, 33 – 60
ISSN : 2355-701X



	Glass	Liver-dis	Job Satisfaction
DBSCAN	138	145	0
FIRES	132	132	68
INSCY	111	200	272
SUBCLU	7	0	0
DAMIRA	0	0	0

Gambar 24. Data A-cluster dataset nyata

Demikian pula, proses pengelompokan data lembaga pendidikan tinggi, oleh FIRES dan INSCY tidak semua data dapat di cluster. Seperti ditunjukkan dalam Gambar 25, metode INSCY paling mungkin untuk menghasilkan contoh un-cluster, sementara SUBCLU berhasil semua data dalam cluster.

	Alumni Amik MEDICOM 2005	Alumni Amik MEDICOM 2006	Alumni Amik MEDICOM 2007	Alumni Amik MEDICOM 2008	Alumni Amik MEDICOM 2009	Alumni Amik MEDICOM 2010
DBSCAN	101	101	101	1	71	44
FIRES	60	80	101	8	0	9
INSCY	101	101	101	52	12	24
SUBCLU	0	0	0	0	0	0

Gambar 25. Data Un-cluster yang lebih tinggi dataset Perguruan tinggi

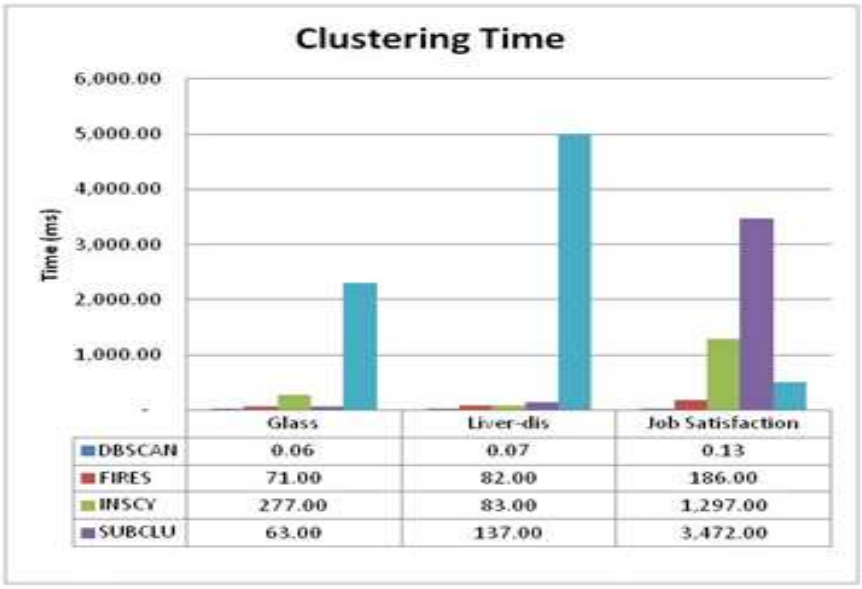
EVALUASI KINERJA

Evaluasi kinerja data mining menjadi sangat penting, prediksi nomor yang benar cluster tanpa pengawasan proses pembelajaran adalah rintangan, namun dapat dibersihkan dengan menggunakan efisiensi, akurasi, cakupan klaster dan indeks F1-Entropi untuk menilai kualitas dari cluster.

Jurnal Teknovasi
Volume 02, Nomor 01, 2015, 33 – 60
ISSN : 2355-701X

Efisiensi Kerja

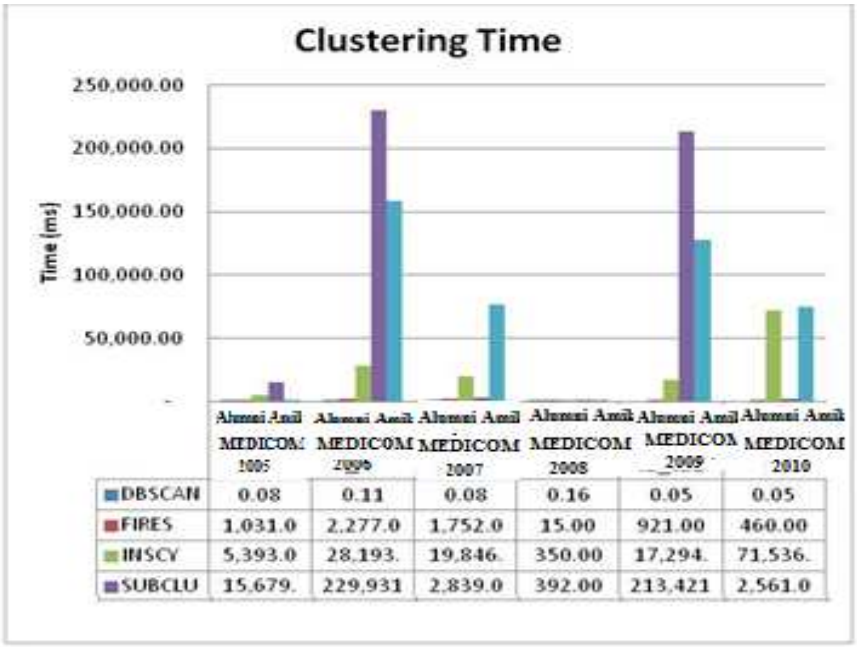
Dari hasil percobaan, kita dapat melihat bahwa waktu untuk pengelompokan dataset kaca dan dataset hati, sedangkan untuk kepuasan kerja dataset Damira membutuhkan waktu singkat daripada metode SUBCLU dan INSCY sebagai ditunjukkan pada gambar 26



	Glass	Liver-dis	Job Satisfaction
DBSCAN	0.06	0.07	0.13
FIRES	71.00	82.00	186.00
INSCY	277.00	83.00	1,297.00
SUBCLU	63.00	137.00	3,472.00

Gambar 26. Waktu proses pengelompokan dataset nyata

Untuk lebih besar dan lebih kompleks data, kinerja Damira terlihat lebih efisien daripada SUBCLU seperti yang ditunjukkan pada gambar 27. Namun, masih lebih rendah dari FIRES dan INSCY, terutama dibandingkan dengan metode DBSCAN, yang dapat dilakukan dengan sangat cepat, rata-rata kurang dari 1 detik, tetapi cenderung banyak data un-cluster

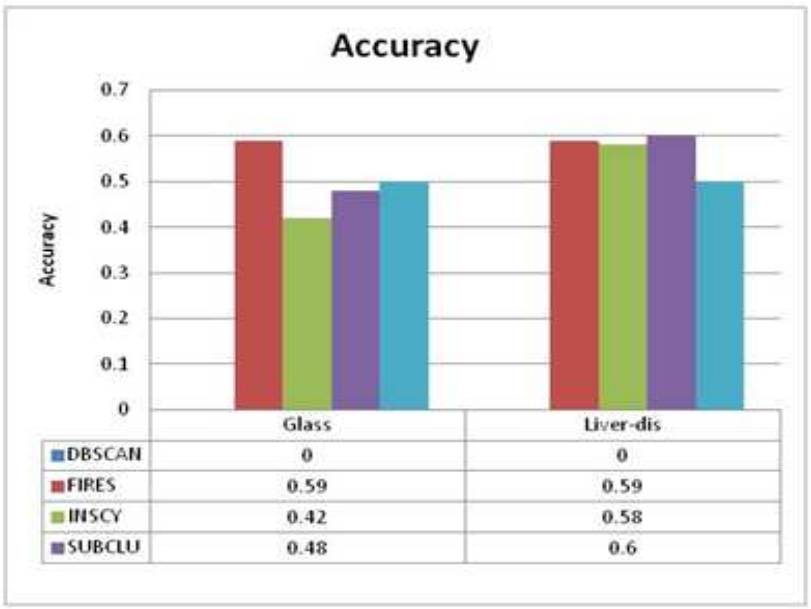


	Alumni Anil MEDICOM 2005	Alumni Anil MEDICOM 2006	Alumni Anil MEDICOM 2007	Alumni Anil MEDICOM 2008	Alumni Anil MEDICOM 2009	Alumni Anil MEDICOM 2010
DBSCAN	0.08	0.11	0.08	0.16	0.05	0.05
FIRES	1,031.0	2,277.0	1,752.0	15.00	921.00	460.00
INSCY	5,393.0	28,193.	19,846.	350.00	17,294.	71,536.
SUBCLU	15,679.	229,931	2,839.0	392.00	213,421	2,561.0

Gambar 27. Waktu proses dari pengelompokan yang tertinggi dari dataset lembaga pendidikan

Data Akurat

Selain mengevaluasi efisiensi kerja metode pengelompokan subspace, penelitian ini membahas parameter terkait hasil clustering. Hasil percobaan menunjukkan bahwa keakuratan metode INSCY lebih akurat dibandingkan SUBCLU dan FIRES, seperti yang ditunjukkan pada Gambar 27. Dalam percobaan pengelompokan untuk dataset kaca, metode FIRES memiliki akurasi yang lebih baik daripada INSCY, dan SUBCLU. Sementara metode Damira itu lebih akurat daripada metode INSCY dan SUBCLU, seperti yang ditunjukkan pada Gambar 28. Tapi untuk percobaan dari dataset pengelompokan akurasi Damira lebih rendah dibandingkan metode FIRES, INSCY dan SUBCLU.



Gambar 28. Hasil akurasi dataset

Hasil Akurasi

Dalam percobaan pengelompokan untuk dataset kaca, metode FIRES memiliki akurasi yang lebih baik daripada INSCY, dan SUBCLU. Sementara metode Damira itu lebih akurat daripada metode INSCY dan SUBCLU, seperti yang ditunjukkan pada Gambar 21. Tapi untuk percobaan dari dataset pengelompokan akurasi Damira lebih rendah dibandingkan metode FIRES, INSCY dan SUBCLU.

KESIMPULAN

Penggunaan algoritma clustering adalah untuk mengukur kesamaan data dimensi tinggi atau atribut sering mencapai hasil yang diinginkan, menyebabkan atribut menjadi tidak berhubungan atau terlalu dekat bersama-sama. Data tertutup dapat membentuk kelompok tumpang tindih baik membentuk kelompok padat, data dapat ditemukan dalam kelompok yang berbeda dan juga di ruang bagian yang berbeda. Subspace clustering diproyeksikan sebagai teknik mencari data atau pengelompokan atribut dalam cluster yang berbeda. Pengelompokan dilakukan dengan menentukan tingkat kepadatan data dan juga dilakukan untuk mengidentifikasi outlier atau data yang tidak relevan yang akan membuat setiap cluster untuk eksis dalam subset terpisah.

Penelitian ini membuktikan inovatif pada algoritma clustering subspace Clustering (SUBCLU) berdasarkan koneksi kepadatan. Kehadiran penelitian memperkirakan dimensi kepadatan dan hasilnya digunakan sebagai data masukan untuk menentukan cluster awal berdasarkan koneksi kepadatan, menggunakan algoritma DBSCAN. Setiap dimensi diuji untuk

menyelidiki apakah memiliki hubungan dengan data pada gugus lain, dengan menggunakan algoritma klasterisasi subruang diusulkan. Jika data memiliki hubungan, maka akan diklasifikasikan sebagai subruang. Penelitian ini menggunakan data multidimensi, seperti dataset patokan dan dataset nyata. Dataset Real dari pendidikan, khususnya mengenai persepsi pelatihan industri siswa dan dari industri.

Untuk memverifikasi kualitas clustering diperoleh melalui teknik penulis dan untuk mempercepat tahap pertama, kita jalankan DBSCAN, INSCY, SUBCLU. kecenderungan untuk gagal untuk membentuk kelompok di setiap ruang bagian. SUBCLU dan Damira tidak memiliki un-cluster dataset nyata, sehingga persepsi hasil cluster akan menghasilkan informasi yang lebih akurat., sedangkan untuk kepuasan kerja dataset DBSCAN membutuhkan waktu lebih lama daripada metode SUBCLU. Untuk lebih besar dan lebih kompleks data, kinerja SUBCLU terlihat lebih efisien daripada DBSCAN.

DAFTAR PUSTAKA

Abonyi, János, Balázs Feil Cluster Analysis forData Mining and System Identification, 2007, Birkhauser Verlag AG, Berlin

Agarwal, C. Procopiuc, J.L. Wolf, and P.S. Yu, Fast Algorithms for Projected Clustering, *Proc. ACM SIGMOD*, 1999

Agrawal, Rakesh, Johannes Gehrke, Automatic Subspace Clustering of High Dimensional Data, *Data Mining and Knowledge Discovery*, 2005, pp.5-33

Baker, Ryan S.J.d., The State of Educational Data Mining in 2009: AReview and Future VisionsData Mining for Education, 2009, *Journal of Educational Data Mining,* October 2009, Volume 1, Issue 1

Berka,Petr, Jan Rauch, Djamel Abdelkader Zighed, *Data Mining And Medical Knowledge Management, Medical Information science reference*, New York, 2009

Berry, W. Michael, Malu Castellanos, *Survey of Text Clustering*, 2008, Springer-Verlag London Limited

Bicici, Ergun, Deniz Yuret, *Local Scaled Density Based Klastering,* ICANNGA, 2007, pp.739-748

Bin, Deng, Shao Peiji, Zhao Dan, Data Mining for Needy Students Identify Based on Improved RFM Model: A Case Study of University, *International Conference on Information Management, Innovation Management and Industrial Engineering* (ICIII), 2008, vol. 1, pp.244-247

Callahan, Dale, Bob Predigo, *Educating Experienced IT Professionals by Addressing Industry's Needs,* 2007

Chakrabarti, Kaushik, Sharad Mehrotra, Local Dimensionality Reduction : A New Approach To Indexing High Dimensional Space, *Proceeding Of The 26th VLDB Conference, Cairo, Egypt,* pp.89-100 (2000)

Cristianini, Nello, John Shawe-Taylor, *An Introduction to Support Vector Machines: And Other Kernel-Based Learning, Cambridge University Press, Cambridge*, UK, 2000

Chu, Yi-Hong, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang, Ming-Syan Chen, Density Conscious Subspace Clusteringfor High-Dimensional Data, 2010, *IEEE Transactions On Knowledge And Data Engineering*, Vol. 22, No. 1, January 2010

Chua, Freddy Chong Tat, (2009), *Dimensionality Reduction and Clustering of Text Document, .*

Cordeiro, Robson, L.F, Agma J.M. Traina, Christos Faloutsos, Caetano Traina Jr., "Finding Clusters in Subspaces of Very Large Multi-dimensional Datasets", 26th *International Conference on Data Engineering (ICDE)*, p.625-636, 2010,

Cortes, Corinna, Vladimir Vapnik, *Support-Vector Networks, Machine Learning*, 20, 273-297 (1995)

Cunningham, Pádraig, (2007), Dimension Reduction, *Technical Report UCD*-CSI-2007-7

Dayan, Peter, *Unsupervised Learning*, The MIT Encyclopedia of the Cognitive Sciences, p.1-7

Devaraj, Sarv, S. Ramesh Babu, How To Measure The Relationship Between Training and Job Performance, *Communications of the ACM*, Volume 47, Issue 5, p.62-67, 2004

Ding, Chris, Tao Li, (2007), Adaptive Dimension Reduction Using Discriminant Analysis and K-means Clustering, *International Conference on Machine Learning, Corvallis,* OR, 2007

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, A Density-Based Algorithm for Discovering Clusters, *2nd International Conference on Knowledge Discovery and Data Mining* (KDD-96), 1996

Farquad, *Machine LearningApplications and Trends:Algorithms*, Methods, andTechniques, University of Valencia, Spain

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth, AI Magazine Volume 17 Number 3, *American Association for Artificial Intelligence* (1996)

Finley, Thomas, Thorsten Joachims, Supervised Clustering with Support Vector Machines, *International Conference on Machine Learning*, Bonn, Germany, 2005. p.1-8

Fodor, I.K: A Survey of Dimension Reduction Techniques. LLNL Technical Report, *UCRL-ID*-148494, p.1-18 (2002)

Fung, Glenn M, Multicategory Proximal Support VectorMachine Classifiers, *Machine Learning,* 59, 77–97, 2005

Gan, Guojun, Jianhong Wu, Zijiang Yang, PARTCAT : A Subspace Clustering Algorithm for High Dimensional Categorical Data, *International Joint Conference on In Neural Networks* (IJCNN), p.4406-4412, 2006

Grossman, R.L. et. al., editors, Data Mining: Challenges and Opportunities, *http://www.ncdm.uic.edu*

Gunnemann, Stephan, Hardy Kremer, Thomas Seidl, Subspace Clustering for Uncertain Data, *SIAM International Conference on Data Mining*, p.245-260, 2009

Han, Jiawei, Micheline Kamber, *Data Mining: Concepts and Techniques*, 2nd Edition, 2006, p.25-26, Morgan Kaufmann

Hand, David, Heikki Mannila, Padhraic Smyth*, Principles of Data Mining*, Massachisetts Institute of Technology Press, Cambridge, Massachusetts, London, England

Heuchan, Archibald, J.F, Industry and Higher Education—Meeting the Needs of the Mining*, Engineering Sector* 105th Annual General Meeting ofthe Canadian Mining, *Metallurgy and Petroleum*, Montreal, Quebec, 2003

Houle, Michael E.; Kriegel, Hans-Peter; Kröger, Peer; Schubert, Erich; Zimek, Arthur (2010), Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?, *Proceedings of the 21th International Conference on Scientific and Statistical Database Management* (SSDBM) (Heidelberg, Germany: Springer)

Maimon, Oded, Lior Rokach, Decomposition Methodology For Knowledge Discovery And Data Mining, World Scientific, 2005.