CATEGORIC DATA GROUPING BY ALGORITHM QUICK ROBUST CLUSTERING USING LINKS (QROCK) (Case Study: Status of Value Addrd Tax Payments at the Samarinda Ulu Primary Tax Office in 2018)

Nana Nirwana[1], Memi Nor Hayati[2], Syaripuddin[3]
[1] Laboratorium Statistika Terapan FMIPA Universitas Mulawarman
[2]Laboratorium Matematika Komputasi FMIPA Universitas Mulawarman
Alamat e-mail : [1] nnwana92@gmail.com, [2]meminorhayati@fmipa.unmul.ac.id

ABSTRACT

Clustering is a method for finding and grouping data that have similar characteristics (similarity) between one data and another. The method of grouping used in this study is the Qrock Algorithm (Quick Robust Using Links).The Qrock Algorithm has a more efficient method to produce the final cluster when the Rock Algorithm has no link beetwen the clusters.The concept of the Qrock Algorithm basically has the same principles as the Rock Algorithm, except that the Qrock Algorithm classifies objects only based on the neighbors of each object. The purpose of this study was to classify 200 Value Added Tax Payment Status data at the Samarinda Ulu Tax Service Office in 2018. Based on the analysis results, the threshold value (θ) = 0.1; 0.2; 0.3; 0.4; 0, 5 and 0.6 produce 1 cluster while the threshold values (θ) = 0.7; 0.8 and 0.9 produce 56 clusters.

Keywords: Qrock Algorithm, Clustering, Categorical Data.

PENDAHULUAN

Data mining adalah suatu istilah yang digunakan untuk penguraian penemuan pengetahuan di dalam *database*. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengindentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar [11].

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu deskripsi, estimasi, prediksi, klasifikasi, asosiasi, dan *clustering. Clustering* merupakan pengelompokkan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan [5].

Clustering atau Pengklasteran adalah proses *data mining* untuk melihat pola pendistribusian data yang akan digunakan untuk melihat karakteristik dari data. Dalam *data mining* ada dua jenis metode *clustering* yang digunakan dalam pengelompokkan data, yaitu pengelompokkan hierarki (*hierarchical clustering)* dan pengelompokkan non hierarki (*non hierarchical clustering)* [8].

Metode pengelompokkan hierarki terbagi lagi menjadi 2, yaitu *agglomerative* (penggabungan) dan *divisif* (pemisahan). Metode *agglomerative* adalah metode yang dimulai dengan kenyataan bahwa setiap objek membentuk kelompoknya masing-masing kemudian dua objek dengan jarak terdekat bergabung. Selanjutnya objek ketiga akan bergabung dengan kelompok yang ada atau bersama objek yang lain membentuk kelompok baru. Metode *agglomerative* terdiri dari *single linkage, complete linkage* dan *average linkage* [6].

Metode *agglomerative* dapat digunakan untuk mengelompokkan data yang bersifat kategorik. Beberapa metode *agglomerative* yang mengelompokkan data yang bersifat kategorik antara lain : metode *Rock (Robust hierarchical-clustering using*

Links) dan metode *Qrock (Quick Rock)*. Algoritma *Rock* adalah teknik pengelompokkan data secara hierarki *agglomerative* berdasarkan ada tidaknya *links* diantara pasangan objek data dan proses pengelompokkan akan berhenti jika tidak ada lagi *links* antar objek data dalam kelompok berbeda yang didapatkan [2].

Algoritma *Qrock* digunakan untuk data dengan variabel berjenis kategorik. Menurut [1] dijelaskan bahwa variabel kategorik adalah variabel yang skala pengukurannya terdiri dari sekumpulan kategorik. Salah satu contoh data yang berjenis kategori yaitu pada bidang pajak.

Pajak merupakan peranan penting dalam pembiayaan pembangunan, dimana wajib pajak merupakan bagian dari penerimaan pajak tersebut. Dengan kata lain, tidak akan ada pajak apabila tidak ada wajib pajak. Wajib pajak adalah orang pribadi atau badan yang menurut ketentuan peraturan perundang-undangan perpajakan ditentukan untuk melakukan kewajiban perpajakan termasuk pemungutan atau pemotong pajak tertentu [4].

Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam *database*. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar [11].

Data mining termasuk dalam proses *Knowledge Discovery in Database* (KDD), yaitu kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data *mining* ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. Beberapa metode yang sering disebut-sebut dalam literatur data *mining* antara lain *clustering*, *classification*, *association rules mining*, *neural network, genetic algorithm* dan lain-lain [8].

Clustering

Clustering merupakan suatu metode untuk mencari dan mengelompokkan data yang memiliki kemiripan karakteristik (*similarity*) antara satu data dengan data yang lain. *Clustering* merupakan salah satu metode *data mining* yang bersifat tanpa arahan *unsupervised*), maksudnya metode ini diterapkan tanpa adanya latihan (*training*) dan tanpa ada guru (*teacher*) serta tidak memerlukan target *output*. *Data mining* memiliki dua jenis metode *clustering* yang digunakan dalam mengelompokkan data, yaitu *hierarchical clustering* dan *non-hierarchical clustering* [8].

Metode hierarki memulai pengelompokan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang mempunyai kedekatan dua tipe dasar yaitu *agglomerative* (pemusatan) dan metode *divisive*. Berbeda dengan metode hierarki, metode non-hierarki dimulai dengan terlebih dahulu menentukan jumlah *cluster* yang diinginkan. Setelah jumlah *cluster* diketahui, kemudian proses *cluster* dilakukan tanpa mengikuti proses hierarki [7].

Terkait dengan pengertian dan tujuan dilakukannya analisis *cluster*, dapat dinyatakan bahwa suatu kelompok yang baik adalah kelompok yang memiliki ciri-ciri sebagai berikut :

a. Homogenitas (kesamaan) yang tinggi antar anggota dalam satu kelompok (*within cluster*).
b. Heterogenitas (perbedaan) yang tinggi antar kelompok yang satu tahap dengan kelompok yang lain (*between cluster*).

Metode dalam Analisis *Cluster*

Terdapat dua metode yang dapat digunakan untuk melakukan analisis

cluster, kedua metode tersebut adalah metode hierarki dan non-hierarki. Metode pengelompokan hierarki (*hierarcichal clustering*) memulai pengelompokan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat, kemudian proses diteruskan ke objek lain yang mempunyai kedekatan kedua. Demikian seterusnya sehingga *cluster* akan membentuk semacam pohon dimana ada hierarki (tingkatan) yang jelas antar objek dari yang paling mirip hingga yang paling tidak mirip. Secara logika semua objek pada akhirnya hanya akan membentuk sebuah *cluster*. *Dendogram* biasanya digunakan untuk membantu memperjelas proses hierarki tersebut [9].

Strategi untuk pengelompokan hierarki pada umumnya dibagi menjadi dua jenis yaitu *agglomerative* (pemusatan) dan *divisive* (penyebaran). Pengelompokan hierarki *agglomerative* merupakan metode pengelompokan hierarki dengan pendekatan bawah-atas (*buttom-up*). Metode hierarki *agglomerative* (pemusatan) terdiri dari metode pautan (*linkage method*) dan metode *centroid* (*centroid method*). Metode pautan meliputi pautan tunggal (*single linkage*), pautan lengkap (*complete linkage*) dan pautan rata-rata (*average linkage*) [6].

Berbeda dengan metode hierarki, metode non-hierarki dimulai dengan menentukan terlebih dahulu jumlah *cluster* yang ditentukan (misalnya 2 *cluster*, 3 *cluster* dan seterusnya). Setelah jumlah *cluster* ditentukan baru proses pengelompokan dilakukan tanpa mengikuti proses hierarki [9].

Menetukan Banyak *Cluster*

Menurut [10], pokok utama dalam analisis *cluster* adalah menentukan berapa banyaknya *cluster*. Tidak ada aturan baku untuk menentukan banyaknya kelompok, namun ada beberapa petunjuk yang bisa digunakan:

1. Pertimbangan teoritis, konseptual, praktis dan bisa disarankan untuk menentukan banyaknya kelompok yang terbentuk.
2. Jarak kelompok yang digabung bisa digunakan sebagai kriteria penentuan banyak
3. kelompok. Informasi ini bisa diperoleh dari susunan dendogram. Cara ini biasanya digunakan dalam metode *hierarchical clustering.*
4. Rasio jumlah varian dalam kelompok dengan jumlah varian antar kelompok dapat dinyatakan sebagai banyaknya kelompok. Cara ini biasa digunakan dalam metode *non-hierarchical clustering.*
5. Banyaknya jumlah kelompok seharusnya berguna dan bermanfaat.

Algoritma *Quick Robust Clustering Using Links (QROCK)*

Analisis *cluster* pada data kategorik dilakukan menggunakan ukuran kemiripan atau jarak data kategorik, kemudian dilanjutkan menggunakan pengelompokkan dengan metode hierarki dan non hierarki, akan tetapi metode hierarki dan non hierarki dinilai tidak cocok digunakan pada data kategorik. Oleh karena itu telah dikembangkan beberapa metode untuk pengelompokkan data kategorik antara lain: Algoritma *Robust Clustering using link (Rock) dan Quick Robust Clustering using links(Qrock).* Algoritma *Qrock* diperkenalkan oleh [3], algoritma ini merupakan percepatan dari algoritma *Rock.* Algoritma *Qrock* sangat efisien untuk membentuk *cluster* sebagaimana *cluster* yang dibentuk oleh algoritma *Rock* ketika tidak ada *link* antara *cluster* satu dengan *cluster* yang lain. Beberapa parameter yang digunakan dalam algoritma *Qrock* antara lain sebagai berikut:

1. *Neighbor* (Tetangga)

Sebagaimana algoritma *Rock*, algoritma *Qrock* juga melibatkan konsep *neighbor* 1 dan 0. Dua objek p_i dan p_j dikatakan *neighbor* jika $sim(p_i, p_j) \geq \theta$

(*Threshold).* *Threshold* merupakan parameter yang ditentukan oleh peneliti yang dapat digunakan untuk mengontrol seberapa dekat hubungan p_i dan p_j sehingga kedua objek tersebut bisa dikatakan sebagai *neighbor*.

2. MFSET

Tidak seperti algoritma *Rock* yang menggunakan informasi *link*, algoritma *Qrock* menggunakan konsep MFSET untuk membentuk *cluster-cluster* nya. MFSET terdiri dari tiga operasi, antara lain:

a. *Initial(x)* : Membentuk himpunan yang hanya beranggotakan elemen *x.*
b. *Find(x)* : Mencari himpunan yang salah satu anggotanya adalah elemen *x.*
c. *Merge(A,B)* : Gabungan dari himpunan *A* dan himpunan *B.*

Algoritma *Qrock* dimulai dengan menghitung kemiripan antar objek. Berdasarkan matriks kemiripan dan *threshold* yang diberikan, dihitung matriks tetangga untuk setiap objek *i*. Inisialisasi setiap objek berfungsi sebagai himpunan. Untuk setiap *i,* diambil objek *x*. kemudian menggabungkan dengan komponen lain.

Pajak Pertambahan Nilai

Menurut [12], Pajak Pertambahan Nilai (PPN) yaitu penggantian pajak penjualan, karena pajak ini tidak bisa memadai dan mencapai sasaran kebutuhan pembangunan masyarakat dan menampung kegiatannya, kegiatan tersebut yaitu pemerataan dalam membebankan pajak, meningkatkan sumber penerimaan negara, dan mendorong produk ekspor. PPN ialah pajak atas konsumsi barang dan jasa yang dikenakan didalam negeri (didalam daerah pabean).

PPN yang diterapkan di Indonesia adalah PPN Tipe Konsumsi *(Consumption Type VAT)*. Dilihat dari sisi perlakuan terhadap barang modal, artinya seluruh biaya yang dikeluarkan untuk perolehan barang modal dapat dikurangi dari dasar pengenaan pajak. Dalam bahasa *indirect subtraction method,* Pajak Masukan (*input tax)* sehingga barang modal dapat dikreditkan dengan Pajak Keluarkan *(output tax).*

Badan Pajak Pertambahan Nilai

Menurut Undang-Undang Nomor 8 Tahun 1983 tentang PPN barang dan jasa & PPnBM Nomor 42 Tahun 2009 badan adalah sekumpulan orang/ modal yang merupakan kesatuan baik yang melakukan usaha maupun tidak melakukan usaha yang meliputi Perseroan Terbatas (PT), Perseroan Komanditer/*Commanditaire Vennootschap* (CV), Badan Usaha Milik Negara(BUMN), Badan Usaha Milik Daerah (BUMD) dengan nama dan dalam bentuk apapun, firma, kongsi, Koperasi, Dana Pensiun, Persekutuan, Perkumpulan, Yayasan, Organisasi massa, Organisasi sosial politik lembaga dan bentuk badan lainnya.

METODOLOGI PENELITIAN

Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini merupakan data wajib pajak badan di KPP Pratama Samarinda Ulu Tahun 2018. Data tersebut akan dianalisis menggunakan analisis *cluster* yaitu algoritma *Quick Robust Clustering using link (QROCK).* Adapun variabel yang digunakan dalam penelitian ini yaitu 4 variabel bertipe kategorik yang terdiri dari Data Status Pembayaran Pajak (X_1), Pendapatan (Rupiah) (X_2), Bentuk Badan(X_3), Status Pelaporan Pajak(X_4).

Metode Analisis

Langkah-langkah dalam Algoritma *Qrock* sebagai berikut:

1. Menghitung *Similarity*

Similarity ukuran kemiripan antara pasangan objek ke-*i* dan objek ke-*j* dengan rumusan yang didefiniskan sebagai berikut:

$$sim(p_i, p_j) = \left|\frac{p_i \cap p_j}{p_i \cup p_j}\right|, i \neq j \qquad (1)$$

dengan,

$i = 1,2,3, \dots, n$ dan $j = 1,2,3, \dots, n$

p_i= Himpunan pengamatan ke-*i*, dengan $p_i = \{p_{1i}, p_{2i}, p_{3i}, \dots, p_{4i}\}$

p_j = Himpunan pengamatan ke-*j*, dengan

$p_j = \{p_{1j}, p_{2j}, p_{3j}, \dots, p_{4j}\}$

2. Menentukan *neighbors* (tetangga)

Pengamatan dinyatakan sebagai *neighbors* jika nilai $sim(p_i, p_j) \geq \theta$

3. Mendapatkan *neighbors* dari masing-masing objek

Neighbors berisi semua objek pengamatan yang mempunyai nilai similaritas dengan objek *i* lebih besar atau sama dengan θ, atau jika nilai $sim(p_i, p_j) \geq \theta$

4. Melakukan Pengelompokkan

Pada *neighbors* ambil komponen yang memuat *initial x* dalam nilai $sim(p_i, p_j) \geq \theta$

A=find(x)

B=find(y)

Jika $A \neq B$, maka merge (*A,B).*

5. Interpretasi karakteristik hasil pengelompokkan

HASIL DAN PEMBAHASAN

Analisis *Algoritma Qrock* dimulai dengan menghitung kemiripan antar objek. Berdasarkan matriks kemiripan dan *threshold* yang diberikan, dihitung matriks tetangga untuk setiap objek *i*. Inisialisasi setiap objek berfungsi sebagai himpunan. Untuk setiap *i*, diambil objek *x*. kemudian menggabungkan dengan komponen lain. Pada penelitian ini menggunakan 200 data pengamatan Wajib Pajak dengan 4 variabel bertipe kategorik yang terdiri dari Data Status Pembayaran Pajak (X_1), Pendapatan (Rupiah) (X_2), Bentuk Badan(X_3), Status Pelaporan Pajak(X_4). Menggunakan nilai θ sebesar 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; dan 0,9. Analisis dilakukan menggunakan bantuan *software R*.

Adapun tahapan dalam analisis *Algoritma Qrock* adalah sebagai berikut:

1. Menghitung *Similarity* antar objek

Contoh perhitungan *similarity* adalah sebagai berikut:

a. $sim(p_1, p_j)$; dengan $j = 2,3, \dots 200$

Untuk objek 1 dan 2

p_1 ={Patuh, <100 Juta, CV,Tepat Waktu}

p_2 ={Patuh, 1-10 M,CV, Tidak Tepat Waktu}

$$sim(p_1, p_2) = \left|\frac{p_1 \cap p_2}{p_1 \cup p_2}\right| = \frac{2}{6} = 0{,}333$$

Untuk objek 1 dan 3

p_1 ={Patuh, <100 Juta, CV,Tepat Waktu}

p_3 ={Tidak Patuh, <100 Juta,BUMN/BUMD, Tepat Waktu}

$$sim(p_1, p_3) = \left|\frac{p_1 \cap p_3}{p_1 \cup p_3}\right| = \frac{2}{6} = 0{,}333$$

⋮

Untuk objek 199 dan 200

p_{199} ={Patuh, <100 Juta, CV,Tepat Waktu}

p_{200} ={Tidak Patuh, >10 M,Koperasi, Tidak Tepat Waktu}

$$sim(p_1, p_{200}) = \left|\frac{p_1 \cap p_{200}}{p_1 \cup p_{200}}\right| = \frac{0}{8} = 0$$

Perhitungan *similarity* dari setiap objek membentuk matriks **sim** yang merupakan matriks simteris 200x200 (banyaknya objek).

2. Menentukan *neighbors* (Tetangga)

Suatu pasangan objek dinyatakan sebagai tetangga jika nilai dari $sim(p_i, p_j) \geq \theta$. Informasi mengenai hubungan tetangga antar objek pengamatan dapat dinyatakan dengan matriks **A**. Matriks **A** merupakan matriks simetris berukuran 200×200 yang bernilai 1 jika objek tersebut memenuhi syarat bertetangga dan bernilai 0 jika objek tersebut tidak memenuhi jarak bertetangga. Di dapatkan Sembilan matriks simetriks **A** berdasarkan θ sebesar 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; dan 0,9.

3. Mendapatkan *Neighbors* dari masing-masing objek

Setelah mendapatkan matriks **A** untuk θ = 0,1;0,2;0,3;0,4;0,5;0,6;0,7;0,8 dan 0,9 maka selanjutnya akan dilakukan penentuan *neighbors* dari masing-masing objek. *Neighbors* $p_i$berisi semua objek amatan yang mempunyai nilai similaritas

dengan objek lebih besar atau sama dengan θ.

a. Nilai θ sebesar 0,1 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_2, p_3, p_4, \dots, p_{199}]$
Neighbors $[p_2]= [p_1, p_2, p_6, p_7, \dots, p_{200}]$
Neighbors $[p_3]= [p_1, p_3, p_4, p_5, \dots, p_{200}]$

⋮

Neighbors $[p_{200}]= [p_2, p_3, p_4, p_5, \dots, p_{200}]$

b. Nilai θ sebesar 0,2 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_2, p_3, p_4, \dots, p_{199}]$
Neighbors $[p_2]= [p_1, p_7, p_{10}, p_{11}, \dots, p_{200}]$
Neighbors $[p_3]= [p_1, p_3, p_4, p_5, \dots, p_{200}]$

⋮

Neighbors $[p_{200}]= [p_2, p_3, p_4, p_5, \dots, p_{200}]$

c. Nilai θ sebesar 0,3 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_2, p_3, p_4, \dots, p_{199}]$
Neighbors $[p_2]= [p_1, p_7, p_{10}, p_{11}, \dots, p_{200}]$
Neighbors $[p_3]= [p_1, p_3, p_4, p_5, \dots, p_{200}]$

⋮

Neighbors $[p_{200}]= [p_2, p_3, p_4, p_5, \dots, p_{200}]$

d. Nilai θ sebesar 0,4 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_7, p_8, p_{10}, \dots, p_{199}]$
Neighbors $[p_2]= [p_2, p_{10}, p_{24}, p_{42}, \dots, p_{197}]$
Neighbors $[p_3]= [p_3, p_4, p_5, p_{14}, \dots, p_{172}]$

⋮

Neighbors $[p_{200}]= [p_{48}, p_{65}, p_{77}, \dots, p_{200}]$

e. Nilai θ sebesar 0,5 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_7, p_8, p_{10}, \dots, p_{199}]$
Neighbors $[p_2]= [p_2, p_{10}, p_{24}, p_{42}, \dots, p_{197}]$
Neighbors $[p_3]= [p_3, p_4, p_5, p_{14}, \dots, p_{172}]$

⋮

Neighbors $[p_{200}]= [p_{48}, p_{65}, p_{77}, \dots, p_{200}]$

f. Nilai θ sebesar 0,6 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_7, p_8, p_{10}, \dots, p_{199}]$
Neighbors $[p_2]= [p_2, p_{10}, p_{24}, p_{42}, \dots, p_{197}]$
Neighbors $[p_3]= [p_3, p_4, p_5, p_{14}, \dots, p_{172}]$

⋮

Neighbors $[p_{200}]= [p_{48}, p_{65}, p_{77}, \dots, p_{200}]$

g. Nilai θ sebesar 0,7 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_{40}, p_{58}, p_{69}, \dots, p_{199}]$
Neighbors $[p_2]= [p_2, p_{89}, p_{90}, p_{153}, p_{183}]$
Neighbors $[p_3]= [p_3, p_4, p_5, p_{62}, \dots, p_{172}]$

⋮

Neighbors $[p_{200}]= [p_{195}, p_{200}]$

h. Nilai θ sebesar 0,8 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_{40}, p_{58}, p_{69}, \dots, p_{199}]$
Neighbors $[p_2]= [p_{89}, p_{90}, p_{153}, p_{183}]$
Neighbors $[p_3]= [p_3, p_4, p_5, p_{62}, \dots, p_{172}]$

⋮

Neighbors $[p_{200}]= [p_{195}, p_{200}]$

i. ilai θ sebesar 0,9 maka akan didapatkan *neighbors* sebagai berikut:

Neighbors $[p_1]= [p_1, p_{40}, p_{58}, p_{69}, \dots, p_{199}]$
Neighbors $[p_2]= [p_{89}, p_{90}, p_{153}, p_{183}]$
Neighbors $[p_3]= [p_3, p_4, p_5, p_{62}, \dots, p_{172}]$

⋮

Neighbors $[p_{200}]= [p_{195}, p_{200}]$

4. Menentukan Pengelompokkan

Setelah mendapatkan *neighbors* dari masing-masing objek maka selanjutnya akan dilakukan pengelompokkan. Pengelompokkan dilakukan dengan memilih objek *x (find(x))* misalkanp_1 dan objek *y (find(y))* misalkan p_2 dengan nilai *x* $\neq$ *y*, kemudian menggabungkan dua objek (*merge(A,B))* jika nilai keduanya berbeda.

a. Mengelompokkan *neighbors* dengan nilai θ = 0,1

Untuk p_i ; dengan i = 1

Ambil satu nilai x dalam *neighbors* $[p_1]$= $[p_1, p_2, p_3, p_4, \dots, p_{199}]$

$x = p_1$

Untuk setiap nilai $y \neq x$ dalam *neighbors* $[p_1]$ = $[p_1, p_2, p_3, p_4, \dots, p_{199}]$

$y = p_2$

$A = find(x) = find\ (p_1) = \{p_1\}$

$B = find(y) = find\ (p_2) = \{p_2\}$

Karena A ≠ B maka *merge* (A,B)= $\{p_1, p_2\}$

Sehingga *cluster* sementara yang terbentuk adalah 199 cluster yaitu: $\{p_1, p_2\}, \{p_3\}, \{p_4\}, \{p_5\}, \{p_6\}, \dots, \{p_{200}\}$

$y = p_3$

$A = find(x) = find\ (p_1) = \{p_1, p_2\}$

$B = find(y) = find\ (p_3) = \{p_3\}$

Karena A ≠ B maka *merge* (A,B)= $\{p_1, p_2, p_3\}$

Sehingga *cluster* sementara yang terbentuk adalah 198 cluster yaitu: $\{p_1, p_2, p_3\}, \{p_4\}, \{p_5\}, \{p_6\}, \dots, \{p_{200}\}$

⋮

$y = p_{199}$

$A = find(x) = find\ (p_1) = \{p_1, .., p_{198}\}$

$B = find(y) = find\ (p_{199}) = \{p_{199}\}$

Karena A ≠ B maka *merge* (A,B)= $\{p_1, p_2, p_3\}$

Sehingga *cluster* sementara yang terbentuk adalah 12 cluster yaitu: $\{p_1, p_2, p_3, \dots, , p_{199}\}, \{p_{30}\}, \{p_{72}\}, \{p_{77}\}, \{p_{104}\}, \{p_{107}\}, \{p_{77}\}, \dots \{p_{200}\}$

Perhitungan dilanjutkan hingga p_i ; dengan i= 200 untuk nilai θ = 0,1. Proses pengelompokkan selesai sehingga menghasilkan *cluster* akhir sebanyak 1 *cluster*. Proses perhitungan pengelompokkan *neighbors* dilanjutkan dengan nilai θ = 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8 hingga 0,9.

5. Interpretasi Karakteristik hasil *clustering*

Berikut merupakan hasil *clustering* data kategorik yang dihasilkan oleh nilai θ = 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8 dan 0,9 sebagai berikut :

Tabel 1. Hasil *Clustering* Data Kategorik

Treshold	***Cluster* yang dihasilkan**
0,1	1
0,2	1
0,3	1
0,4	1
0,5	1
0,6	1
0,7	56
0,8	56
0,9	56

Berdasarkan **Tabel 1**. dapat diketahui bahwa ketika nilai θ = 0,1;0,2;0,3;0,4;0,5 dan 0,6 *cluster* yang dihasilkan hanya 1 *cluster* , sedangkan ketika nilai θ = 0,7;0;8 dan 0,9 menghasilkan 56 *cluster*. Adapun anggota pengamatan dari 56 *cluster* masing-masing dapat dilihat pada **Tabel 2**. sebagai berikut:

Tabel 2. Anggota Pengamatan Masing-masing *cluster*

Cluster	**Anggota Pengamatan**
1	1,40,58,69,82,91,106,115,116,120,123,149,150,173,177,190,199
2	2,89,90,153,183
3	3,4,5,62,68,71,136,171,172
4	6,33,95,130,141,147,159,160,165
5	7,15,57,111,138,184,187,198
6	8,18,25,36,38,50,55,74,87,131,139,163,164,179,182
7	9,73
8	10,51,167
9	11,39,81,100,101,186
10	12
11	13,23,85,134
12	14,135
13	16
14	17,189
15	19
16	20,21,26,53,78,142,145,181,192,193
17	22
18	24,42,97,119,185

Cluster	Anggota Pengamatan
19	27,29,60,110,117,168,174
20	28,44,49,75
21	30
22	31,34,52,113,121,144,169
23	32,43,66,158
24	35,64
25	37,70,112
26	41,79,133
27	45,140,157
28	46,99,128,194
29	47,84,122
30	48,108
31	80
32	56,155,188
33	59,143
34	61,88,94,105,152,154,166
35	63,92
36	65
37	67,98,162
38	72
39	76
40	77
41	83,118,124,178
42	86,148,156
43	54,93,125,170
44	96,151,197
45	102,109
46	103,161
47	104,107,132
48	144
49	126,137
50	127,129
51	146
52	175
53	176
54	180,191
55	195,200
56	196

Berdasarkan **Tabel 2**. dapat diketahui anggota pengamatan dari masing-masing *cluster*. Adapun contoh beberapa karakteristik yaitu *cluster* 1 sampai *cluster* 20 sebagai berikut:

1. *Cluster* 1
   Cluster 1 merupakan *cluster* yang terdiri dari 17 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan <100 Juta, bentuk badan CV dan status pelaporan pajak tepat waktu.
2. *Cluster* 2
   Cluster 2 merupakan *cluster* yang terdiri dari 5 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan 1 – 10 Miliar, bentuk badan CV dan status pelaporan pajak tidak tepat waktu.
3. *Cluster* 3
   Cluster 3 merupakan *cluster* yang terdiri dari 9 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan <100 Juta, bentuk badan BUMN/BUMD dan status pelaporan pajak tepat waktu.
4. *Cluster* 4
   Cluster 4 merupakan *cluster* yang terdiri dari 9 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan <100 Juta, bentuk badan BUMN/BUMD dan status pelaporan pajak tepat waktu.
5. *Cluster* 5
   Cluster 5 merupakan *cluster* yang terdiri dari 8 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan 500 Juta – 1 Miliar, bentuk badan CV dan status pelaporan pajak tepat waktu.
6. *Cluster* 6
   Cluster 6 merupakan *cluster* yang terdiri dari 15 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan <100 Juta, bentuk badan PT dan status pelaporan pajak tepat waktu
7. *Cluster* 7
   Cluster 7 merupakan *cluster* yang terdiri dari 2 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan 100 - 500 Juta, bentuk badan PT dan status pelaporan pajak tepat waktu.
8. *Cluster* 8
   Cluster 8 merupakan *cluster* yang terdiri dari 3 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan <100 Juta, bentuk

badan CV dan status pelaporan pajak tidak tepat waktu.

9. *Cluster* 9
   Cluster 9 merupakan *cluster* yang terdiri dari 6 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan <100 Juta, bentuk badan CV dan status pelaporan pajak tidak tepat waktu.
10. *Cluster* 10
    Cluster 10 merupakan *cluster* yang terdiri dari 1 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan <100 Juta, bentuk badan PT dan status pelaporan pajak tidak tepat waktu.
11. *Cluster* 11
    Cluster 11 merupakan *cluster* yang terdiri dari 4 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan 500 Juta – 1 Miliar, bentuk badan CV dan status pelaporan pajak tepat waktu.
12. *Cluster* 12
    Cluster 12 merupakan *cluster* yang terdiri dari 2 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan <100 Juta, bentuk badan CV dan status pelaporan pajak tepat waktu.
13. *Cluster* 13
    Cluster 13 merupakan *cluster* yang terdiri dari 1 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan 100 – 500 Juta, bentuk badan Koperasi dan status pelaporan pajak tepat waktu.
14. *Cluster* 14
    Cluster 14 merupakan *cluster* yang terdiri dari 2 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan 100 - 500 Juta, bentuk badan Koperasi dan status pelaporan pajak tepat waktu.
15. *Cluster* 15
    Cluster 15 merupakan *cluster* yang terdiri dari 1 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan <100 Juta, bentuk badan lainnya dan status pelaporan pajak tidak tepat waktu.
16. *Cluster* 16
    Cluster 16 merupakan *cluster* yang terdiri dari 10 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan 100 – 500 Juta, bentuk badan CV dan status pelaporan pajak tepat waktu.
17. *Cluster* 17
    Cluster 17 merupakan *cluster* yang terdiri dari 1 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan <100 Juta, bentuk badan PT dan status pelaporan pajak tepat waktu.
18. *Cluster* 18
    Cluster 18 merupakan cluster yang terdiri dari 5 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan 1 - 10 Miliar, bentuk badan CV dan status pelaporan pajak tepat waktu.
19. *Cluster* 19
    Cluster 19 merupakan *cluster* yang terdiri dari 7 wajib pajak yang memiliki karakteristik status pembayaran PPN patuh, pendapatan >10 Miliar, bentuk badan PT dan status pelaporan pajak tepat waktu.
20. *Cluster* 20

Cluster 20 merupakan *cluster* yang terdiri dari 4 wajib pajak yang memiliki karakteristik status pembayaran PPN tidak patuh, pendapatan 1 - 10 Miliar, bentuk badan CV dan status pelaporan pajak tepat waktu.

KESIMPULAN

Berdasarkan hasil analisis dan pembahasan, maka didapatkan kesimpulan sebagai berikut:

1. Hasil Pengelompokkan data wajib pajak badan di KPP Pratama Samarinda Ulu tahun 2018 dengan menggunakan algoritma *Qrock* menghasilkan 1 *cluster* pada nilai *threshold* (θ) 0,1;0,2;0,3;0,4;0,5 dan

0,6 sedangkan pada nilai *threshold* (θ) 0,7;0,8 dan 0,9 menghasilkan 56 *cluster*.

2. Karakteristik yang dihasilkan dari 56 *cluster* beraneka ragam sebagai contoh untuk karakteristik *cluster* 18 didapatkan hasil *cluster* yang terdiri dari 5 wajib pajak badan dengan status pembayaran PPN patuh, pendapatan 1 - 10 Miliar, bentuk badan CV dan status pelaporan pajak tepat waktu.

DAFTAR PUSTAKA

[1]Agresti, A. (1990). *Categorical Data Analysis.* USA: John Wiley and Sons.

[2]Alamsyah, M. (2006). *Pengelompokaan Data Kategoril dengan Algoritma Qrock.* Surabaya: Universitas Airlangga.

[3] Dutta., M.Mahanta, A. K., & Arun, K. P. (2005). *QROCK: A Quick Version of he ROCK Algorithm for Clustering of Categorical Data. Proceedings of the 15 IE International Conference on Data Engineering*.

[4] Erly, S. (2002). *Perpajakan.* Jakarta: Salemba Empat.

[5] Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons Inc.

[6] Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

[7]_____(2014). *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

[8] Santosa, B. (2007). Data Mining: *Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.

[9] Santoso, S. (2015). *Menguasai Statistik Multivariat Konsep Dasar dan Aplikasi dengan SPSS*. Jakarta:PT Elex Media Komputindo.

[10] Supranto, J. (2010). *Analisis Multivariat Arti dan Interpretasi.* Jakarta: Rineka Cipta.

[11] Turban, E., Aronson, Jay, E. & Peng, L.T. (2005). *Decision Support Systems and Intelligent Systems*. Yogyakarta: Andi Offset.

[12] Waluyo. (2011). *Perpajakan Indonesia Edisi 10 Buku 1.* Jakarta : Selemba Empat.