TESIS – EE185401

CLUSTERING KLASIFIKASI LAPANGAN USAHA (KLU) WAJIB PAJAK BADAN POTENSIAL MENGGUNAKAN METODE K-MEANS

JESSICA RAHMAWATI NUGROHO
07111950067012

DOSEN PEMBIMBING
Prof. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
Eko Setijadi, ST., MT., Ph.D

PROGRAM MAGISTER
BIDANG KEAHLIAN TELEMATIKA (PETIK)
DEPARTEMEN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2021

TESIS – EE185401

CLUSTERING KLASIFIKASI LAPANGAN USAHA (KLU) WAJIB PAJAK BADAN POTENSIAL MENGGUNAKAN METODE K-MEANS

JESSICA RAHMAWATI NUGROHO
07111950067012

DOSEN PEMBIMBING
Prof. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
Eko Setijadi, ST., MT., Ph.D

PROGRAM MAGISTER
BIDANG KEAHLIAN TELEMATIKA (PETIK)
DEPARTEMEN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI ELEKTRO DAN INFORMATIKA CERDAS
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2021

LEMBAR PENGESAHAN TESIS

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Teknik (MT)
di
Institut Teknologi Sepuluh Nopember

Oleh

JESSICA RAHMAWATI NUGROHO
NRP: 07111950067012

Tanggal Ujian: 06 Agustus 2021
Periode Wisuda: Oktober 2021

Disetujui oleh
Pembimbing:

1. **Prof.Dr.Ir. Yoyon Kusnendar Suprapto, MSc.**
   NIP: 195409251978031001

2. **Eko Setijadi, ST.,MT.,Ph.D.**
   NIP: 197210012003121002

Penguji:

1. **Dr.Ir. Endroyono, DEA.**
   NIP: 196504041991021001

2. **Reza Fuad Rachmadi, ST, MT., Ph.D**
   NIP: 198504032012121000

3. **Dr. Adhi Dharma Wibawa, S.T., M.T.**
   NIP: 197605052008121003

4. **Dr. Supeno Mardi Susiki Nugroho, ST., M.T.**
   NIP: 197003131995121001

Kepala Departemen Teknik Elektro
Dedet Candra Riawan, S.T., M.Eng., Ph.D.
NIP: 197311192000031001

Halaman ini sengaja dikosongkan

PERNYATAAN KEASLIAN TESIS

Dengan ini saya menyatakan bahwa isi keseluruhan Tesis saya dengan judul "**CLUSTERING KLASIFIKASI LAPANGAN USAHA (KLU) WAJIB PAJAK BADAN POTENSIAL MENGGUNAKAN METODE K-MEANS**" adalah benar-benar hasil karya intelektual mandiri, diselesaikan tanpa menggunakan bahan-bahan yang tidak diijinkan dan bukan merupakan karya pihak lain yang saya akui sebagai karya sendiri.

Semua referensi yang dikutip maupun dirujuk telah ditulis secara lengkap pada daftar pustaka. Apabila ternyata pernyataan ini tidak benar, saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Surabaya, Juni 2021

Jessica Rahmawati Nugroho
NRP. 07111950067012

Halaman ini sengaja dikosongkan

CLUSTERING KLASIFIKASI LAPANGAN USAHA (KLU) WAJIB PAJAK BADAN POTENSIAL MENGGUNAKAN METODE K-MEANS

Nama mahasiswa : Jessica Rahmawati Nugroho
NRP : 07111950067012
Pembimbing : 1. Prof. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
2. Eko Setijadi, ST., MT., Ph.D

ABSTRAK

Belum optimalnya penerimaan pajak yang terlihat dari *tax gap* dan *tax ratio* menunjukkan bahwa tingkat kepatuhan di Indonesia masih rendah. Salah satu upaya untuk meminimalisir resiko ketidakpatuhan Wajib Pajak (WP) adalah dengan melakukan pengawasan dan pemeriksaan terhadap WP. Sebagai penopang penerimaan terbesar, WP badan berdasarkan sektor usahanya/klasifikasi lapangan usahanya mempunyai kontribusi dominan terhadap penerimaan negara setiap tahunnya. Namun, terbatasnya jumlah pemeriksa pajak menyebabkan kegiatan pemeriksaan dan pengawasan menjadi kurang optimal. Dari latar belakang tersebut, diambil pendekatan menggunakan metode *clustering* dengan algoritma K-Means untuk mengelompokkan Klasifikasi Lapangan Usaha (KLU) yang memiliki potensi bagi penerimaan pajak dari data yang dimiliki oleh Direktorat Jenderal Pajak (DJP). Hasil pengujian *clustering* dengan algoritma K-Means menunjukkan, 173 KLU memiliki tingkat kepatuhan rendah, 156 KLU memiliki tingkat kepatuhan sedang, 684 KLU memiliki tingkat kepatuhan tinggi, 967 KLU memiliki dampak fiskal rendah, 38 KLU memiliki dampak fiskal sedang, dan 8 KLU memiliki dampak fiskal tinggi. Validasi *clustering* menggunakan uji *silhouette*, diperoleh nilai 0.69 untuk variabel x dan 0.92 untuk variabel y. Informasi yang dihasilkan dari penelitian ini dapat digunakan untuk mendukung pengambilan keputusan dalam penentuan daftar KLU yang perlu diprioritaskan untuk dilakukan pemeriksaan dan pengawasan.

Kata kunci: clustering, DJP, KLU, K-Means, pajak

Halaman ini sengaja dikosongkan

CLUSTERING BUSINESS CLASSIFICATION FOR POTENTIAL CORPORATE TAXPAYERS USING THE K-MEANS METHOD

By : Jessica Rahmawati Nugroho
Student Identity Number : 07111950067012
Supervisor(s) : 1. Prof. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
2. Eko Setijadi, ST., MT., Ph.D

ABSTRACT

Not yet optimal tax revenue, which can be seen from the tax gap and tax ratio, indicates that the compliance level in Indonesia is still low. One of the efforts to minimize the risk of non-compliance taxpayers is to carry out supervision and inspection of taxpayers. As one of the largest sources of revenue, corporate taxpayers based on their business sector have a dominant contribution to state revenues every year. With limited human resources, it makes supervision and inspection, less than optimal. Thus, in this study, the researcher tries to implement the clustering method with the K-Means algorithm to grouping business classification (Klasifikasi Lapangan Usaha/KLU) with the potential for tax from the Directorate General of Taxes (DGT) data. This study used the K-Means clustering method. The clustering test results utilizing the K-Means algorithm revealed that 173 KLUs had a low level of compliance, 156 KLUs had a medium level of compliance, 684 KLUs had a high level of compliance, 967 KLUs had a low fiscal impact, 38 KLUs had a medium fiscal impact, and 8 KLUs had a high fiscal impact. Clustering validation using the silhouette index, obtained values of variable x and variable y, respectively 0,69 and 0,92. The information provided from this study can be used to support decision making in determining the list of KLUs that need to be prioritized for supervising and inspecting.

Key words: clustering, DGT, KLU, K-Means, tax

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Alhamdulillah segala puji bagi Allah SWT, Shalawat serta salam semoga senantiasa tercurah kepada junjungan kita Nabi Muhammad SAW beserta keluarga, sahabat dan seluruh pengikutnya. Syukur senantiasa Penulis ucapkan kepada Allah Yang Maha Penyayang atas petunjuk, rahmat, kasih sayang, karuniaNya penulis dapat menyelesaikan penyusunan tesis ini, dengan judul "**CLUSTERING KLASIFIKASI LAPANGAN USAHA (KLU) WAJIB PAJAK BADAN POTENSIAL MENGGUNAKAN METODE K-MEANS**". Tesis ini disusun sebagai syarat kelulusan pada Program Magister Bidang Keahlian Telematika, Departemen Teknik Elektro, Fakultas Teknologi Elektro dan Informatika Cerdas, Institut Teknologi Sepuluh Nopember, Surabaya.

Penyusunan tesis ini tidak terlepas dari dukungan berbagai pihak. Pada kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Bapak Prof. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc. dan Bapak Eko Setijadi, ST., MT., Ph. D. selaku dosen pembimbing, atas waktu, motivasi dan bimbingannya dalam penyusunan tesis ini.
2. Bapak Dr. Adhi Dharma Wibawa, S.T., M.T. selaku koordinator Bidang Keahlian Telematika atas arahannya selama perkuliahan.
3. Almarhum Bapak Dr. Istas Pratomo, ST., MT. selaku dosen Manajemen Jaringan. Terima kasih atas arahan, saran, dan masukannya dalam penyusunan awal proposal tesis penulis. Semangat beliau dalam berbagi dengan mahasiswa menjadi salah satu inspirasi bagi penulis.
4. Para dosen Program Magister Bidang Keahlian Telematika, Departemen Teknik Elektro, Fakultas Teknologi Elektro dan Informatika Cerdas, Institut Teknologi Sepuluh Nopember, Surabaya atas ilmu, motivasi, pengalaman, dan arahannya selama perkuliahan.
5. Kepala Badan Litbang SDM Kementrian Komunikasi dan Informatika yang telah memberikan bantuan berupa beasiswa studi.

6. Direktorat Data dan Informasi Perpajakan, Direktorat Jenderal Pajak, Kementrian Keuangan Republik Indonesia yang telah memberi dukungan data untuk penelitian ini.
7. Papa Nugroho, mama Lis Eni Farida, ibu mertua Chusnul Chotimah, dan bapak mertua Harun Al Rosyid, terima kasih atas dukungan dan doa nya yang tidak pernah putus. Penulis percaya tidak ada yang lebih dahsyat daripada doa orang tua kepada anaknya.
8. Suami tercinta, Ahmad Baihaqi, dan terkhusus anakku tersayang, Aisha Qiana Ahmad, terima kasih atas dukungan, motivasi, serta pengertian selama mendampingi dalam penyelesaian studi S2 ini.
9. Rekan-rekan S2 Telematika angkatan 2019 yang selalu kompak dan saling membantu selama proses perkuliahan. Serta angkatan 2018 dan angkatan lainnya, terima kasih atas kebersamaan, keceriaan, dan kekompakan selama ini.
10. Semua pihak yang telah membantu dalam proses penyelesaian studi dan penyelesaian tesis ini.

Penulis menyadari tesis ini masih jauh dari sempurna, maka perlu masukan maupun kritikan yang membangun untuk penelitian ini. Semoga karya ini dapat bermanfaat bagi pembaca. Mohon maaf atas segala kesalahan dan kekurangan, Semoga rahmat Allah senantiasa tercurah kepada kita semua.

Surabaya, 17 Juni 2021

Jessica Rahmawati Nugroho

DAFTAR ISI

DAFTAR GAMBAR

Halaman ini sengaja dikosongkan

DAFTAR TABEL

NOMENKLATUR

Simbol	Definisi
d	Jarak titik
c_i	Centroid ke- i
x_i	Objek ke- i
v	Centroid pada *cluster*
n	Jumlah objek dalam satu *cluster*
$a(i)$	Rata -rata jarak entitas ke-i
$b(i)$	Nilai minimum dari jarak rata-rata entitas i
$s(i)$	Shilouette index
v'	Nilai baru
μ	Rata-rata populasi
σ	Standar deviasi populasi
s^2	Varian

Halaman ini sengaja dikosongkan

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Pajak adalah sumber utama pendapatan negara. Otoritas pajak di seluruh dunia berfungsi untuk mengurangi kesenjangan pajak (*tax gap*) yaitu perbedaan antara kewajiban pajak yang seharusnya dibayar ke negara, dan jumlah yang sebenarnya dibayar [13]. *Tax gap* dibagi menjadi tiga komponen, yaitu: *nonfilling gap, underreporting gap,* dan *underpayment gap*. *Non filling gap* adalah kesenjangan karena Wajib Pajak (WP) tidak tepat waktu dalam melakukan laporan atau tidak melapor sama sekali. *Underreporting gap* adalah kesenjangan karena WP mengurangi jumlah penghasilan atau melebihkan pengurangan pajaknya. Dan *underpayment gap* adalah kesenjangan karena WP gagal melakukan pembayaran pajak pada tanggal jatuh tempo pembayaran [2].

Salah satu faktor yang menyebabkan adanya kesenjangan pajak adalah tingkat kepatuhan Wajib Pajak (WP). Kepatuhan WP memerlukan kesadaran dan kepatuhan pembayar pajak terhadap norma dan aturan perundang-undangan yang berlaku dalam pemenuhan kewajiban perpajakan. Sedangkan WP yang tidak patuh, cenderung akan melakukan kegiatan penghindaran pajak baik secara formal maupun informal. Termasuk dalam penghindaran pajak, yaitu pengurangan pajak yang legal secara hukum, maupun penghindaran pajak dengan melakukan pelalaian pembayaran pajak. Pada akhirnya serangkaian kegiatan ini akan merugikan negara dan menyebabkan penerimaan pajak negara berkurang. Mengingat penerapan *Self Asessment* pada sistem perpajakan Indonesia, dimana WP dipercaya untuk menghitung, membayar, dan melaporkan pajaknya sendiri sesuai dengan peraturan perundang-undangan perpajakan yang berlaku, maka kepatuhan WP menjadi aspek penting.

Meskipun kesadaran WP semakin meningkat dari tahun ke tahun, namun Direktorat Jenderal Pajak (DJP) yang berada di bawah Kementrian Keuangan sebagai perwakilan negara dalam menghimpun pajak di Indonesia tetap berkewajiban untuk memperkuat strategi kepatuhan dengan tujuan menimalisir

kesenjangan pajak. Karena pada tahun 2019, jumlah penduduk Indonesia hasil proyeksi Badan Pusat Statistik (BPS) mencapai 266,911 juta jiwa. Dengan sekurangnya 100 juta pelaku bisnis, hanya 45.343.014 penduduk yang terdaftar menjadi WP. Dan dari jumlah WP terdaftar tersebut hanya 30.334.994 WP atau sekitar 66,9% yang patuh menyampaikan kewajiban perpajakannya. Selain itu, dari *tax gap* dan *tax ratio* terlihat bahwa penerimaan pajak belum optimal, hal ini menunjukkan bahwa tingkat kepatuhan di Indonesia masih rendah. Berdasarkan data DJP, rasio pajak Indonesia pada tahun 2019 adalah sebesar 10,7%. Angka ini turun 0,3% jika dibandingkan tahun sebelumnya yang berada di kisaran 11%. Rasio pajak ini tercatat sebagai yang terendah se-Asean. Sebagai perbandingan, pada tahun yang sama data *Organisation for Economic Co-operation and Development* (OECD) menunjukkan rasio pajak Singapura berada di level 13,2%, Malaysia pada level 12,5%, dan Thailand pada level 17,5%.

Tingkat kepatuhan yang rendah mungkin disebabkan oleh kondisi sistem administrasi perpajakan yang belum maksimal dan lemahnya kebijakan serta peraturan yang menyertai. Faktor-faktor seperti kurangnya pengawasan, kurang menyeluruhnya sosialisasi, kurangnya tingkat kesadaran WP, serta kurangnya kemudahan layanan juga dapat berpengaruh terhadap kepatuhan WP. Sistem pelayanan yang terlalu rumit akan mengakibatkan penerimaan pajak yang rendah sehingga menurunkan tingkat penerimaan.[7].

Salah satu upaya untuk meminimalisir resiko ketidakpatuhan WP adalah dengan melakukan pengawasan dan pemeriksaan terhadap WP [14]. Berdasarkan PMK nomor 17 tahun 2013 terdapat dua sebab WP diperiksa, yang pertama adalah untuk menilai tingkat kepatuhan WP dalam melaksanakan pemenuhan kewajiban perpajakannya, yang kedua dalam hal WP mengajukan restitusi atau pengembalian kelebihan pembayaran pajak.

Sebagai penopang penerimaan terbesar, WP badan berdasarkan sektor usahanya mempunyai kontribusi dominan terhadap penerimaan negara setiap tahunnya. Banyaknya klasifikasi lapangan usaha (KLU) atau jenis usaha WP, yang harus diawasi oleh pegawai pajak serta belum adanya sistem yang dapat membantu dalam mengetahui sektor yang potensial merupakan salah satu faktor yang mempengaruhi penerimaan pajak tidak mencapai target yang telah ditetapkan.

Dengan adanya keterbatasan SDM membuat pengawasan dan pemeriksaan, khususnya terhadap WP Badan, menjadi kurang optimal. Pengawasan dan pemeriksaan yang kurang optimal dapat berakibat tidak tertagihnya penerimaan pajak negara.

Pemeriksaan adalah pembuktian yang dilakukan secara obyektif dan profesional dengan cara mengumpulkan dan mengolah data, pernyataan, dan/atau ketetapan yang dilakukan oleh pemeriksa pajak berdasarkan standar pemeriksaan. Tujuan utama dari dilaksanakannya pemeriksaan pajak adalah untuk menumbuhkan perilaku kepatuhan WP dalam memenuhi kewajiban perpajakan (*tax compliance*) yaitu dengan jalan penegakkan hukum (*law enforcement*). Sedangkan pengawasan dilakukan oleh *Account Representative* (AR) berdasarkan WP yang masuk dalam wilayah pengawasan AR tersebut. Satu AR harus mengawasi ratusan WP. Sehingga untuk memudahkan pengawasan, AR dapat membuat daftar prioritas WP yang akan di tindaklanjuti.

Saat ini DJP telah melakukan reformasi perpajakan dengan melakukan digitalisasi pada sistem administrasi perpajakan. Dengan adanya ketersedian data dan teknologi yang ada saat ini pengelompokan KLU dapat dilakukan dengan metode *data mining*. *Data mining* merupakan serangkaian proses yang bertujuan untuk menemukan informasi dari sekumpulan data, yang dapat digunakan dalam proses pendukung keputusan. *Data mining* juga dapat digunakan untuk memproses data atau *records* dalam jumlah yang besar [6].

Metode *clustering* adalah salah satu metode yang dapat digunakan dalam proses *data mining*. Metode *clustering* dapat berupa *hierarchical* atau *partitional*. Metode *hierarchical* menentukan cluster secara berurutan menggunakan *cluster* yang telah ditentukan sebelumnya, sedangkan metode *partitional* menentukan semua *cluster* sekaligus. Pada dasarnya c*lustering* bertujuan mengelompokkan data berdasarkan kemiripan antar data, sehingga data dengan kemiripan paling dekat berada dalam satu kelompok sedangkan data yang berbeda dalam kelompok lainnya. K-Means dan K-Medoids merupakan metode *partitional* paling umum digunakan yang memiliki algoritma yang masih saling berkaitan.

K-Means menggunakan *mean* sebagai titik pusat *cluster* sedangkan K-Medoids memilih data point sebagai pusatnya. Dalam penelitian ini akan dijelaskan

perbandingan metode *clustering* menggunakan *Hierarchical Clustering* dan *Partitional Clustering* yang meliputi K-Means dan K-Medoids.

Berdasarkan penelitian sebelumnya yang dilakukan di KPP Pratama Sukoharjo, data penerimaan pajak per KLU dapat dibagi menjadi empat area, yaitu daerah prima, daerah potensial, daerah berkembang, dan daerah terbelakang [4]. Hasil penelitian dituangkan ke dalam matriks potensi berdasarkan 4 area tersebut. Sehingga dapat dilakukan strategi-strategi khusus untuk mempertahankan pertumbuhan penerimaan pajak terhadap KLU yang berada di daerah prima. Dan meningkatkan strategi terhadap KLU yang berada di daerah terbelakang. Dapat disimpulkan bahwa data penerimaan per kelompok KLU dapat dimanfaatkan dalam pengelompokan potensi pajak.

Penelitian lainnya membahas penggunaan algoritma K-Means. Dalam penelitian yang dilakukan oleh Linda Maulida, algoritma K-Means digunakan dalam pengelompokan kunjungan wisatawan ke objek unggulan di provinsi DKI Jakarta [10]. Penelitian ini mengelompokkan berdasarkan tingkat kunjungan, yaitu tinggi, sedang, dan rendah. Begitu juga dengan penelitian yang dilakukan oleh Miftachur Robani dan Achmad Widodo, algoritma K-Means digunakan dalam pengelompokan ayat Al Quran [8]. Penelitian dilakukan dengan terlebih dahulu melakukan prapemrosesan pada data teks dan melakukan pembobotan terhadap setiap kata yang dihasilkan.

Karena latar belakang tersebut, peneliti mencoba menggunakan algoritma K-Means untuk mengelompokkan Klasifikasi Lapangan Usaha (KLU) atau jenis usaha WP yang memiliki potensi bagi penerimaan pajak dari data yang dimiliki oleh DJP. Kelebihan K-Means adalah mudah untuk diadaptasi dan waktu yang dibutuhkan untuk menjalankan relatif cepat. Selain itu untuk dataset yang besar, algoritma ini relatif sederhana, efisien dan mudah dipahami. Suatu dataset dikatakan besar ketika berukuran besar dan memiliki karakteristik seperti jenis data sangat beragam dan laju pertumbuhan maupun frekuensi perubahan sangat tinggi.

1.2 Rumusan Masalah

Pajak merupakan sumber pendapatan negara terbesar di Indonesia. Untuk mencapai target penerimaan pajak, DJP perlu melakukan penggalian potensi pajak

yang ada di wilayah kerjanya dengan melakukan kegiatan pengawasan dan pemeriksaan. Namun, dengan adanya keterbatasan SDM membuat pengawasan dan pemeriksaan, khususnya terhadap WP Badan, menjadi kurang optimal. Berangkat dari hal tersebut adapun rumusan masalah dari penelitian ini adalah bagaimana memetakan WP potensial berdasarkan risiko tingkat ketidakpatuhan dan dampak fiskal per sektor usaha atau Klasifikasi Lapangan Usaha (KLU) dapat di selesaikan dengan menggunakan metode K-Means *clustering*. Dan apakah dengan menggunakan metode K-Means dapat diperoleh hasil pengelompokan yang optimal.

1.3 Tujuan

Tujuan dari penelitian ini adalah melakukan implementasi *clustering* KLU potensial sesuai risiko tingkat ketidakpatuhan serta dampak fiskalnya dengan memanfaatkan algoritma K-Means dan menguji tingkat kekompakan *cluster* yang dihasilkan. Informasi dari hasil penelitian ini dapat dipergunakan untuk mengetahui KLU yang terindikasi tidak memenuhi kewajiban perpajakannya sesuai undang-undang dan peraturan yang berlaku.

1.4 Batasan Masalah

Dalam perancangan penelitian ini, ada beberapa hal yang dibatasi oleh penulis. Di antaranya adalah:

1. Data yang diambil merupakan data WP per KLU tahun 2016 sampai dengan 2019, yang diperoleh dari Direktorat Data dan Informasi Perpajakan, Direktorat Jenderal Pajak, Kementrian Keuangan Republik Indonesia.
2. Data yang digunakan adalah data klasifikasi lapangan usaha (KLU) atau data jenis usaha wajib pajak di Kantor Wilayah DJP Wajib Pajak Besar dan Kantor Wilayah DJP Jakarta Khusus.

1.5 Kontribusi

Hasil dan informasi yang diperoleh dari penelitian ini dapat digunakan untuk mengelompokkan WP per kelompok KLU yang potensial bagi penerimaan negara. Pengelompokan ini diharapkan mampu dimanfaatkan sebagai informasi

yang mendukung dalam pengambilan keputusan untuk meningkatkan kinerja pemeriksaan dan pengawasan di Kantor Pelayanan Pajak Pratama sebagai perpanjangan dari kantor pusat DJP.

BAB 2
KAJIAN PUSTAKA

Pada bab ini membahas tentang landasan teori yang mendasari penulis dalam pengerjaan penelitian ini.

2.1 Kajian Penelitian Terkait

Penggunaan algoritma K-Means telah dibahas di beberapa penelitian. Salah satunya penelitian dengan judul "Penerapan Datamining dalam Mengelompokkan Kunjungan Wisatawan ke Objek Wisata Unggulan di Provinsi DKI Jakarta dengan K-Means" [10]. Penelitian ini dilakukan di provinsi DKI Jakarta. Tujuan penelitian ini adalah untuk mengelompokkan dan mencari tahu potensi yang paling rendah yang dimiliki oleh objek wisata unggulan di provinsi DKI Jakarta dalam kunjungan wisatawan ke Indonesia, dalam kasus ini khususnya di provinsi DKI Jakarta. K-Means digunakan untuk pengelompokan jumlah wisatawan asing yang datang ke provinsi DKI Jakarta. Data penelitian menggunakan data jumlah wisatawan yang berkunjung dalam rentang waktu tahun 2007-2013. Dari data tersebut dilakukan pengelompokan menjadi 3 *cluster*, yakni jumlah kunjungan tinggi, sedang, dan rendah. Hasil *clustering* menunjukkan terdapat lima objek wisata unggulan yang berada di *cluster* dengan kategori jumlah kunjungan wisatawan rendah. Hal ini menjadi catatan bagi pemerintah provinsi DKI Jakarta untuk memperbaiki sarana dan prasarana objek wisata dan lebih gencar dalam pengenalan objek wisata. Diharapkan langkah yang diambil oleh pemerintah setempat dapat meningkatkan jumlah kunjungan sehingga berdampak pada peningkatan devisa negara.

Penelitian lainnya dengan judul "Algoritma K-Means Clustering Untuk Pengelompokan Ayat Al Quran Pada Terjemahan Bahasa Indonesia" [8]. Penelitian ini membahas tentang pengelompokan ayat Al Quran yang memiliki kemiripan dengan tujuan memberikan kemudahan bagi pengguna untuk menemukan suatu tema tertentu di dalam Al Quran. Penelitian ini melakukan pengelompokan menggunakan metode *clustering* K-Means. Penelitian ini dilakukan dengan 4

tahapan, yaitu prapemrosesan data teks, pembobotan dengan TFIDF, pengelompokan dengan K-Means, dan pelabelan dengan kata kunci. Penulis memilih menggunakan metode K-Means karena sederhana, efisien, dan mudah dipahami. Penulis menggunakan *silhouette index* untuk pengujian hasil *cluster* dan menghasilkan nilai positif sebesar 0.336 yang artinya data telah berada di kelompok yang tepat. Pengujian juga menunjukkan bahwa hasil pengujian silhouette akan berbanding lurus dengan jumlah *cluster* dan berbanding terbalik dengan jumlah dimensi data.

2.2 Klasifikasi Lapangan Usaha (KLU)

Klasifikasi Lapangan Usaha atau KLU Pajak merupakan kode khusus yang diterbitkan oleh DJP. Pengklasifikasian dilakukan dengan mengacu pada Klasifikasi Baku Lapangan Usaha Indonesia (KBLI) Badan Pusat Statistik. KLU diterbitkan untuk memudahkan dalam melakukan klasifikasi WP ke dalam jenis badan usaha berdasarkan kategori tertentu.

Penyusunan Klasifikasi Lapangan Usaha Wajib Pajak (KLU WP) diatur dalam Keputusan Dirjen Pajak Nomor KEP - 233/PJ/2012. Dalam surat keputusan tersebut, KLU disusun berdasarkan kategori, golongan pokok, golongan sub golongan dan kelompok kegiatan ekonomi. Pada umumnya kode KLU dipergunakan sebagai:

a. Penatausahaan Wajib Pajak, seperti data kelompok kegiatan ekonomi Wajib Pajak dalam master file Wajib Pajak serta kelompok kegiatan ekonomi pada SPT Pajak Penghasilan (PPh);
b. Dasar penyusunan norma penghitungan penghasilan netto;
c. Keperluan khusus lainnya, seperti evaluasi penerimaan pajak sektoral, *mapping* potensi pajak sektoral, dan lain-lain.

Penyusunan kode KLU menggunakan kode angka 5-digit dan 1 kode alfabet yang disebut kategori. Kode alfabet dicantumkan untuk mempermudah dalam penyusunan sektor, namun sebenarnya bukan termasuk bagian dari kode KLU. Struktur KLU secara lebih terperinci adalah sebagai berikut:

1. Kategori, seluruh kegiatan ekonomi di Indonesia digolongkan menjadi 21 jenis kategori sebagai berikut:

a. Kategori A yaitu Pertanian, Kehutanan dan Perikanan
b. Kategori B yaitu Pertambangan dan Penggalian
c. Kategori C yaitu Industri Pengolahan
d. Kategori D yaitu Pengadaan Listrik, Gas, Uap/Air Panas dan Udara Dingin
e. Kategori E yaitu Pengadaan Air,Pengelolaan Sampah dan Daur Ulang, Pembuangan dan Pembersihan Limbah dan Sampah
f. Kategori F yaitu Konstruksi
g. Kategori G yaitu Perdagangan Besar dan Eceran; Reparasi dan Perawatan Mobil dan Sepeda Motor
h. Kategori H yaitu Transportasi dan Pergudangan
i. Kategori I yaitu Penyediaan Akomodasi dan Penyediaan Makan Minum
j. Kategori J yaitu Informasi dan Komunikasi
k. Kategori K yaitu Jasa Keuangan dan Asuransi
l. Kategori L yaitu Real Estate
m. Kategori M yaitu Jasa Profesional, Ilmiah dan Teknis
n. Kategori N yaitu Jasa Persewaan, Ketenagakerjaan, Agen Perjalanan dan Penunjang Usaha Lainnya
o. Kategori O yaitu Administrasi Pemerintahan, dan Jaminan Sosial Wajib
p. Kategori P yaitu Jasa Pendidikan
q. Kategori Q yaitu Jasa Kesehatan dan Kegiatan Sosial
r. Kategori R yaitu Kebudayaan, Hiburan dan Rekreasi
s. Kategori S yaitu Kegiatan Jasa Lainnya
t. Kategori T yaitu Jasa Perorangan Yang Melayani Rumah Tangga, Kegiatan Yang menghasilkan Barang dan Jasa
u. Kategori U yaitu Kegiatan Badan Internasional dan Badan Ekstra Internasional Lainnya

2. Golongan pokok, merupakan uraian lebih lanjut dari kategori. Setiap kategori diuraikan menjadi satu atau beberapa golongan pokok (sebanyak-banyaknya 5 golongan pokok, kecuali industri pengolahan) menurut sifat dari masing-masing golongan pokok. Setiap golongan pokok diberi kode 2-digit angka.
3. Golongan, merupakan uraian lebih lanjut dari golongan pokok. Kode golongan terdiri dari 3-digit angka, yaitu 2-digit angka pertama menunjukkan Golongan

Pokok yang berkaitan dan 1-digit angka terakhir menunjukkan kegiatan ekonomi dari setiap golongan bersangkutan. Setiap golongan pokok dapat diuraikan menjadi sebanyak-banyaknya 9 golongan.

4. Subgolongan, merupakan uraian lebih lanjut dari golongan. Kode subgolongan terdiri dari 4-digit, yaitu kode 3-digit angka pertama menunjukkan golongan yang berkaitan, dan 1-digit angka terakhir yang menunjukkan kegiatan ekonomi dari subgolongan bersangkutan. Setiap golongan dapat diuraikan lebih lanjut menjadi sebanyak-banyaknya 9 subgolongan.
5. Kelompok, dimaksudkan untuk memilah lebih lanjut kegiatan yang tercakup dalam suatu subgolongan, menjadi beberapa kegiatan yang lebih homogen.

Contoh penyusunan kode KLU secara lengkap dapat dilihat pada gambar 2-1.

GP	G	SG	KEL	URAIAN KLASIFIKASI LAPANGAN USAHA
KATEGORI A : PERTANIAN, KEHUTANAN DAN PERIKANAN				
01				**PERTANIAN TANAMAN, PETERNAKAN, PERBURUAN DAN KEGIATAN YBDI**
	011			**PERTANIAN TANAMAN SEMUSIM**
		0111		**PERTANIAN TANAMAN SEREALIA (BUKAN PADI), KACANG-KACANGAN DAN BIJI-BIJIAN PENGHASIL MINYAK**
			01111	**PERTANIAN TANAMAN JAGUNG**
				Kelompok ini mencakup usaha pertanian mulai dari kegiatan pengolahan lahan, penanaman, pemeliharaan, dan juga pemanenan dan pasca panen jika menjadi satu kesatuan kegiatan tanaman serealia jagung.
			01112	**PERTANIAN TANAMAN GANDUM**
				Kelompok ini mencakup usaha pertanian mulai dari kegiatan pengolahan lahan, penanaman, pemeliharaan, dan juga pemanenan dan pasca panen jika menjadi satu kesatuan kegiatan tanaman serealia gandum, seperti sorgum/cantel, gandum (wheat/oats), jelai (barley), gandum hitam (rye), jawawut (millet) dan sejenisnya.

Gambar 2-1 Penyusunan Kode KLU

2.3 *Data Mining*

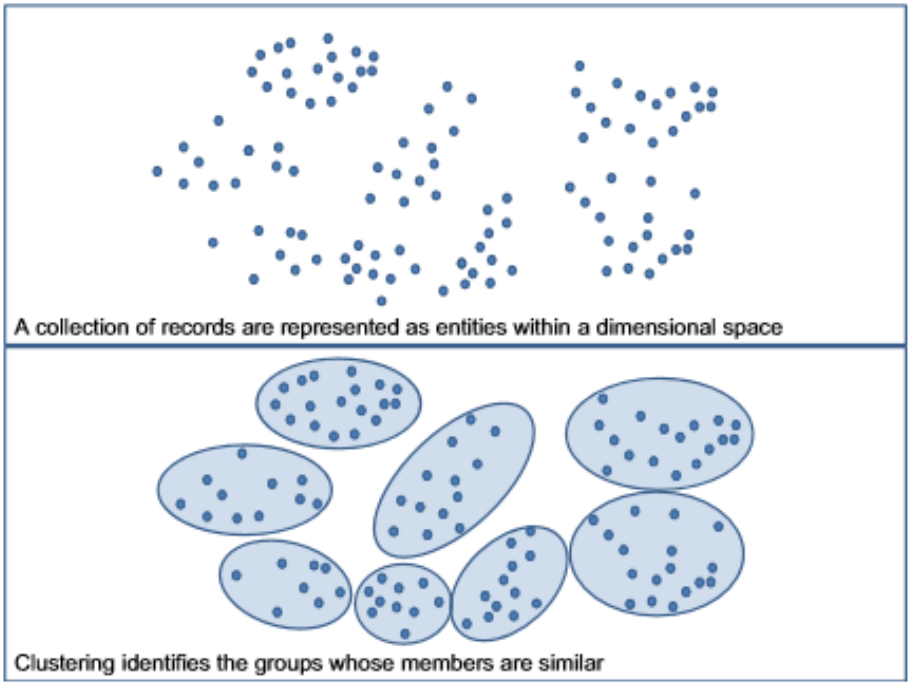
Teknologi pengumpulan dan penyimpanan informasi yang terus berkembang telah membuat lonjakan data dalam jumlah besar di sebagian besar bidang aplikasi, baik dalam kegiatan bisnis, komunitas ilmiah dan medis, maupun administrasi publik. Rangkaian kegiatan yang dilakukan untuk menganalisis *database* dalam jumlah besar ini dapat menggunakan berbagai cara, salah satunya

adalah *data mining*, yang bertujuan untuk mengekstraksi pengetahuan yang dapat berguna dalam mendukung pengambilan keputusan.

Data Mining adalah serangkaian proses yang dilakukan secara berulang untuk menganalisis *database* dalam jumlah besar, dengan tujuan mengekstraksi informasi dan pengetahuan baru yang berpotensi bermanfaat dalam pengambilan keputusan dan penyelesaian masalah [1]. *Data Mining* dibagi menjadi 6 fungsi utama, yaitu:

1. *Clustering* dan Segmentasi

   Clustering merupakan pengelompokan entitas dari suatu kumpulan *record* dalam jumlah besar dan membagi *record* tersebut menjadi kelompok-kelompok kecil yang memiliki kemiripan antar satu entitas dalam satu *cluster* dan memiliki ketidakmiripan dengan entitas dalam *cluster* lain. Contoh *clustering* ditampilkan pada Gambar 2-2.



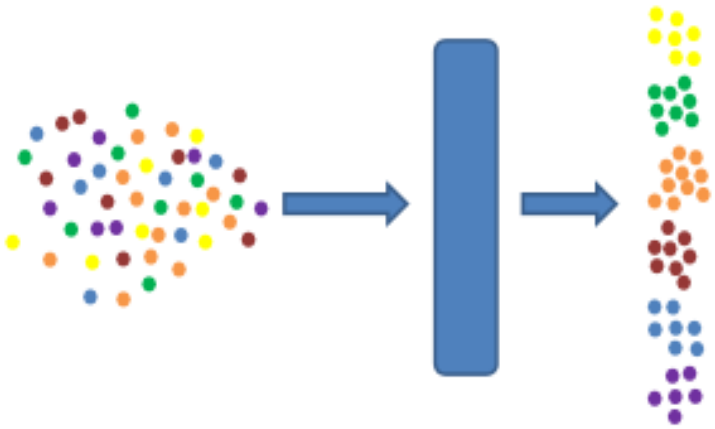
Gambar 2-2 Contoh *Clustering*

Clustering dapat digunakan ketika tidak yakin dengan apa yang dicari namun tetap ingin melakukan segmentasi. Misalnya, ketika ingin mengevaluasi data kesehatan berdasarkan penyakit tertentu dan variabel lain untuk melihat apakah terdapat korelasi yang dapat disimpulkan melalui proses pengelompokan

tersebut. *Clustering* dapat digunakan bersamaan dengan teknik *data mining* lainnya dengan tujuan mengidentifikasi dan mengekplorasi lebih lanjut sebuah informasi.

2. Klasifikasi

   Klasifikasi adalah proses penggolongan entitas ke dalam kelas yang telah ditentukan. Kelas-kelas tersebut dapat ditentukan oleh analis menggunakan atribut yang telah dipilih, atau dapat juga berdasarkan hasil pemodelan menggunakan *clustering*. Nilai-nilai variabel dependen yang telah diidentifikasi dapat digunakan untuk klasifikasi. Contoh klasifikasi dapat dilihat pada Gambar 2-3. Langkah awal adalah mendeskripsikan data ke dalam sejumlah atribut berjumlah n. Kemudian tingkat kemiripan atau kedekatan dihitung dari data baru terhadap seluruh data pelatihan. Selanjutnya kelas terdekat yang memiliki jumlah bobot paling besar digunakan untuk menentukan hasil klasifikasi.



Gambar 2-3 Klasifikasi

Klasifikasi dapat digunakan untuk mengevaluasi perusahaan BUMN menjadi 3 kelas investasi, yaitu baik, sedang, dan buruk. Penggunaan klasifikasi lainnya banyak digunakan dalam menentukan transaksi kartu kredit, dimana hasil klasifikasi membagi transaksi menjadi transaksi ilegal atau bukan.

3. Estimasi

   Estimasi adalah proses menetapkan beberapa nilai numerik yang dinilai terus menerus untuk suatu objek. Misalnya, penilaian risiko kredit tidak selalu berbentuk pertanyaan ya / tidak, tetapi penilaian yang melihat kecenderungan gagal bayar. Perbedaan estimasi dengan klasifikasi yaitu estimasi memiliki

variabel target yang lebih mengarah pada hasil numerik, sedangkan klasifikasi mempunyai variabel target yang mengarah pada kategori.

Nilai estimasi dilakukan dengan memberikan nilai numerik pada variabel kontinu, yang kemudian hasilnya diberi peringkat berdasarkan skor.

Misalnya, nilai indeks prestasi seorang mahasiswa saat mengikuti program sarjana dapat digunakan untuk mencari nilai estimasi indeks prestasi mahasiswa tersebut saat mengikuti program pasca sarjana.

4. Prediksi

   Prediksi adalah upaya untuk mengklasifikasikan objek berdasarkan perilaku yang diharapkan di masa depan. Prediksi dapat dilakukan dengan menggunakan data historis dari proses klasifikasi dan estimasi. Pada klasifikasi dibentuk sebuah model yang disebut juga dengan *training*. Model tersebut kemudian diterapkan pada data baru untuk memprediksi perilaku di masa depan. Dalam melakukan prediksi disarankan untuk memakai kumpulan data yang berbeda untuk *training* dan tes.

5. Asosiasi

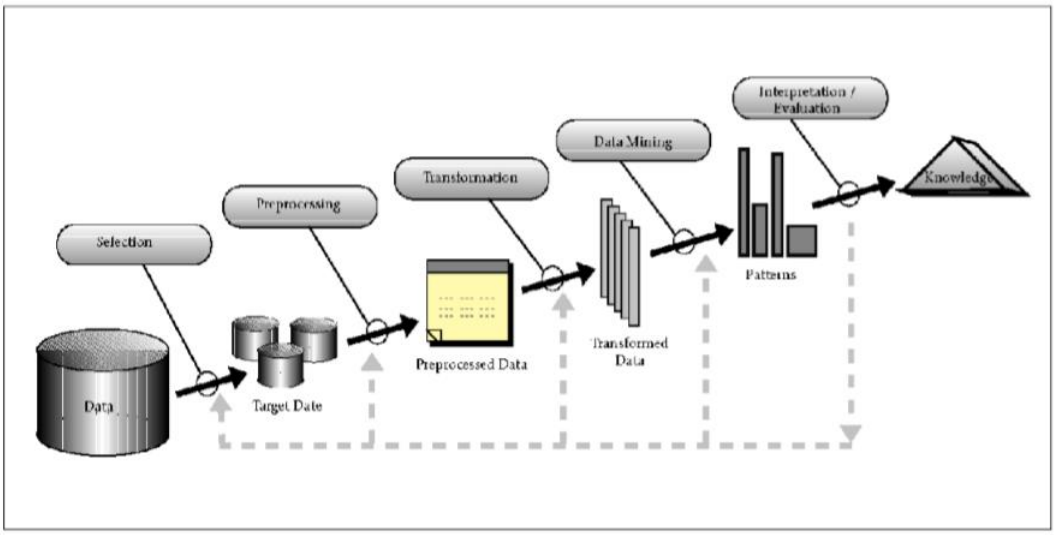
   Asosiasi adalah proses mengevaluasi hubungan atau asosiasi antara elemen data yang menunjukkan afinitas antar objek. Pengelompokan berdasarkan afinitas ini dapat digunakan untuk menentukan kemungkinan bahwa orang yang membeli satu produk akan cenderung untuk mencoba produk yang berbeda. Jenis analisis ini berguna untuk pemasaran ketika sebuah perusahaan berusaha *cross-selling* atau menjual produk tambahan pada pelanggan.

6. Deskripsi

   Deskripsi merupakan proses yang menggambarkan pola dan kecenderungan yang terdapat dalam data secara sederhana dari proses *data mining*. Mampu menggambarkan perilaku atau aturan bisnis adalah langkah lain yang dapat mengidentifikasi pengetahuan, menuangkannya, dan kemudian mengevaluasi tindakan yang dapat diambil. Deskripsi pengetahuan yang ditemukan juga dapat dimasukkan ke dalam metadata yang terkait dengan kumpulan data tersebut.

2.4 Knowledge Discovery in Database (KDD)

KDD dan *data mining* merupakan dua istilah yang berbeda. KDD mengacu pada keseluruhan proses dalam menemukan pengetahuan yang bermanfaat dari suatu kumpulan data. Sedangkan *data mining* merupakan salah satu proses dalam KDD yang berkaitan dengan menemukan pola-pola baru dari kumpulan data dalam database untuk mengekstraksi pengetahuan yang berguna.



Gambar 2-4 Alur KDD

Berdasarkan gambar 2-4, alur KDD terdiri dari metode yang berulang sebagai berikut [1]:

1. *Selection*. Pada tahap ini dilakukan pemilihan data yang relevan dari database untuk dianalisis.
2. *Preprocessing* / *Data Cleaning*. Di tahap ini data yang tidak konsisten maupun mengalami duplikasi dihapus, dan data yang rusak diperbaiki. Pada tahap ini juga dilakukan penggabungan beberapa sumber data.
3. *Transformation*. Proses ini mengubah data menjadi bentuk yang sesuai kebutuhan dalam proses *data mining*.
4. *Data mining*. Proses ini dilakukan dengan memilih algoritma *data mining* yang sesuai dengan pola dalam data. Di tahap ini juga dilakukan ekstraksi pola suatu data.
5. *Interpretation* / Evaluasi: Proses ini untuk menafsirkan pola menjadi pengetahuan dengan menghapus pola yang tidak relevan atau berulang,

Kemudian menerjemahkan pola yang berguna ke dalam istilah yang dapat dipahami manusia dalam bentuk visualisasi.

2.5 *Clustering*

Pada dasarnya *clustering* adalah teknik umum untuk analisis data statistik, yang digunakan di banyak bidang, termasuk *machine learning*, *data mining*, pengenalan pola, analisis gambar, dan bioinformatika. *Clustering* adalah proses pengelompokan objek-objek serupa ke dalam *cluster* yang berbeda, atau pembagian suatu kumpulan data ke dalam *cluster* menurut ukuran jarak yang ditentukan. Dimana dalam sebuah *cluster* terdapat kumpulan objek yang memiliki kemiripan antar satu entitas dengan yang lainnya dan ketidakmiripan dengan entitas di *cluster* lain [5].

Clustering dapat juga disebut sebagai proses di mana sekelompok pola yang tidak berlabel dibagi menjadi beberapa dataset sehingga pola yang sama ditempatkan ke dalam *cluster* yang sama, dan pola yang berbeda ditempatkan ke dalam *cluster* yang berbeda. Terdapat dua tujuan dari algoritma *clustering*, yaitu menentukan *cluster* yang baik dan melakukannya secara efisien [17].

Clustering adalah sebuah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang maksimum dan data antar *cluster* memiliki kemiripan yang minimum [23].

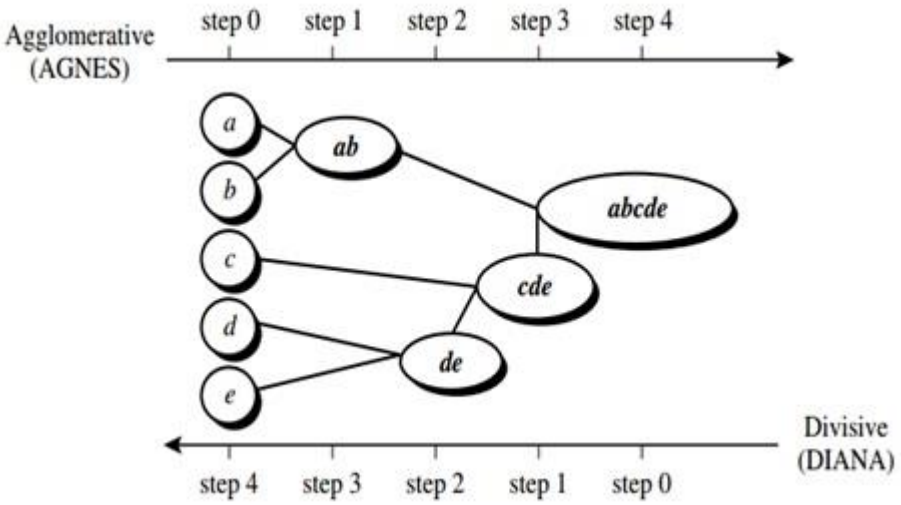
Istilah *cluster* mengacu pada subkelompok homogen yang ada dalam suatu populasi. Karena itu, teknik pengelompokan ditujukan untuk mensegmentasi populasi yang heterogen ke dalam sejumlah subkelompok yang didapatkan dari pengamatan yang memiliki karakteristik serupa. Berbeda dengan klasifikasi, dalam *clustering* tidak ada kelas yang ditentukan sebelumnya [3]. *Clustering* sering disebut sebagai *Unsupervised Classification*.

Metode *clustering* dapat berupa *hierarchical* atau *partitional*. Metode *hierarchical* menentukan cluster secara berurutan menggunakan *cluster* yang telah ditentukan sebelumnya, sedangkan metode *partitional* menentukan semua *cluster* sekaligus [5]. Adapun penjelasan kedua metode tersebut adalah sebagai berikut:

2.5.1 *Hierarchical Clustering*

Pada *hierarchical clustering* data dikelompokkan dengan memilih ukuran jarak. Terdapat dua ukuran yang digunakan dalam menentukan *cluster*, *manhattan* dan *euclidean*. Jarak *manhattan* adalah sama dengan jumlah jarak absolut untuk setiap objek. Sedangkan jarak *euclidean*, dihitung dengan mencari kuadrat jarak antara setiap variabel, menjumlahkan kuadrat, dan menghitung akar kuadrat dari jumlah tersebut.

Dalam *hierarchical clustering* setiap objek diperlakukan sebagai *cluster* yang terpisah. Penggabungan dilakukan terhadap dua atau lebih objek yang saling berdekatan melalui suatu bagan yang berupa hirarki. Kemudian dilanjutkan dengan objek lain dengan jarak terdekat kedua, proses iterasi ini dilakukan sampai seluruh objek tergabung dalam suatu cluster. Hasil akhir yang didapatkan berupa pohon hirarki yang memiliki tingkatan antar objek, dari yang memiliki kemiripan maksimal sampai dengan kemiripan minimal seperti dapat dilihat pada contoh gambar 2-5.



Gambar 2-5 Hierarchical clustering

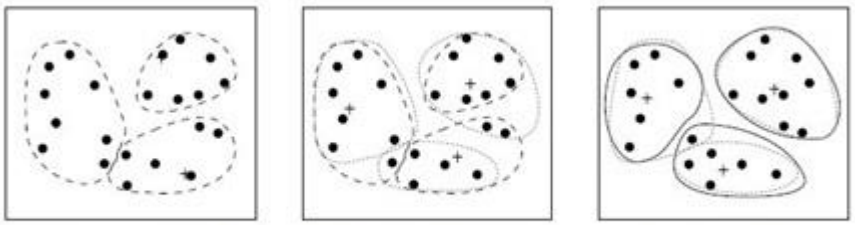
Dalam *hierarchical clustering* langkah yang harus dilakukan adalah sebagai berikut:

1. Mengidentifikasi dua *cluster* yang paling dekat,
2. Menggabungkan dua *cluster* yang paling mirip.
3. Menghitung jarak antar *cluster*

4. Melanjutkan proses iteratif sampai semua *cluster* terhubung

2.5.2 *Partitional Clustering*

Partitional clustering dilakukan berdasarkan penentuan jumlah awal *cluster*, kemudian secara iteratif menempatkan ulang suatu objek di antara *cluster* tanpa menggunakan pohon hirarki. Metode *partitional* menentukan semua *cluster* secara sekaligus. Dalam metode ini, sebelum proses pengelompokan dilakukan, jumlah *cluster* yang akan dibentuk ditentukan terlebih dahulu. Setiap *cluster* memiliki titik pusat *cluster* (centroid) dan secara umum metode ini memiliki fungsi tujuan yaitu meminimumkan jarak (*dissimilarity*) dari seluruh data ke pusat *cluster* masing-masing seperti terlihat pada gambar 2-6. Metode *partitional* yang paling umum digunakan adalah K-Means dan K-Medoids.



Gambar 2-6 *Partitional Clustering*

2.6 K-Means *Clustering*

Metode K-Means merupakan salah satu metode *partitional clustering* atau *non-hierarchical clustering*. K-Means merupakan salah satu algoritma yang berbasis titik pusat (centroid). Metode ini secara iteratif membagi kumpulan objek ke dalam *cluster* berbeda yang telah ditentukan sebelumnya di mana setiap objek hanya dimiliki oleh satu *cluster* dan tidak tumpang tindih. K-Means membagi objek memiliki tingkat kemiripan maksimal dalam satu *cluster* dan memiliki kemiripan minimal dengan objek dari *cluster* lainnya. Metode ini mengelompokan objek ke sebuah *cluster* sedemikian rupa sehingga jumlah jarak kuadrat antara objek dan centroid *cluster* adalah minimum. Semakin sedikit variasi yang dimiliki dalam

sebuah *cluster*, semakin homogen (mirip) objek yang berada dalam *cluster* yang sama.

Terdapat tiga parameter yang dibutuhkan dalam penerapan algoritma K-Means, yaitu jumlah cluster K, inisialisasi *cluster*, dan ukuran jarak. Ketiga parameter tersebut ditentukan oleh pengguna. Penjelasan dari tiga parameter tersebut adalah sebagai berikut:

1. Jumlah *Cluster* K

   Dalam melakukan pengelompokkan dengan algoritma K-Means, perlu ditentukan terlebih dahulu jumlah *cluster* K. Penentuan jumlah *cluster* K dapat dilakukan melalui pendekatan metode hirarki atau ditentukan langsung oleh pengguna. Hal ini dikarenakan tidak terdapat aturan khusus dalam menentukan jumlah *cluster* K.

2. Inisialisasi *Cluster*

   Karena *clustering* K-Means bertujuan untuk menghasilkan *cluster* konvergen yang optimal dan keanggotaan *cluster* berdasarkan jarak dari centroid melalui iterasi yang berurutan, maka semakin optimal penempatan centroid awal, semakin sedikit iterasi dari algoritma *clustering* K-Means yang diperlukan untuk mencapai konvergensi. Terdapat beberapa cara dalam memilih inisialisasi centroid metode K-Means, diantaranya adalah sebagai berikut:

   1) Menentukan pusat *cluster* awal dengan menggunakan interval dari jumlah setiap objek [18].
   2) Menentukan pusat *cluster* awal dengan menggunakan pengelompokan hirarki (metode Ward) [19].

3. Jarak

   Ukuran jarak yang umum digunakan dalam algoritma K-Means adalah jarak *euclidean*. Jarak *euclidean* adalah jarak dua titik dalam bidang atau ruang 3 dimensi untuk mengukur panjang bidang yang menghubungkan kedua titik. Ukuran ini kemudian digunakan untuk menempatkan sebuah objek ke dalam *cluster* yang memiliki jarak terdekat dengan centroid. Semakin besar jaraknya maka semakin tinggi pula tingkat perbedaanya. Untuk mengukur jarak menggunakan *euclidean* digunakan Persamaan 1:

$$d(x_i, c_i) = \sqrt{(x_i - c_i)^2} \qquad (1)$$

Keterangan:

d = Jarak titik x ke c

c_i = Centroid ke- i

x_i = Objek ke- i

Langkah yang harus dilakukan dalam pembentukan *cluster* menggunakan algoritma K-Means adalah sebagai berikut:

a. Menentukan jumlah *cluster* K

b. Inisialisasi centroid (titik pusat *cluster*) awal dengan terlebih dahulu mengacak dataset dan kemudian memilih K centroid secara acak sebanyak K *cluster*. Kemudian untuk menghitung centroid *cluster* ke-i berikutnya, digunakan persamaan 2:

$$v = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2)$$

Keterangan:

v = centroid pada *cluster*

x_i = objek ke-i

n = jumlah objek dalam satu *cluster*

c. Menghitung jumlah kuadrat jarak antara objek ke semua centroid. Menghitung jarak dilakukan menggunakan rumus *euclidean* seperti pada persamaan 3:

$$\mathrm{d(x,y)} = \|\mathrm{x} - \mathrm{y}\| = \sqrt{\sum_{i=1}^{n}(\mathrm{x}_i - \mathrm{y}_i)^2} \quad \mathrm{i} = 1,2,3,\ldots,\mathrm{n} \qquad (3)$$

Keterangan:

x = objek *x* ke-i

y = objek *y* ke-i

n = jumlah objek

d. Setiap objek ditempatkan ke dalam *cluster* terdekat (centroid)

e. Lakukan proses iterasi, hitung centroid untuk *cluster* dengan mengambil rata-rata dari semua objek yang dimiliki setiap *cluster* sebagai posisi centroid baru
f. Ulangi langkah c sampai tidak ada perubahan pada centroid

Pengelompokan data menggunakan K-Means bertujuan untuk memaksimalkan kemiripan objek dalam satu *cluster* dan meminimalkan kemiripan objek antar *cluster*. Fungsi jarak digunakan untuk mengukur tingkat kemiripan. Sehingga tingkat kemiripan maksimal didapatkan dengan mencari jarak terpendek antar objek dari semua centroid.

Proses iteratif menempatkan kembali objek ke dalam cluster untuk meningkatkan variasi nilai dalam tiap *cluster*. Sehingga tidak ada distribusi ulang terhadap objek pada *cluster* mana pun yang terjadi dan proses berhenti [20].

2.7 K-Medoids *Clustering*

Sama halnya dengan K-Means, K-Medoids juga termasuk dalam algoritma *partitional clustering* atau *non-hierarchical clustering*. K-Medoids atau yang dikenal juga dengan sebutan *Partitioning Around Method* (PAM) ini berbeda dari algoritma K-Means dalam hal cara memilih pusat *cluster* (centroid). K-Means menentukan titik pusat *cluster* dengan mencari rata-rata dari suatu populasi, sedangkan K-Medoids menggunakan objek dari suatu populasi sebagai titik pusat *cluster* yang disebut dengan medoid.

Langkah yang harus dilakukan dalam pembentukan *cluster* menggunakan algoritma K-Medoids adalah sebagai berikut:

a. Menentukan jumlah *cluster* K
b. Memilih secara acak k objek sebagai medoid (titik pusat *cluster*) awal
c. Menghitung jarak antara objek ke semua medoid menggunakan rumus *euclidean* seperti pada persamaan 3 dan menempatkan objek ke dalam *cluster* dengan jarak medoid terdekat
d. Memilih secara acak objek pada masing-masing *cluster* sebagai medoid baru
e. Menghitung total simpangan (S) dengan persamaan 4:

$$S = b - a \qquad (4)$$

Keterangan:

S = total simpangan

b = jumlah jarak terdekat antara obyek ke medoid baru

a = jumlah jarak terdekat antara obyek ke medoid lama

f. Jika $S < 0$, maka tukar objek dengan data untuk membentuk sekumpulan *k* baru sebagai medoid

g. Ulangi langkah c sampai tidak ada perubahan pada medoid

2.8 *Silhouette Index*

Metode validasi *silhouette index* mengevaluasi penempatan setiap objek dalam setiap *cluster* dengan membandingkan jarak rata-rata entitas di dalam satu *cluster* dan jarak antar entitas dalam *cluster* yang berbeda [21]. Tahapan untuk menghitung *silhouette* adalah:

1. Menghitung rata-rata jarak entitas dengan entitas lain yang berada dalam satu *cluster* dengan persamaan 5

$$a(i) = \frac{1}{|A|-1} \sum j \in_{A, j \neq i} \; d(i,j) \qquad (5)$$

Dimana $a(i)$ rata -rata jarak entitas ke-i dengan semua entitas yang berada di dalam satu *cluster*

2. Menghitung rata-rata jarak entitas dengan entitas lain yang berada pada *cluster* lain dengan persamaan 6

$$d(i,C) = \frac{1}{|A|} \sum j \; \in \; C \; d(i,j) \qquad (6)$$

Kemudian diambil jarak yang paling pendek, dengan persamaan 7

$$b(i) = \min d(i,C) \qquad (7)$$

Dimana $b(i)$ adalah nilai minimum dari jarak rata-rata entitas i dengan semua entitas pada *cluster* lain.

3. Menghitung nilai koefisien *Silhouette* dengan persamaan 8

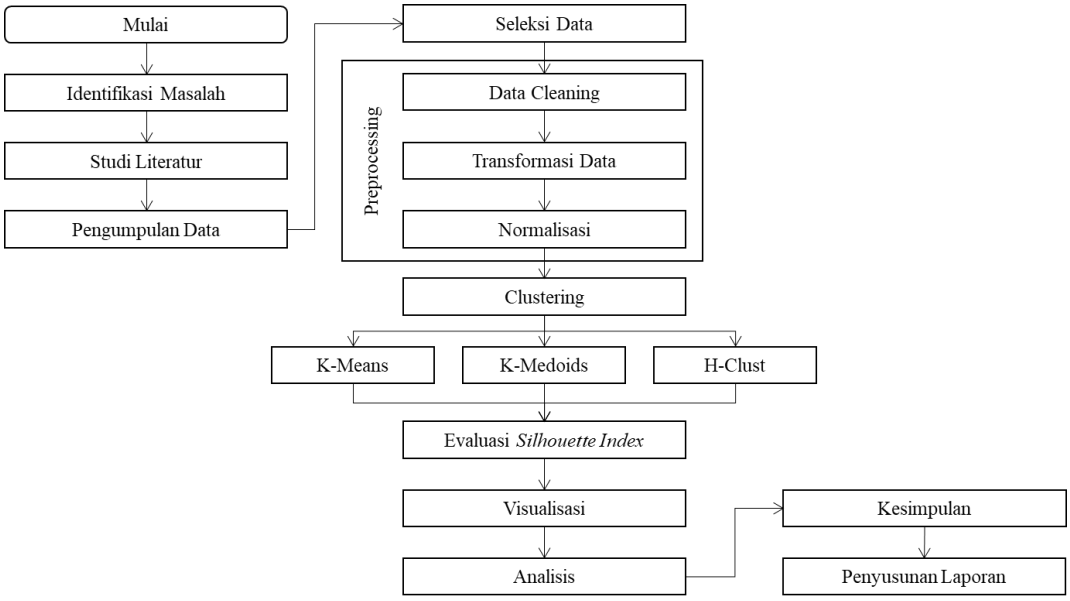
$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \qquad (8)$$

Nilai *silhouette* berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa objek cocok dengan *cluster*-nya dan tidak cocok dengan *cluster* lainnya. Jika sebagian besar objek memiliki nilai tinggi, maka *clustering* dianggap sudah sesuai. Jika banyak titik memiliki nilai rendah atau negatif, maka kemungkinan jumlah *cluster* yang ditentukan terlalu banyak atau terlalu sedikit. Pendekatan ini memberikan representasi tentang seberapa baik kualitas dari setiap *cluster* yang terbentuk. Semakin tinggi nilai rata-rata nya maka akan semakin baik

BAB 3
METODE PENELITIAN

Pada bagian ini akan diuraikan mengenai metode penelitian yang digunakan penulis dalam menyusun laporan penelitian ini. Metode ini juga digunakan sebagai panduan dalam pengerjaan agar terarah dan sistematis.

Alur metodologi yang digunakan dalam pengerjaan penelitian ini dijabarkan dalam Gambar 3-1:



Gambar 3-1 Diagram Alir Penelitian

Adapun penjelasan masing-masing tahapan dalam diagram alir penelitian pada gambar 3-1 adalah:

3.1 Identifikasi Masalah

Identifikasi masalah penelitian dilakukan melalui pengamatan dan survey pada instansi terkait. Perlu dilakukan pemahaman mengenai data dan bisnis dari DJP meliputi kebijakan, proses bisnis, dan informasi-informasi terkait sesuai tugas dan fungsi di DJP, yaitu pemeriksaan, pengawasan, dan penegakan hukum.

Aktifitas yang dilakukan meliputi pemahaman kebijakan dan regulasi di DJP yang telah ditetapkan dalam Peraturan Menteri Keuangan maupun kebijakan lainnya. Kemudian dilanjutkan dengan mempelajari kebijakan internal DJP seperti ketentuan–ketentuan umum perpajakan yang mengatur tentang pelayanan, pengawasan, dan penegakan hukum di kantor pelayanan pajak. Di samping itu juga melakukan wawancara dan diskusi dengan pegawai dari unit yang berkaitan.

3.2 Studi Literatur

Pada tahap ini, penulis melakukan pengumpulan data pustaka yang berhubungan dengan permasalahan yang diangkat. Penulis mencari referensi dan landasan teori yang relevan dari berbagai sumber informasi, seperti buku, jurnal, artikel, serta aturan perundang-undangan. Tujuan dari tahapan ini adalah untuk memahami dan mengevaluasi metode yang akan digunakan dalam penelitian.

3.3 Pengumpulan Data

Proses pengumpulan data dilakukan dari sumber yang relevan untuk mencari solusi atas masalah yang telah teridentifikasi sebelumnya. Pada tahap ini, penulis melakukan pengumpulan data yang lebih detail. Adapun metode pengumpulan data yang dilakukan adalah sebagai berikut:

1. Studi Kepustakaan

   Teknik pengumpulan data dengan mempelajari buku, literatur, serta aturan perundang-undangan yang berhubungan dengan masalah yang ingin dipecahkan.

2. Pengambilan Data

   Data yang digunakan adalah data registrasi, data tanda terima, data Surat Pemberitahuan (SPT), data riwayat pemeriksaan, data penerimaan, dan data pengembalian tahun 2016 s.d. 2019 berdasarkan Klasifikasi Lapangan Usaha (KLU) yang dihimpun oleh kantor pusat DJP. Data akan diperoleh dari Direktorat Data dan Informasi Perpajakan, Direktorat Jenderal Pajak, Kementrian Keuangan Republik Indonesia.

3.4 Pengolahan Data

Langkah selanjutnya adalah pengolahan data yang dilakukan sesuai tahapan *Knowledge Discovery in Database* (KDD), adapun tahapan-tahapannya adalah:

3.4.1 Seleksi Data

Pemilihan atau seleksi data adalah proses di mana data yang relevan diambil dari database untuk digunakan dalam proses data mining. Tahap ini dilakukan sebelum tahap penggalian informasi. Data yang akan digunakan antara lain data registrasi, tanda terima, Surat Pemberitahuan (SPT), riwayat pemeriksaan, penerimaan, dan pengembalian.

Seluruh data diambil dalam rentang waktu 3 tahun (2016 s.d 2019) di dua Kantor Wilayah (Kanwil) DJP yang terdiri dari sembilan Kantor Pelayanan Pajak (KPP). Kanwil yang digunakan yaitu Kanwil Wajib Pajak Besar yang terdiri dari KPP Wajib Pajak Besar Satu (kode: 091), KPP Wajib Pajak Besar Dua (kode: 092), KPP Wajib Pajak Besar Tiga (kode: 051), KPP Wajib Pajak Besar Empat (kode: 093), dan Kanwil Jakarta Khusus, yang terdiri dari KPP Penanaman Modal Asing Satu (kode: 052), KPP Penanaman Modal Asing Dua (kode: 055), KPP Penanaman Modal Asing Tiga (kode: 056), KPP Penanaman Modal Asing Empat (kode: 057), dan KPP Penanaman Modal Asing Lima (kode: 058).

Adapun rincian dan penjelasan dari masing-masing data adalah:

1. Registrasi, berisi data jumlah WP badan.
2. Tanda Terima, berisi data jumlah pelaporan SPT Tahunan dan SPT Masa PPN.
3. SPT, berisi data formulir Surat Pemberitahuan (SPT) baik tahunan maupun masa yang oleh WP digunakan untuk melaporkan pembayaran pajak penghasilan, pengeluaran, dan informasi terkait pajak lainnya sesuai dengan aturan perundang-undangan perpajakan. Data SPT yang akan digunakan antara lain; nilai peredaran usaha, nilai Dasar Pengenaan Pokok (DPP) PPN, nilai perolehan PPN, biaya lain-lain, dan biaya total.
4. Riwayat pemeriksaan, berisi data tentang histori surat perintah pemeriksaan, baik yang pemeriksaan rutin yaitu pemeriksaan yang dilakukan sehubungan dengan pemenuhan hak dan/atau pelaksanaan kewajiban perpajakan, maupun

pemeriksaan khusus yang berdasarkan keterangan lain. Termasuk di dalamnya terdapat data jumlah surat ketetapan pajak yang dikeluarkan oleh DJP.

5. Penerimaan, berisi data pembayaran pajak yang dilakukan WP badan per kelompok KLU pada tahun yang bersangkutan/tahun bayar. Data pembayaran yang diambil berdasarkan data MPN.
6. Pengembalian, berisi data besarnya restitusi/imbalan bunga yang diterima oleh WP badan.

Seluruh tabel yang digunakan kemudian diolah agar dapat digunakan untuk proses *clustering* menggunakan metode K-Means. *Clustering* WP badan per kelompok KLU dibuat dengan mengkategorikan berdasarkan variabel x dan variabel y. Variabel x melihat resiko kemungkinan hilangnya penerimaan pajak dari segi tingkat ketidakpatuhan WP, variabel-variabel yang akan dibentuk pada variabel x adalah sebagai berikut:

1. Variabel laporan SPT tahunan sebagai variabel x1. Variabel ini menilai WP Badan per kelompok KLU dalam melaksanakan kepatuhan formalnya. Risiko yang dilihat adalah risiko pelaksanaan kepatuhan formal berupa pelaporan SPT tahunan PPh badan. Karena seluruh WP wajib melaporkan SPT tahunan PPh tanpa terkecuali. Variabel ini menggunakan data dari tabel registrasi dan tabel tanda terima.
2. Variabel ekualisasi omset PPh dan penyerahan PPN sebagai variabel x2. Variabel ini menunjukkan risiko ketidakpatuhan WP badan per kelompok KLU dengan tidak melaporkan seluruh penghasilannya di SPT tahunan ataupun seluruh penyerahan Barang Kena Pajak (BKP) / Jasa Kena Pajak (JKP) di SPT masa PPN.
3. Variabel ekualisasi perolehan PPN dan peredaran usaha PPh sebagai variabel x3. Variabel ini untuk melihat secara spesifik WP badan per kelompok KLU yang melakukan banyak pembelian tetapi penyerahannya sangat sedikit.
4. Variabel histori pemeriksaan sebagai variabel x4. Variabel ini mengindikasikan bahwa WP badan per kelompok KLU yang terdapat koreksi termasuk kelompok KLU yang berisiko tidak patuh.
5. Variabel rasio biaya lain-lain terhadap biaya total sebagai variabel x5. Variabel ini bertujuan untuk memberi indikasi bahwa WP badan per kelompok KLU

berisiko membebankan biaya secara berlebihan. Beberapa contoh jenis biaya yang umumnya terdapat di biaya lainnya adalah: biaya, air, listrik, telpon, biaya bensin dan biaya ATK.

Sedangkan variabel y digunakan untuk melihat tingkat resiko kemungkinan hilangnya penerimaan pajak dilihat dari dampak fiskalnya. Variabel-variabel yang dibentuk dalam variabel y adalah sebagai berikut:

1. Variabel jumlah peredaran usaha sebagai variabel y1. Variabel ini untuk memperoleh peredaran usaha / penghasilan bruto WP badan per kelompok KLU.
2. Variabel total biaya sebagai variabel y2. Variabel ini untuk memperoleh total biaya yang dilaporkan oleh WP badan per kelompok KLU.
3. Variabel pengembalian pajak sebagai variabel y3. Variabel ini untuk memperoleh gambaran besarnya restitusi yang diterima oleh WP badan per kelompok KLU.
4. Variabel pajak yang disetor sebagai variabel y4. Variabel ini untuk memperoleh gambaran besarnya pembayaran pajak yang dilakukan WP badan per kelompok KLU pada tahun yang bersangkutan. Data pembayaran yang diambil berdasarkan data MPN. Data yang dipergunakan adalah data pembayaran seluruh jenis pajak yang dilakukan.

3.4.2 *Preprocessing*

Tahap dalam KDD selanjutnya adalah *preprocessing*, pada tahap ini dilakukan integrasi data untuk menggabungkan data dari tabel *database* yang berbeda. Selanjutnya dilakukan *data cleaning* untuk menghasilkan dataset yang bersih. Penjelasan kedua proses tersebut sebagai berikut:

1. Integrasi Data

   Proses integrasi bertujuan untuk menggabungkan data dari berbagai sumber ke dalam satu database baru, sehingga data-data tersebut dapat saling terhubung. Dalam hal ini, langkah yang dilakukan adalah mempersiapkan *database* lokal. *Database* yang digunakan adalah MySQL dengan aplikasi web phpMyAdmin. Tabel yang disiapkan pada database antara lain:

   a. Tabel WP untuk menyimpan data registrasi

b. Tabel SPT untuk menyimpan data tanda terima
c. Tabel SPT PPN untuk menyimpan data SPT yang berisi nilai peredaran usaha, nilai Dasar Pengenaan Pokok (DPP) PPN, nilai perolehan PPN, biaya lain-lain, dan biaya total
d. Tabel SP2 untuk menyimpan data riwayat pemeriksaan
e. Tabel Penerimaan untuk menyimpan data penerimaan
f. Tabel Pengembalian untuk menyimpan data pengembalian
g. Tabel Biaya Lain untuk menyimpan data biaya lain-lain

2. Data Cleaning

Pada tahap ini dilakukan pembersihan data yang mempunyai *missing value*, *redundant*, atau tidak relevan dengan penelitian. *Missing value* adalah nilai kosong, atau atribut yang tidak memiliki nilai pada suatu dataset. Sedangkan *redundant* adalah nilai yang sama yang lebih dari satu *record* dalam satu dataset.

Tujuan *preprocessing* dalam *data mining* adalah mentransformasi data ke suatu format yang menjadikan proses *data mining* lebih mudah dan efektif sesuai kebutuhan. *Preprocessing* juga membantu dalam mendapatkan hasil yang lebih akurat, mengurangi waktu untuk penghitungan dalam skala besar, dan membuat nilai data menjadi lebih kecil tanpa merubah informasi di dalamnya.

3.4.3 Transformasi Data

Tahap selanjutnya adalah tahap transformasi, yaitu merubah data yang telah dipilih. Tahapannya adalah pengubahan format data tersimpan sesuai bentuk yang dibutuhkan, dimana format file dari data ditransformasikan menjadi bentuk standar sesuai dengan aplikasi yang akan digunakan. Kemudian dilakukan skoring terhadap variabel yang digunakan berdasarkan ketentuan yang telah dideskripsikan. Transformasi data dilakukan menggunakan bahasa pemrograman PHP dan SQL.

3.4.4 Normalisasi

Normalisasi adalah sebuah proses merubah nilai atribut data. Proses ini dilakukan dengan pengskalaan terhadap nilai atribut menjadi range tertentu [22]. Dalam penelitian ini metode normalisasi yang digunakan adalah *z-score*. Metode *z-*

score merupakan metode normalisasi berdasarkan nilai rata-rata dan deviasi standar dari populasi data. Metode ini memungkinkan untuk membandingkan dua skor yang berasal dari populasi yang berbeda. Persamaan yang digunakan dalam metode *z-score* adalah:

$$v' = \frac{v-\mu}{\sigma} \qquad (4)$$

Keterangan:

v' = nilai yang baru

v = nilai lama

μ = rata-rata populasi

σ = standar deviasi populasi

Standar deviasi (σ) adalah ukuran jumlah variasi atau sebaran dari sekumpulan nilai yang paling sering digunakan dengan melihat akar kuadrat dari varian-nya. Rumus standar deviasi adalah:

$$s = \sqrt{s^2} \qquad (5)$$

Sedangkan untuk menghitung varian (s^2) rumus yang digunakan adalah:

$$s^2 = \frac{\sum_{i=1}^{n}(x_1-x')^2}{n-1} \qquad (6)$$

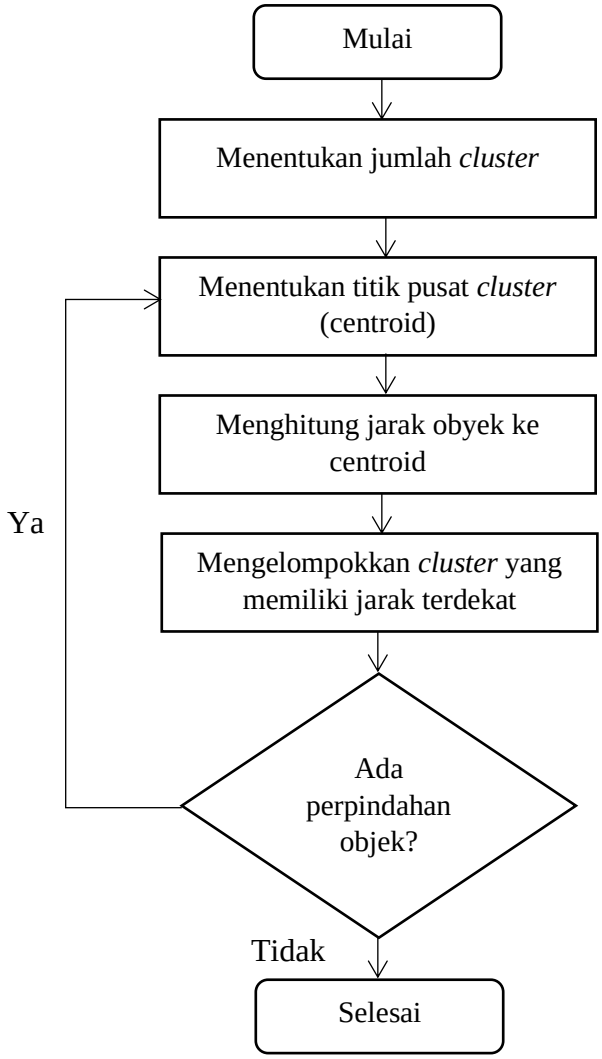
3.4.5 *Data Mining*

Tahap ini merupakan proses menemukan informasi, pola, dan korelasi dalam dataset menggunakan teknik atau metode tertentu. Penelitian ini akan melakukan perbandingan algoritma K-Means, K-Medoids, dan Hierarchical Clustering dalam pengelompokan (*clustering*) WP badan per kelompok KLU sebagai dasar pertimbangan dalam penggunaan algoritma K-Means.

Gambar 3-2 menjelaskan tahapan yang akan dilakukan pada proses *clustering* dengan algoritma K-Means, yaitu:

1. Menentukan jumlah *cluster* K yang akan dibentuk
2. Melakukan pembentukan *cluster* dengan algoritma K-Means, yaitu:

a. Inisialisasi centroid (titik pusat *cluster*) awal sebanyak K *cluster*
b. Menghitung jarak antara objek ke semua centroid
c. Menempatkan setiap objek ke dalam kelompok *cluster* yang memiliki jarak terdekat dengan centroid
d. Menghitung kembali centroid yang terbentuk dengan mengambil rata-rata dari semua objek yang dimiliki setiap *cluster* sebagai posisi centroid baru
e. Mengulang proses iterasi sampai tidak ada perpindahan objek antar *cluster*

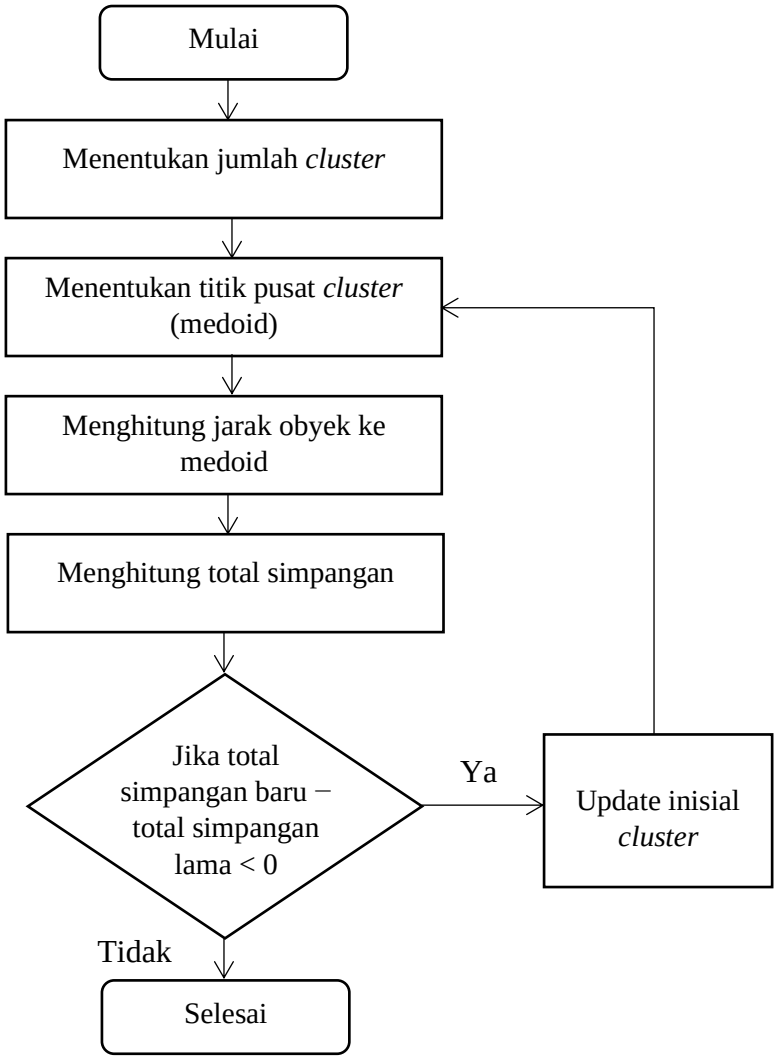


Gambar 3-2 Diagram Alir K-Means

Sedangkan tahapan pada proses *clustering* dengan algoritma K-Medoids dapat dilihat pada gambar 3-3 dengan penjelasan sebagai berikut:

1. Menentukan jumlah *cluster* K yang akan dibentuk

2. Melakukan pembentukan *cluster* dengan algoritma K-Medoids, yaitu:
   a. Inisialisasi k objek sebagai medoid secara acak
   b. Menghitung jarak antara objek ke medoid terdekat
   c. Menempatkan setiap objek ke dalam kelompok *cluster* yang memiliki jarak terdekat dengan medoid
   d. Memilih secara acak objek pada masing-masing *cluster* sebagai medoid baru
   e. Menghitung total simpangan (S), jika $S < 0$, maka tukar objek untuk membentuk sekumpulan k baru sebagai medoid
   f. Mengulang proses iterasi sampai tidak ada perpindahan objek

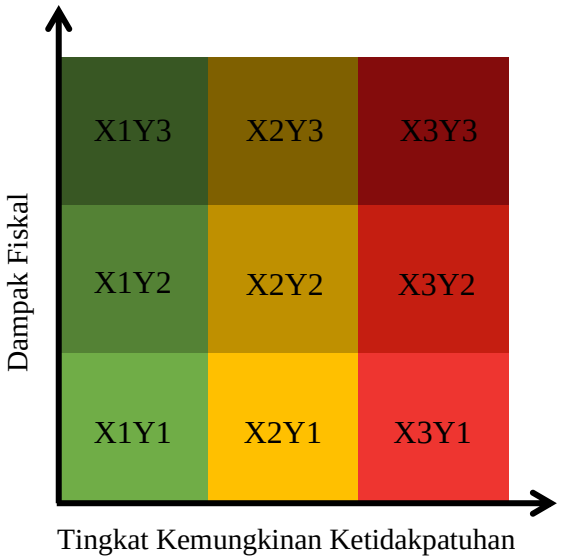


Gambar 3-3 Diagram Alir K-Medoids

Clustering WP badan per kelompok KLU dibuat dengan mengelompokkan KLU yang tidak memiliki resiko menyebabkan hilangnya penerimaan pajak sampai dengan KLU yang memiliki resiko tinggi dapat menyebabkan hilangnya penerimaan pajak.

Tingkat resiko kemungkinan hilangnya penerimaan pajak dapat diperoleh dari data ketidakpatuhan dalam melakukan pelaporan dan pembayaran pajak. Data pelaporan dan pembayaran tersebut dikelompokan menjadi 2 kategori yakni tingkat ketidakpatuhan dengan dampak fiskal yang disebabkan.

Dari data tersebut dibentuk kuadran KLU berdasarkan risiko kemungkinan hilangnya penerimaan pajak seperti Gambar 3-4:



Gambar 3-4 Kuadran KLU

3.4.5.1 *Clustering* Variabel x

Clustering variabel x mencerminkan tingkat ketidakpatuhan. Tingkat ketidakpatuhan WP badan per kelompok KLU dapat dilihat menggunakan beberapa variabel nilai seperti telah dijelaskan pada bab 3.4.1

Setelah tahapan *clustering* dilakukan maka daftar KLU dapat di kelompokan menjadi 3 *cluster* (risiko tinggi, sedang dan rendah):

1. kuadran x1 (masuk dalam *cluster* 1)
2. kuadran x2 (masuk dalam *cluster* 2)

3. kuadran x3 (masuk dalam *cluster* 3)

3.4.5.2 *Clustering* Variabel y

Clustering pada variabel y memperlihatkan dampak fiskal terhadap penerimaan negara. Variabel y dikategorikan berdasarkan konsekuensi tidak terpenuhinya kewajiban perpajakan. Berdasarkan kuadran yang dibentuk maka daftar KLU akan di *cluster* (kelompok) menjadi 3. Dari mulai kelompok yang konsekuensinya kecil sampai dengan konsekuensi tertinggi. Variabel yang digunakan seperti telah dijelaskan pada bab 3.4.1.

Setelah tahapan *clustering* dilakukan maka daftar KLU dapat di kelompokan menjadi 3 *cluster* (risiko tinggi, sedang dan rendah):

1. kuadran y1 (masuk dalam *cluster* 1)
2. kuadran y2 (masuk dalam *cluster* 2)
3. kuadran y3 (masuk dalam *cluster* 3)

3.4.5.3 *Silhouette Index*

Evaluasi dilakukan untuk mengetahui seberapa kompak/dekat jarak antar entitas dalam sebuah *cluster* dan seberapa jauh jarak antar *cluster*. Metode evaluasi yang akan digunakan pada penelitian ini adalah metode koefisien *silhouette* atau *silhouette index.*

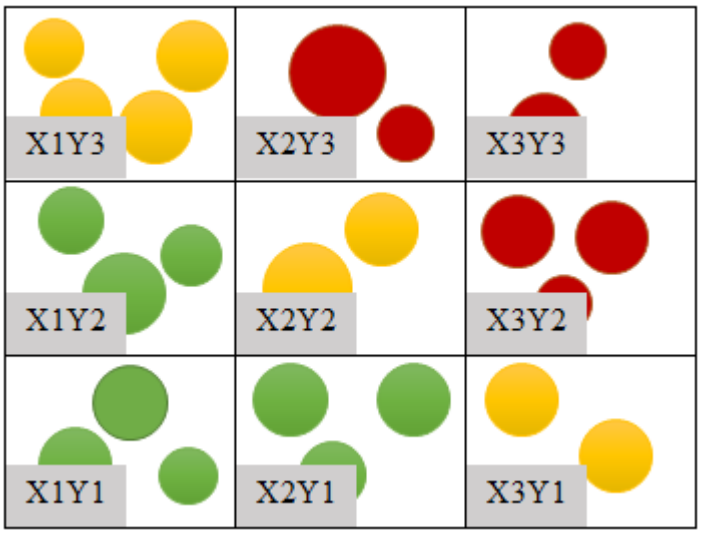
Kriteria koefisien *silhouette* dihitung berdasarkan jarak antar entitas, yaitu seberapa dekat jarak ke entitas lain di *cluster* yang sama dan dibandingkan dengan jarak terhadap jarak ke entitas di *cluster* lain. Nilai *silhouette* berkisar antara -1 hingga 1, di mana nilai yang mendekati 1 menunjukkan bahwa objek cocok dengan *cluster*-nya. Namun jika bernilai negatif mendekati -1 menandakan entitas tersebut berada di *cluster* yang salah. Pendekatan ini memberikan representasi tentang seberapa baik kualitas dari setiap *cluster* yang terbentuk. Semakin tinggi nilai rata-rata nya maka akan semakin baik. Tabel 3-1 menunjukkan kriteria nilai *silhoutte* berdasarkan Kaufman dan Rousseeuw.

Tabel 3-1 Kriteria Nilai *Silhoutte*

Silhouette Coeficient (SC)	Interpretation
0.71 – 1.00	Strong Stucture
0.51 − 0.70	Reasonable/Medium Structure
0.26 − 0.50	Weak Structure
≤ 0.25	No structure

3.4.6 Visualisasi

Pada tahap visualisasi, pola informasi yang dihasilkan dari tahap sebelumnya ditampilkan dalam bentuk kuadran agar lebih mudah dipahami oleh pihak yang berkepentingan. Hasil algoritma *clustering* yang telah dilakukan atas variabel x dan variabel y selanjutnya dapat digabungkan sehingga dapat membuat pengelompokan KLU seperti pada Gambar 3-5. Visualisasi ini dapat digunakan untuk mengetahui KLU yang perlu dilakukan pengawasan dan pemeriksaan lebih lanjut.



Gambar 3-5 Contoh Hasil Pengelompokan

3.4.7 Analisis

Tahap ini merupakan tahap interpretasi dari visualisasi yang telah dilakukan sebelumnya. Hasil analisis berupa sektor kelompok KLU dari *cluster* yang perlu diprioritaskan untuk dilakukan pemeriksaan dan pengawasan, serta penghitungan angka potensi penerimaan dari KLU tersebut.

3.5 Kesimpulan

Pada tahap ini dilakukan pembahasan hasil *data mining* yang telah dilakukan dengan menggunakan metode yang dipilih. Tahap ini dilakukan ketika model *data mining* yang sudah dibuat telah divalidasi.

3.6 Penyusunan Laporan

Penyusunan laporan dilakukan dengan menyusun dan mendokumentasikan dalam sebuah laporan langkah-langkah yang telah dilakukan sebelumnya. Laporan disusun berdasarkan aturan dan standard yang berlaku.

Halaman ini sengaja dikosongkan

BAB 4
HASIL DAN PEMBAHASAN

4.1 Gambaran Proses Pemeriksaan

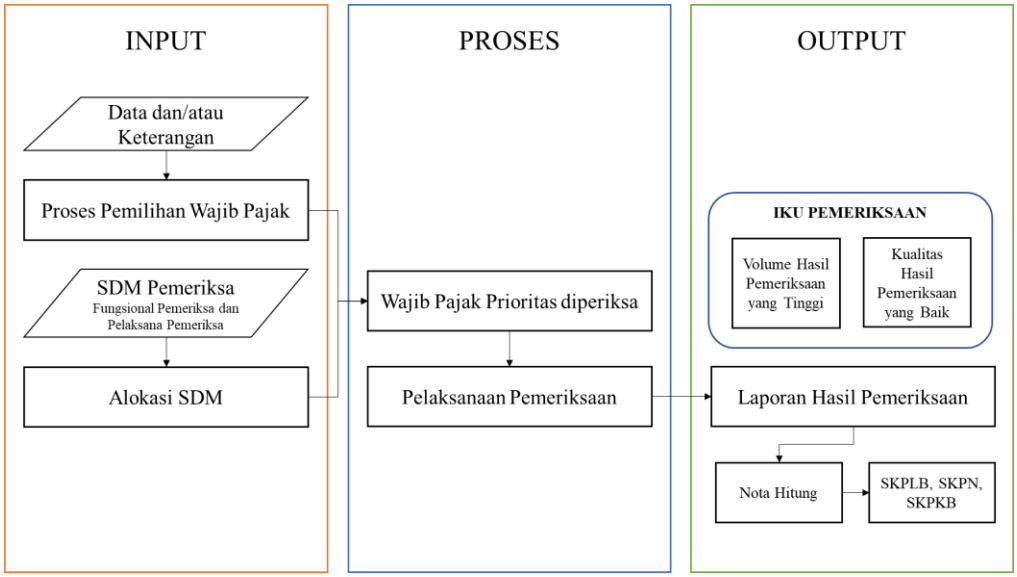
Berdasarkan Pasal 29 Undang-Undang No.28 Tahun 2007 tentang Perubahan Ketiga Undang-Undang No.6 Tahun 1983 Tentang Ketentuan Umun dan Tata Cara Perpajakan (Undang-undang KUP) menyatakan Direktorat Jenderal Pajak dalam rangka pengawasan berwenang melakukan pemeriksaan untuk menguji kepatuhan pemenuhan kewajiban perpajakan dan tujuan lain untuk melaksanakan ketentuan peraturan perundang-undangan yang berlaku.

Kemudian pada pasal 4 huruf a Peraturan Direktur Jenderal Pajak Nomor: PER-23/PJ/2013 tentang Standar Pemeriksaan, dijelaskan bahwa “Pelaksanaan pemeriksaan harus didahului dengan persiapan pemeriksaan yang baik sesuai dengan tujuan pemeriksaan, yang paling sedikit meliputi kegiatan mengumpulkan dan mempelajari data Wajib Pajak". Dijelaskan kembali pada pasal 4 huruf a angka 1, bahwa kegiatan mengumpulkan dan mempelajari data Wajib Pajak, meliputi antara lain mempelajari profil Wajib Pajak. Profil Wajib Pajak yang dimaksud termasuk jenis usaha dan Klasifikasi Lapangan Usaha (KLU).

Merujuk pada Surat Edaran nomor SE-15/PJ/2018 tentang kebijakan pemeriksaan, penyusunan Daftar Sasaran Prioritas Penggalian Potensi (DSP3) diperlukan dalam rangka meningkatkan kualitas penggalian potensi sehubungan dengan optimalisasi penerimaan pajak dari kegiatan pengawasan serta pencairan surat ketetapan pajak (pemeriksaan dan penagihan) dalam tahun berjalan. Penyusunan peta kepatuhan atas Wajib Pajak yang terdaftar pada KPP tersebut berdasarkan Klasifikasi Lapangan Usaha (KLU)/sektor/subsektor/industri, letak geografis, Produk Domestik Regional Bruto (PDRB), dan/atau fakta lapangan. Masih merujuk pada Surat Edaran yang sama, diperlukan pula peningkatan kualitas penentuan wajib pajak yang dilakukan pemeriksaan.

Kegiatan pemeriksaan terdiri dari tiga komponen, yakni melakukan pemilihan WP yang akan diperiksa secara obyektif, mengoptimalkan alokasi Sumber Daya Manusia (SDM) pemeriksa pajak, dan perbaikan terus menerus

terhadap aturan perpajakan khususnya terkait pemeriksaan. Adapun garis besar proses bisnis pemeriksaan adalah seperti pada gambar 4-1.



Gambar 4-1 Proses Bisnis Pemeriksaan

Proses pemilihan Wajib Pajak yang akan diperiksa dilakukan dengan menyusun peta kepatuhan dan Daftar Sasaran Prioritas Penggalian Potensi (DSP3). Tujuan disusunnya DSP3 ini adalah supaya setiap KPP dapat menentukan secara spesifik daftar WP yang akan dilakukan penggalian potensi. Penyusunan ini dilakukan dengan melakukan analisis terhadap seluruh data dan informasi yang dimiliki oleh KPP yang berasal dari sistem informasi yang dimiliki DJP dan dikombinasikan dengan data fakta lapangan.

Kepala KPP menentukan populasi WP atau sektor usaha/KLU yang akan dimasukkan ke dalam DSP3 berdasarkan variabel-variabel seperti berikut:

1. Indikasi ketidakpatuhan tinggi

   Indikasi ketidakpatuhan yang dilihat dari ketidakpatuhan material. Indikasi ini ditunjukkan dengan adanya kesenjangan pajak (*tax gap*) antara profil yang dilaporkan di SPT dengan profil ekonomi sebenarnya yang dapat diketahui dari data internal maupun eksternal.
2. Indikasi modus ketidakpatuhan WP

   Indikasi ini meliputi WP tidak melaporkan omset yang sebenarnya atau WP membebankan biaya yang tidak seharusnya.

3. Identifikasi nilai potensi pajak

   WP yang menjadi prioritas adalah WP yang memiliki potensi pajak besar.
4. Identifikasi kemampuan WP untuk membayar ketetapan pajak (collectability)

   Saat menentukan DSP3, kepala KPP harus memperhatikan risiko ketertagihan atau kemampuan WP dalam membayar ketetapan pajaknya.
5. Pertimbangan Direktur Jenderal Pajak

   Direktur Jenderal Pajak sesuai dengan kewenangannya dapat menetapkan WP yang akan menjadi DSP3 berdasarkan pertimbangan tertentu.

Untuk melaksanakan pemeriksaan, maka perlu dilakukan pengalokasian SDM. SDM pemeriksa pajak terdiri dari fungsional pemeriksa dan pelaksana pemeriksa. Pemeriksa Pajak bertanggung jawab atas pelaksanaan pemeriksaan termasuk atas seluruh dokumentasi pemeriksaan sampai dengan SP2 telah selesai dilaksanakan dan telah dibuat Laporan Hasil Pemeriksaan (LHP), termasuk pengiriman Kertas Kerja Pemeriksaan (KKP), LHP, dan nota hitung ketetapan pajak kepada pejabat yang bertanggung jawab untuk melaksanakan administrasi pemeriksaan.

4.2 Pengolahan Data

Tahapan awal yang dilakukan adalah seleksi fitur data, integrasi data dan pembersihan data. Dari tahapan awal ini diharapkan akan menghasilkan dataset yang bersih sehingga bisa diproses di tahapan selanjutnya. Penerapan konsep *data mining* menggunakan metode K-Means ini adalah untuk mengetahui KLU mana yang potensial.

4.2.1 Seleksi Data

Data yang akan digunakan dalam penelitian ini adalah data yang berasal dari DJP, diantaranya data registrasi, data tanda terima, data Surat Pemberitahuan (SPT), data riwayat pemeriksaan, data penerimaan, dan data pengembalian. Jumlah data awal dari masing-masing tabel yang diambil dapat dilihat pada tabel 4-1.

Tabel 4-1 Jumlah Data Awal

NO	Tabel	Jumlah Data Awal
1	Registrasi	232,507
2	Tanda Terima	19,970,992
3	SPT – Biaya Lain	524,228
4	Riwayat Pemeriksaan	140,945
5	Penerimaan	6,112,471
6	Pengembalian	49,909
7	SPT	526,451
Total		**27,557,503**

Data yang didapat berbentuk csv dan txt, data-data yang belum sesuai dirubah ke dalam format csv. Kemudian dilakukan seleksi fitur data dengan memilih kolom-kolom yang diperlukan. Pada masing-masing tabel registrasi dan SPT semua kolom akan digunakan. Sehingga seleksi fitur data hanya akan dilakukan pada tabel penerimaan, pengembalian, tanda terima, dan riwayat pemeriksaan. Data awal penerimaan dapat dilihat pada tabel 4-2.

Tabel 4-2 Data Awal Penerimaan

TAHUN_PAJAK	KPPADM	KD_ KLU	KJS	JML SSP	JML NILAI BAYAR
2016	221	96303	411128100	1	2000000
2017	437	47191	411128403	15	43664636
2017	806	45403	411122100	1	545455
2017	102	47414	411211930	21	20423945
2019	86	84111	411128100	5	766631200
2020	418	85499	411124104	1	200000
2016	522	46339	411128409	1	718873
2016	624	64123	411611201	6	90000000
2017	542	47591	411211100	122	751375818
…	…	…	…	…	…

Dari data awal dilakukan pemilihan kolom yang akan dijadikan fitur untuk tahapan selanjutnya. Data yang dijadikan fitur adalah data TAHUN, KPP, KD KLU, dan JML NILAI BAYAR seperti yang terlihat pada tabel 4-3.

Tabel 4-3 Tabel Penerimaan setelah Seleksi Fitur

TAHUN_PAJAK	KPPADM	KD_KLU	JML NILAI BAYAR
2016	221	96303	2000000
2017	437	47191	43664636
2017	806	45403	545455
2017	102	47414	20423945
2019	86	84111	766631200
2020	418	85499	200000
2016	522	46339	718873
2016	624	64123	90000000
2017	542	47591	751375818
…	…	…	…

Kemudian data awal pengembalian seperti terlihat pada table 4-4 dilakukan seleksi fitur terhadap data TAHUN, KPP, KD KLU, dan JML NILAI BAYAR yang terlihat pada table 4-5.

Tabel 4-4 Data Awal Pengembalian

TAHUN_PAJAK	KPPADM	KD_ KLU	KJS	JML_SSP	JML_NILAI_BAYAR
2016	335	47726	411211000	1	-18,495,855
2016	618	96999	411125000	3	-7,936,000
2016	511	21022	411126000	2	-1,652,298,983
2016	431	17019	411211000	2	-256,872,968
2016	941	70100	411211000	1	-123,764,545
2016	645	41012	411211000	1	-3,115,445
2016	44	46319	411211000	1	-1,240,678,974
2016	824	41019	411126000	1	-4,939,669
2016	453	47245	411126000	1	-758,514,375
2016	28	47712	411125000	7	-665,750
…	…	…	…	…	…

Tabel 4-5 Tabel Pengembalian setelah Seleksi Fitur

TAHUN_PAJAK	KPPADM	KD_KLU	JML_NILAI_BAYAR
2016	335	47726	-18,495,855
2016	618	96999	-7,936,000
2016	511	21022	-1,652,298,983
2016	431	17019	-256,872,968
2016	941	70100	-123,764,545
2016	645	41012	-3,115,445
2016	44	46319	-1,240,678,974
2016	824	41019	-4,939,669
2016	453	47245	-758,514,375
2016	28	47712	-665,750
…	…	…	…

Data awal riwayat pemeriksaan terlihat pada table 4-6. Setelah dilakukan seleksi fitur didapatkan table seperti pada table 4-7.

Tabel 4-6 Data Awal Pemeriksaan

TAHUN_PAJAK	KD_KLU	KPPADM	JML_ SP2	JML_ SKPKB	JML_ SKPLB	JML_ SKPN
2016	45103	901	1	0	0	0
2016	46593	724	0	0	0	29
2016	47722	505	1	0	0	0
2016	64140	72	2	0	0	1
2016	12091	527	2	0	0	0
2016	45403	901	3	2	0	13
2016	07102	434	1	2	8	7
2016	33121	417	2	8	0	3
2016	62010	741	1	30	1	7
2016	49431	712	10	57	6	85
…	…	…	…	…	…	…

Tabel 4-7 Tabel Pemeriksaan setelah Seleksi Fitur

TAHUN_PAJAK	KD_KLU	KPPADM	JML_SKPKB	JML_SKPLB
2016	45103	901	0	0
2016	46593	724	0	0
2016	47722	505	0	0
2016	64140	72	0	0
2016	12091	527	0	0
2016	45403	901	2	0
2016	07102	434	2	8
2016	33121	417	8	0
2016	62010	741	30	1
2016	49431	712	57	6
…	…	…	…	…

Data awal tanda terima seperti pada table 4-8. Dilakukan seleksi fitur hingga didapatkan tabel seperti pada table 4-9.

Tabel 4-8 Data Awal Tanda Terima

ID JNS SPT	NM_SPT	ID_MS_TH _PAJAK	KD_ KLU	KPP ADM	NM_KPP	Jumlah
11	SPT Tahunan PPh Badan	202000		648	PRATAMA TUBAN	1
11	SPT Tahunan PPh Badan	202000		649	PRATAMA JOMBANG	1

11	SPT Tahunan PPh Badan	202000		706	PRATAMA SINTANG	2
11	SPT Tahunan PPh Badan	202000		912	PRATAMA RABA BIMA	1
12	SPT Masa PPN dan PPnBM	201500	41012	122	PRATAMA MEDAN KOTA	1
12	SPT Masa PPN dan PPnBM	201500	41012	416	PRATAMA TGR TIMUR	1
12	SPT Masa PPN dan PPnBM	201500	41012	811	PRATAMA KENDARI	1
13	SPT Masa PPN Pemungut	201907	84112	626	PRATAMA JEMBER	43
13	SPT Masa PPN Pemungut	201907	84112	627	PRATAMA BANYUWANGI	14
13	SPT Masa PPN Pemungut	201907	84112	628	PRATAMA BATU	6

Tabel 4-9 Tabel Tanda Terima setelah Seleksi Fitur

ID JNS SPT	ID_MS_TH_PAJAK	KD_KLU	KPPADM	Jumlah
11	202000		648	1
11	202000		649	1
11	202000		706	2
11	202000		912	1

4.2.2 *Data Cleaning*

Pada tahap ini dilakukan pembersihan data yang mempunyai *missing value*, *redundant*, atau tidak relevan dengan penelitian. *Missing value* adalah nilai kosong, atau atribut yang tidak memiliki nilai pada suatu dataset. Sedangkan *redundant* adalah nilai yang sama yang lebih dari satu *record* dalam satu dataset.

Data yang akan digunakan disimpan terlebih dahulu ke dalam *database* untuk memudahkan dalam pengolahan data. *Database* yang digunakan adalah MySQL dengan aplikasi web phpMyAdmin. PhpMyAdmin merupakan bagian untuk mengelola *database* MySQL yang termasuk dalam perangkat lunak xampp. Xampp mempunyai fungsi sebagai server yang berdiri sendiri (localhost), yang terdiri dari program MySQL database, Apache HTTP Server, dan ditulis dalam bahasa pemrograman PHP.

Data dalam format .csv diimpor dalam *database* dan tabel yang telah dibuat sebelumnya seperti yang terlihat pada gambar 4-2 sampai dengan gambar 4-7.

#	Name	Type
1	id_wp	int(11)
2	tahun	year(4)
3	klu	varchar(10)
4	kppadm	varchar(10)
5	jumlah_wp	int(11)

Gambar 4-2 Struktur Tabel Registrasi

#	Name	Type
1	id_wp	int(11)
2	tahun	year(4)
3	klu	varchar(10)
4	kppadm	varchar(10)
5	jumlah_spt	int(11)

Gambar 4-3 Struktur Tabel Tanda Terima

#	Name	Type
1	id_wp	int(11)
2	tahun	year(4)
3	klu	varchar(10)
4	kppadm	varchar(10)
5	jumlah_dpp_ppn	bigint(20)
6	biaya_total	bigint(20)
7	biaya_perolehan	bigint(20)
8	nilai_peredaran_usaha	bigint(20)

Gambar 4-4 Struktur Tabel SPT

#	Name	Type
1	id_wp	int(11)
2	tahun	year(4)
3	klu	varchar(10)
4	kppadm	varchar(10)
5	jumlah_sp2	int(11)
6	jumlah_skpkb	int(11)
7	jumlah_skplb	int(11)
8	jumlah_skpn	int(11)

Gambar 4-5 Struktur Tabel Riwayat Pemeriksaan

#	Name	Type
1	id_wp	int(11)
2	tahun	year(4)
3	klu	varchar(10)
4	kppadm	varchar(10)
5	jumlah_pengembalian	bigint(20)

Gambar 4-6 Struktur Tabel Pengembalian

#	Name	Type
1	id_wp	int(11)
2	tahun	year(4)
3	klu	varchar(10)
4	kppadm	varchar(10)
5	jumlah_penerimaan	bigint(20)

Gambar 4-7 Struktur Tabel Penerimaan

Pada data awal penerimaan terdapat kolom KJS dan jumlah SSP. Kedua kolom tersebut tidak digunakan dalam pengolahan data, namun berkaitan dengan kolom jumlah nilai bayar. Untuk mendapatkan jumlah nilai bayar, dilakukan penjumlahan untuk setiap JML NILAI BAYAR dengan KJS dan JML SSP yang memiliki TAHUN, KPP, dan KD KLU yang sama dengan menggunakan bahasa pemrograman PHP yang dapat dilihat pada nomor 1 tabel 4-11. Penjumlahan tersebut juga dilakukan untuk data pengembalian seperti pada *script* nomor 2 tabel 4-11.

Sedangkan untuk tanda terima dengan jumlah records 19.970.992, data yang akan digunakan hanya data yang memiliki ID JNS SPT = 11 atau SPT tahunan badan yang dituliskan pada *script* nomor 3 dalam tabel 4-11.

Dikarenakan keterbatasan kemampuan perangkat yang digunakan, maka jumlah KPP ADM yang digunakan dalam penelitian berjumlah 2 Kantor Wilayah (Kanwil) yang dibawahnya mencakup 9 KPP ADM, yaitu Kanwil Wajib Pajak Besar yang terdiri dari KPP Wajib Pajak Besar Satu (091), KPP Wajib Pajak Besar Dua (092), KPP Wajib Pajak Besar Tiga (051), KPP Wajib Pajak Besar Empat (093), dan Kanwil Jakarta Khusus, yang terdiri dari KPP Penanaman Modal Asing Satu (052), KPP Penanaman Modal Asing Dua (055), KPP Penanaman Modal Asing Tiga (056), KPP Penanaman Modal Asing Empat (057), dan KPP Penanaman Modal Asing Lima (058). Digunakan *script* pada nomor 6 dalam table 4-10 untuk mengambil data hanya dengan kode KPP yang telah disebutkan.

Tabel 4-10 Script PHP SQL

	Keterangan dan Script yang digunakan	
1	Ket.	Menjumlahkan JML NILAI BAYAR dengan KJS dan JML SSP yang memiliki TAHUN, KPP, dan KD KLU yang sama pada data Penerimaan

	Script	`foreach ($kelompok as $tahun => $value1) {`
		`  foreach ($value1 as $klu => $value2) {`
		`  foreach ($value2 as $kppadm => $value3) {`
		`  $jumlah_penerimaan = array_sum($value3);`
		`  $mysqli->query("INSERT INTO penerimaan (tahun, klu, kppadm, jumlah_penerimaan) VALUES ('$tahun', '$klu', '$kppadm', '$jumlah_penerimaan')");`
		`  }`
		`  }`
		`}`

2	Ket.	Menjumlahkan JML NILAI BAYAR dengan KJS dan JML SSP yang memiliki TAHUN, KPP, dan KD KLU yang sama pada data Pengembalian

	Script	`foreach ($kelompok as $tahun => $value1) {`
		`  foreach ($value1 as $klu => $value2) {`
		`  foreach ($value2 as $kppadm => $value3) {`
		`  $jumlah_pengembalian= array_sum($value3);`
		`  $mysqli->query("INSERT INTO pengembalian (tahun, klu, kppadm, jumlah_pengembalian) VALUES ('$tahun', '$klu', '$kppadm', '$jumlah_pengembalian)");`
		`  }`
		`  }`
		`}`

3	Ket.	Mengambil data SPT dengan ID JNS SPT = 11 (SPT Tahunan Badan)

	Script	`if ($pecah[0] == 11){`
		`  $tahun = substr($pecah[2], 0, 4);`
		`  $klu = $pecah[3];`
		`  $kppadm = $pecah[4];`
		`  $jumlah_spt = $pecah[6];`
		`  $id_jns_spt = $pecah[0];`

		`$mysqli->query("INSERT INTO spt (tahun, klu, kppadm, jumlah_spt) VALUES ('$tahun', '$klu', '$kppadm', '$jumlah_spt')");` `}`
4	Ket.	Menghapus row yang mengandung data KPP atau KLU yang kosong (dilakukan untuk semua tabel)
	Script	`DELETE FROM table_name WHERE klu IS NULL or kppadm IS NULL`
5	Ket.	Menghapus row yang mengandung data KPP atau KLU = ERR00 (dilakukan untuk semua tabel)
	Script	`DELETE FROM table_name WHERE klu = 'ERR00' or kppadm = 'ERR00'`
6	Ket.	Mengambil hanya data dengan Kode KPPADM tertentu

	Script	`$kanwil = array('51','91','92','93','52','55','56','57','58');`
		`if (in_array($kppadm, $kanwil)) {`
		`$mysqli->query("INSERT INTO wp (tahun, klu, kppadm, jumlah_wp) VALUES ('$tahun', '$klu', '$kppadm', '$jumlah_wp')"); }`

Sehingga jumlah *records* dari masing-masing tabel data yang telah diimpor dapat dilihat pada tabel 4-11.

Tabel 4-11 Jumlah Data setelah *Cleaning*

NO	Tabel	Jumlah Data Uji
1	Registrasi	4,101
2	Tanda Terima	4,618
3	SPT − Biaya Lain	4,594
4	Riwayat Pemeriksaan	2,520
5	Penerimaan	4,840
6	Pengembalian	2,243
7	SPT	4,624
Total		**27,540**

4.2.3 Transformasi Data

Langkah selanjutnya akan dilakukan proses skoring dari masing-masing variabel. Data akan dikelompokan menjadi 2 kategori yakni variabel x untuk melihat tingkat ketidakpatuhan dan variabel y untuk melihat dampak fiskal yang disebabkan. Tabel 4-13 menunjukkan range skoring untuk variabel x. Masing-masing variabel akan di skoring berdasarkan hal-hal sebagai berikut:

1. Variabel pelaporan SPT tahunan (x1) diambil dari jumlah WP pada tabel registrasi dan jumlah laporan SPT tahunan badan pada tabel tanda terima. Variabel ini berfungsi untuk mendapatkan WP badan yang tidak melaporkan SPT tahunannya sesuai ketentuan UU. Tanda terima SPT tahunan WP badan per kelompok KLU disandingkan dengan data registrasi WP badan per kelompok KLU.

   Untuk kelompok KLU yang melaporkan SPT-nya selama 4 tahun berturut-turut maka diberi skor 0, untuk kelompok KLU yang tidak melaporkan SPT nya selama 1 tahun diberikan skor 50, dan untuk kelompok KLU yang tidak melaporkan SPT nya selama lebih dari 1 tahun maka diberi skor 100.

   Jumlah kelompok KLU pada masing-masing KPP ADM dapat dilihat pada tabel 4-12.

Tabel 4-12 Jumlah KLU Masing-masing KPP

Kode KPP	Nama KPP	Jumlah KLU
51	KPP Wajib Pajak Besar Tiga	584
52	KPP Penanaman Modal Asing Satu	430
55	KPP Penanaman Modal Asing Dua	468
56	KPP Penanaman Modal Asing Tiga	483
57	KPP Penanaman Modal Asing Empat	448
58	KPP Penanaman Modal Asing Lima	572
91	KPP Wajib Pajak Besar Satu	200
92	KPP Wajib Pajak Besar Dua	440
93	KPP Wajib Pajak Besar Empat	476
	Total	**4101**

2. Variabel ekualisasi omset PPh dan penyerahan PPN (x2) diambil dari nilai peredaran usaha dan nilai DPP PPN pada tabel SPT. Nilai peredaran usaha dilihat dari form 1771-I SPT Tahunan Badan, sedangkan DPP PPN dapat dilihat dari jumlah satu tahun objek PPN dalam SPT Masa PPN. Variabel ini untuk menentukan apakah terjadi *underreporting* peredaran usaha di SPT tahunan atau *underreporting* pelaporan penyerahan BKP dan/atau JKP di SPT masa PPN. Variabel ini didapatkan dengan menghitung selisih omset SPT tahunan dan jumlah DPP SPT masa 12 bulan dibandingkan dengan omset SPT tahunan seperti pada persamaan 1, angka ini disebut nilai ekualisasi.

$$ekualisasi\ Omset\ PPh\ dan\ Penyerahan\ PPN = \frac{Omset\ PPh - Penyerahan\ PPN}{Omset\ PPh}\ x\ 100\% \qquad (1)$$

   Untuk ekualisasi dengan nilai kurang dari -10% atau lebih dari 10% diberikan skor 80, untuk ekualisasi dengan nilai kurang dari -20% atau lebih dari 20% diberikan skor 100, dan untuk range nilai diluar itu diberikan skor 0.

   Dasar hukum dari ekualisasi pajak adalah Surat Edaran Direktur Jenderal Pajak Nomor SE-10/PJ/2017 tentang Petunjuk Teknis Pemeriksaan Lapangan dalam Rangka Pemeriksaan untuk Menguji Kepatuhan Pemenuhan Kewajiban Perpajakan.

3. Variabel ekualisasi perolehan PPN terhadap peredaran usaha PPh (x3) diambil dari nilai peredaran usaha dan nilai perolehan PPN pada tabel SPT. Nilai peredaran usaha dilihat dari form 1771-I SPT Tahunan Badan, sedangkan nilai perolehan PPN dilihat dari form SPT Masa PPN 1111 AB bagian II.D. Variabel ini untuk melihat perbandingan antara pembelian dari data perolehan SPT masa PPN terhadap peredaran usaha WP. WP yang melakukan pembelian dengan nilai lebih dari 100% peredaran usahanya dianggap berisiko. Variabel ini didapat dengan menghitung nilai ekualisasi perolehan PPN terhadap peredaran usaha PPh. Selisih omset SPT tahunan dan jumlah perolehan PPN dibandingkan dengan omset SPT tahunan seperti pada persamaan 2.

$$ekualisasi\ Omset\ PPh\ dan\ Perolehan\ PPN = \frac{Omset\ PPh - Perolehan\ PPN}{Omset\ PPh}\ x\ 100\% \qquad (2)$$

Untuk ekualisasi dengan nilai kurang dari sama dengan 50% diberikan skor 0, untuk ekualisasi dengan nilai lebih dari 50% dan kurang dari 100% diberikan skor 75, dan untuk range nilai lebih dari sama dengan 100% diberikan skor 100. Dasar hukum dari ekualisasi pajak adalah Surat Edaran Direktur Jenderal Pajak Nomor SE-10/PJ/2017 tentang Petunjuk Teknis Pemeriksaan Lapangan dalam Rangka Pemeriksaan untuk Menguji Kepatuhan Pemenuhan Kewajiban Perpajakan.

4. Variabel riwayat pemeriksaan (x4) diambil dari tabel riwayat pemeriksaan. Variabel ini untuk menentukan WP badan per kelompok KLU yang dilakukan pemeriksaan dengan melihat apakah terdapat data ketetapan atas wajib pajak.

   Untuk kelompok KLU yang tidak ada koreksi selama 4 tahun berturut-turut diberikan skor 0, untuk kelompok KLU yang terdapat koreksi selama 1 tahun diberikan skor 50, sedangkan untuk kelompok KLU yang terdapat koreksi lebih dari 1 tahun diberi skor 100.

   Dasar hukum dari ekualisasi pajak adalah Surat Edaran Direktur Jenderal Pajak Nomor SE - 08/PJ/2012 tentang Pedoman Penyusunan Kertas Kerja Pemeriksaan Untuk Menguji Kepatuhan Pemenuhan Kewajiban Perpajakan.

5. Variabel rasio biaya lain-lain terhadap biaya total (x5) diambil dari jumlah biaya lain pada tabel SPT−Biaya Lain dan nilai biaya total pada tabel SPT. Biaya lain-lain dapat dilihat dari form SPT 1771 Lampiran II nomor 11, sedangkan biaya total dilihat dari form SPT 1771 Lampiran II nomor 14. Variabel ini digunakan untuk mengidentifikasi WP badan per kelompok KLU yang memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar. Dalam hal tersebut ketidakwajaran dianggap terjadi jika porsi biaya lain-lain dibanding total biaya lebih dari sama dengan 30%. Rasio Biaya lain-lain terhadap biaya total dihitung dengan persamaan 3.

$$rasio\ biaya\ lain\ terhadap\ biaya\ total = \frac{Biaya\ Lain-lain}{Biaya\ Total} x\ 100\% \quad (3)$$

   Untuk Rasio kurang dari 30% diberikan skor 0, sedangkan untuk rasio yang lebih dari sama dengan 30% diberikan skor 100.

Dasar hukum dari ekualisasi pajak adalah Surat Edaran Direktur Jenderal Pajak Nomor SE-96/PJ/2009 tentang Rasio Total Benchmarking.

Tabel 4-13 Tabel Skoring

Variabel	Keterangan	Skor
x1	Lapor 4 tahun	0
	Tidak lapor 1 tahun	50
	Tidak lapor > 1 tahun	100
x2	-10% < Ekualisasi omset < 10%	0
	Ekualisasi omset < -10% or > 10%	80
	Ekualisasi omset < -20% or > 20%	100
x3	Ekualisasi perolehan ≤ 50%	0
	50% < Ekualisasi perolehan < 100%	75
	Ekualisasi perolehan ≥ 100%	100
x4	Tidak ada ketetapan	0
	1 Tahun ketetapan	50
	> 1 Tahun ketetapan	100
x5	Ekualisasi Biaya < 30%	0
	Ekualisasi Biaya ≥ 30%	100

Hasil dari tahapan skoring seperti terlihat pada tabel 4-14.

Tabel 4-14 Hasil Skoring Variabel x

id_x	tahun	klu	kppadm	x1	x2	x3	x4	x5
1	2016	1262	92	0	0	100	1	0
2	2016	1262	51	0	0	100	1	0
3	2016	2117	92	0	80	0	1	0
4	2016	2117	91	0	0	0	0	0
5	2016	10320	92	0	0	100	1	0
6	2016	10330	92	0	0	0	0	0
7	2016	10431	92	0	0	75	1	0
8	2016	10432	92	0	0	0	1	0
9	2016	10490	92	0	0	0	1	0
10	2016	10520	92	0	0	0	1	100
11	2016	10590	92	0	0	0	0	0
12	2016	10617	92	0	0	0	1	0
13	2016	10721	92	0	0	100	1	100
14	2016	10721	51	0	80	75	1	100
15	2016	10732	92	0	0	0	1	0
16	2016	10740	92	1	0	75	1	0
17	2016	10761	92	0	0	0	1	0
18	2016	10772	92	0	0	0	1	0
19	2016	10793	92	1	0	75	1	0
20	2016	10802	92	0	0	100	1	0

…	…	…	…	…	…	…	…	…
5711	2019	85602	58	0	0	0	0	0
5712	2019	86904	58	0	0	0	0	0
5713	2019	95120	58	0	0	0	0	0

Setelah mencari skor dari masing-masing kelompok KLU, langkah selanjutnya adalah mencari rata-rata nilai selama 4 tahun. Variabel x2, x3, dan x5 masing-masing dijumlahkan dan dibagi dengan jumlah tahun seperti dapat dilihat pada persamaan (4).

$$rata - rata\ ekualisasi = \frac{Jumlah\ ekualisasi\ selama\ n\ tahun}{n\ tahun} \quad (4)$$

Menggunakan perintah sql pada *database* mysql:

```
SELECT AVG(x2) as jml_x2, AVG(x3) as jml_x3, AVG(x5) as jml_x5
FROM sumbu_x_skoring WHERE klu='$klu' AND kppadm='$kppadm'
```

Hasil dari pencarian rata-rata untuk seluruh variabel dapat dilihat pada tabel 4-15.

Tabel 4-15 Rata-rata Variabel x Tahun 2016-2019

id_x_all	klu	kppadm	x1	x2	x3	x4	x5
1	1262	92	0	0	43.75	100	0
2	1262	51	0	0	100	100	0
3	2117	92	100	90	0	100	0
4	2117	91	0	0	0	50	0
5	10320	92	0	20	93.75	100	0
6	10330	92	0	65	25	100	0
7	10431	92	0	0	37.5	100	0
8	10432	92	0	0	0	100	0
9	10490	92	0	25	37.5	100	0
10	10520	92	0	0	0	100	25
11	10590	92	0	0	0	0	0
12	10617	92	0	0	0	100	0
13	10721	92	0	70	93.75	100	100
14	10721	51	0	95	93.75	100	75
15	10732	92	0	0	0	50	0
16	10740	92	100	0	75	50	0
17	10761	92	0	0	0	50	0
18	10772	92	0	50	0	100	50
19	10793	92	100	20	37.5	50	0
20	10802	92	0	0	87.5	100	0
…	…	…	…	…	…	…	…
1011	86904	58	0	0	0	0	0
1012	95120	58	0	0	0	0	0
1013	47511	56	0	0	0	0	0

Pada variabel y (dampak fiskal) akan dikategorikan berdasarkan konsekuensi tidak terpenuhinya kewajiban perpajakan. Variabel yang digunakan untuk *clustering* variabel y adalah data pembayaran yang meliputi:

1. Variabel jumlah peredaran usaha (y1) diambil dari nilai peredaran usaha pada tabel SPT. Variabel ini untuk memperoleh peredaran usaha / penghasilan bruto WP badan per kelompok KLU.
2. Variabel total biaya (y2) diambil dari total biaya pada tabel SPT. Variabel ini untuk memperoleh total biaya yang dilaporkan oleh WP badan per kelompok KLU.
3. Variabel pengembalian pajak (y3) diambil dari nilai jumlah pengembalian pada tabel pengembalian. Variabel ini untuk memperoleh gambaran besarnya restitusi yang diterima oleh WP badan per kelompok KLU.
4. Variabel pajak yang disetor (y4) diambil dari nilai jumlah penerimaan pada tabel penerimaan. Variabel ini untuk memperoleh gambaran besarnya pembayaran pajak yang dilakukan WP badan per kelompok KLU. Data pembayaran yang diambil berdasarkan data MPN. Data yang dipergunakan adalah data pembayaran seluruh jenis pajak yang dilakukan.

Sehingga didapatkan tabel variabel y seperti terlihat pada tabel 4-16.

Tabel 4-16 Tabel Variabel y

id_y	tahun	klu	kppadm	y1	y2	y3	y4
1	2016	1262	92	1.76E+12	1.61E+12	-3.1E+10	9.27E+10
2	2016	1262	51	2.2E+13	2.05E+13	-1.9E+11	5.35E+11
3	2016	2117	92	2.37E+12	2.15E+12	-1.7E+09	5.84E+10
4	2016	2117	91	0	8.76E+09	0	86307200
5	2016	10320	92	2.7E+12	1.77E+12	-1.7E+10	2.49E+11
6	2016	10330	92	4.48E+12	3.89E+12	-3.9E+09	4.29E+11
7	2016	10431	92	5.67E+13	5.28E+13	-2.7E+12	-6.6E+11
8	2016	10432	92	1.46E+14	1.44E+14	-2.2E+12	-3.2E+11
9	2016	10490	92	3.36E+13	3.3E+13	-7E+11	-3.8E+11
10	2016	10520	92	4.08E+13	3.34E+13	0	5.23E+12
11	2016	10590	92	0	0	0	0
12	2016	10617	92	2.02E+13	2.02E+13	-5.7E+08	2.45E+12
13	2016	10721	92	7.75E+12	6.23E+12	-1.6E+10	1.07E+12
14	2016	10721	51	5.42E+12	5.78E+12	-3.2E+10	5E+11
15	2016	10732	92	2.76E+12	2.33E+12	0	1.99E+11

16	2016	10740	92	3.42E+13	2.89E+13	0	2.96E+12
17	2016	10761	92	1.24E+13	1.11E+13	0	6.37E+11
18	2016	10772	92	6.56E+12	5.74E+12	0	5.52E+11
19	2016	10793	92	4.44E+12	4.29E+12	0	2.23E+11
20	2016	10802	92	4.73E+13	4.34E+13	-1.5E+11	1.74E+12
…	…	…	…	…	…	…	…
5711	2019	85602	58	0	0	0	3.98E+08
5712	2019	86904	58	0	0	0	4.17E+09
5713	2019	95120	58	0	0	0	-7.3E+08

Setelah menentukan nilai y untuk masing-masing kelompok KLU, langkah selanjutnya adalah mencari rata-rata nilai variabel y selama 4 tahun. Untuk memudahkan dalam membaca angka maka seluruh data dibuat per 100,000. Menggunakan perintah sql pada database mysql:

```
SELECT AVG(y1)/100000 as jml_y1, AVG(y2)/100000 as jml_y2,
AVG(y3)/100000 as jml_y3, AVG(y4)/100000 as jml_y4 FROM sumbu_y
WHERE klu='$klu' AND kppadm='$kppadm'
```

Hasil dari pencarian rata-rata untuk seluruh variabel y dapat dilihat pada tabel 4-17.

Tabel 4-17 Rata-rata Variabel y Tahun 2016-2019

id_y_all	klu	kppadm	y1	y2	y3	y4
1	1262	92	1.56E+07	1.52E+07	-5.26E+05	1.23E+05
2	1262	51	2.10E+08	2.02E+08	-6.01E+05	6.40E+06
3	1262	57	0	0	0	1.81E+05
4	1262	58	0	0	-4.57E+06	1.16E+07
5	2117	92	2.32E+07	2.19E+07	-4.55E+04	9.79E+05
6	2117	91	0	1.54E+05	0	7.31E+02
7	2117	58	0	0	0	6.56E+04
8	10320	92	3.50E+07	3.11E+07	-1.59E+05	1.88E+06
9	10320	57	0	0	0	3.64E+03
10	10330	92	4.68E+07	4.05E+07	-3.39E+04	4.68E+06
11	10431	92	6.45E+08	6.24E+08	-2.42E+07	-1.03E+07
12	10431	58	0	0	-1.00E+06	1.34E+06
13	10432	92	1.57E+09	1.55E+09	-2.46E+07	-1.42E+06
14	10432	57	0	0	-4.64E+06	1.34E+06
15	10432	58	0	0	-3.73E+06	2.00E+05
16	10490	92	4.02E+08	3.92E+08	-6.02E+06	-1.91E+06
17	10520	92	4.36E+08	3.59E+08	-6.19E+05	5.60E+07
18	10520	56	0	0	-3.36E+03	1.54E+03
19	10520	57	0	0	-3.63E+03	3.60E+05
20	10590	92	0	0	0	0
…	…	…	…	…	…	…
1011	86904	58	0	0	0	1.62E+05

1012	95120	58	0	0	0	2.98E+04
1013	47511	56	0	0	0	1.27E+03

4.2.4 Normalisasi

Setelah dilakukan transformasi data, maka setiap data pada variabel x dan variabel y dinormalisasi menggunakan metode *z-score* sehingga menghasilkan data sebagai berikut:

Tabel 4-18 Hasil Normalisasi Variabel x

id_x_all	klu	kppadm	x1	x2	x3	x4	x5
1	1262	92	-0.350	-0.498	0.704	1.526	-0.322
2	1262	51	-0.350	-0.498	2.302	1.526	-0.322
3	2117	92	-0.350	-0.498	-0.539	-0.654	-0.322
4	2117	91	-0.350	-0.498	-0.539	-0.654	-0.322
5	10320	92	2.920	2.298	-0.539	1.526	-0.322
6	10330	92	-0.350	-0.498	-0.539	1.526	-0.322
7	10431	92	-0.350	-0.498	-0.539	-0.654	-0.322
8	10432	92	-0.350	0.123	2.125	1.526	-0.322
9	10490	92	-0.350	-0.498	-0.539	-0.654	-0.322
10	10520	92	-0.350	1.521	0.171	1.526	-0.322
11	10590	92	-0.350	-0.498	0.526	1.526	-0.322
12	10617	92	-0.350	-0.498	-0.539	-0.654	-0.322
13	10721	92	-0.350	-0.498	-0.539	1.526	-0.322
14	10721	51	-0.350	-0.498	-0.539	-0.654	-0.322
15	10732	92	-0.350	-0.498	-0.539	-0.654	-0.322
16	10740	92	-0.350	0.278	0.526	1.526	-0.322
17	10761	92	-0.350	-0.498	-0.539	1.526	0.773
18	10772	92	-0.350	-0.498	-0.539	-0.654	-0.322
19	10793	92	-0.350	-0.498	-0.539	-0.654	-0.322
20	10802	92	-0.350	-0.498	-0.539	-0.654	-0.322
…	…	…	…	…	…	…	…
1011	86904	58	-0.350	-0.498	-0.539	-0.654	-0.322
1012	95120	58	-0.350	-0.498	-0.539	-0.654	-0.322
1013	47511	56	-0.350	-0.498	-0.539	-0.654	-0.322

Tabel 4-19 Hasil Normalisasi Variabel y

id_y_all	klu	kppadm	y1	y2	y3	y4
1	1262	92	-0.131	-0.128	0.013	-0.165
2	1262	51	0.597	0.659	-0.016	0.088
3	1262	57	-0.189	-0.192	0.212	-0.162
4	1262	58	-0.189	-0.192	-1.518	0.297
5	2117	92	-0.102	-0.099	0.195	-0.130

6	2117	91	-0.189	-0.191	0.212	-0.169
7	2117	58	-0.189	-0.192	0.212	-0.167
8	10320	92	-0.058	-0.061	0.152	-0.094
9	10320	57	-0.189	-0.192	0.212	-0.169
10	10330	92	-0.014	-0.021	0.199	0.019
11	10431	92	2.219	2.437	-8.950	-0.585
12	10431	58	-0.189	-0.192	-0.168	-0.116
13	10432	92	5.664	6.324	-9.100	-0.227
14	10432	57	-0.189	-0.192	-1.545	-0.115
15	10432	58	-0.189	-0.192	-1.199	-0.161
16	10490	92	1.312	1.462	-2.067	-0.247
17	10520	92	1.439	1.321	-0.022	2.087
18	10520	56	-0.189	-0.192	0.211	-0.169
19	10520	57	-0.189	-0.192	0.211	-0.155
20	10590	92	-0.189	-0.192	0.212	-0.169
…	…	…	…	…	…	…
1011	86904	58	-0.189	-0.192	0.212	-0.163
1012	95120	58	-0.189	-0.192	0.212	-0.168
1013	47511	56	-0.189	-0.192	0.212	-0.169

4.2.5 Data Mining

Clustering dibuat dengan mengelompokkan KLU yang tidak memiliki resiko menyebabkan hilangnya penerimaan pajak sampai dengan KLU yang memiliki resiko tinggi dapat menyebabkan hilangnya penerimaan pajak. *Clustering* dikelompokan menjadi 2 kategori yakni tingkat ketidakpatuhan dan dampak fiskal yang disebabkan, masing-masing disimbolkan dengan variabel x dan variabel y. Variabel x terdiri dari beberapa variable nilai yakni, laporan SPT tahunan, peredaran usaha, dan riwayat pemeriksaan. Sedangkan variabel yang digunakan untuk *clustering* variabel y adalah data pembayaran. Selanjutnya KLU akan di *cluster* (kelompok) menjadi 3, yaitu risiko tinggi, sedang dan rendah.

Langkah-langkah yang ditempuh adalah menyusun data yang telah diolah ke dalam format yang dapat dibaca oleh program R dan di simpan dalam format comma separated file (csv). *Clustering* dilakukan menggunakan perangkat lunak R Studio version 1.4.1103.

Sebelum menentukan algoritma yang digunakan dalam pengelompokan, terlebih dahulu dilakukan perbandingan hasil clustering dengan algoritma K-Means, K-Medoids, dan Hierarchical Clustering. Perbandingan ini bertujuan untuk

memperoleh hasil *clustering* yang paling baik dilihat dari nilai *silhouette index* serta waktu komputasinya. *Average silhouette* memiliki rentang 1 hingga -1, jika semakin tinggi nilainya (mendekati 1) semakin baik hasil klasternya. Performa yang baik dari ketiga metode dilihat dari nilai *average silhouette* yang lebih tinggi. Sedangkan waktu komputasi dihitung untuk mengukur kinerja algoritma tersebut. Tabel 4-20 adalah *script* untuk melihat nilai *average silhouette* dari masing-masing metode.

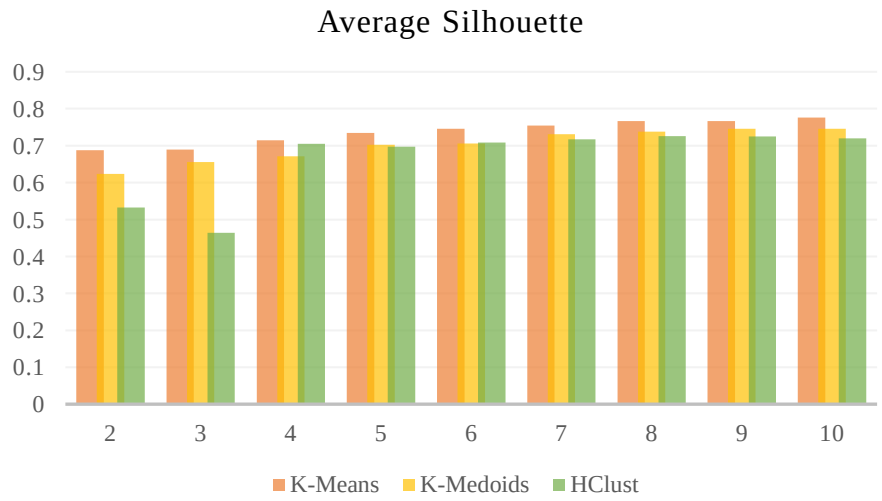
Tabel 4-20 *Script Average Silhouette* K-Means, K-Medoids, dan HClust

Baris	Script
1	install.packages("purrr")
2	library(purrr)
3	#--- Mendeskripsikan jumlah cluster 2 hingga 10 ------
4	k.values <- 2:10
5	#--- K-Means ----------
6	avg_sil_kmeans <- function(k) {
7	km.res <- kmeans(z, centers = k, nstart = 25)
8	ss <- silhouette(km.res$cluster, dist(z))
9	mean(ss[, 3]) }
10	sil_kmeans <- map_dbl(k.values, avg_sil_kmeans)
11	#--- K-Medoids (PAM) ----------
12	avg_sil_kmedoids <- function(k) {
13	km.res <- pam(z, k = k)
14	ss <- silhouette(km.res$clustering, dist(z))
15	mean(ss[, 3]) }
16	sil_kmedoids <- map_dbl(k.values, avg_sil_kmedoids)
17	#--- Hierarchical Clustering----------
18	avg_sil_hclust <- function(k) {
19	km.res <- hclust(dist(z))
20	grp <- cutree(km.res, k = k)
21	ss <- silhouette(grp, dist(z))
22	mean(ss[, 3]) }
23	sil_hclust <- map_dbl(k.values, avg_sil_hclust)
24	#--- Melihat hasil Silhouette semua metode ----------
25	sil_all=data.frame(klaster=k.values,Kmeans=sil_kmeans,
26	KMedoids=sil_kmedoids, HClust=sil_hclust)
27	sil_all

Tabel 4-21 menunjukkan nilai average silhouette untuk algoritma K-Means, K-Medoids, dan hierarchical clustering. Nilai tersebut kemudian disajikan dalam bentuk grafik seperti pada gambar 4-8 untuk memudahkan dalam melihat perbandingan pada ketiga metode tersebut.

Tabel 4-21 *Average Silhouette* K-Means, K-Medoids, dan HClust

Cluster	**K-Means**	**K-Medoids**	**HClust**
2	0.687611	0.623152	0.532923
3	0.689323	0.655634	0.464103
4	0.714459	0.671566	0.705375
5	0.734269	0.702591	0.697075
6	0.745868	0.706111	0.708834
7	0.754336	0.731161	0.717547
8	0.766574	0.73756	0.725668
9	0.766244	0.746076	0.724637
10	0.775706	0.746128	0.719438



Gambar 4-8 Grafik *Average Silhouette* K-Means, K-Medoids, dan HClust

Selanjutnya untuk menghitung waktu komputasi masing-masing algoritma digunakan *script* seperti pada tabel 4-22. *Script* ini diletakkan di awal dan akhir dari *script clustering* masing-masing metode. Didapatkan waktu komputasi yang disajikan dalam tabel 4-23. Gambar 4-9 adalah grafik dari waktu komputasi yang dibutuhkan masing-masing metode dalam melakukan proses *clustering*, grafik ditampilkan untuk memudahkan dalam melihat perbandingan ketiga metode *clustering*.

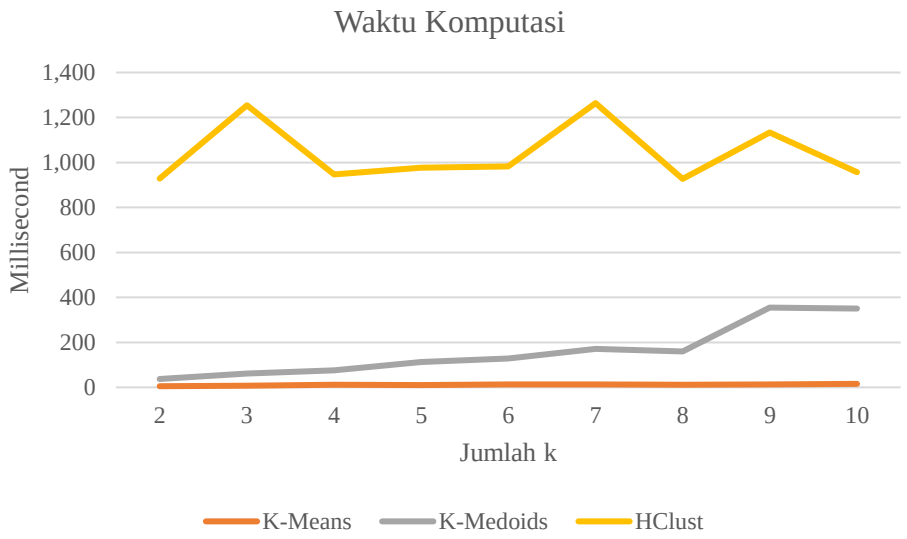
Tabel 4-22 Script Waktu Komputasi K-Means, K-Medoids, dan HClust

Baris	**Script**
1	• start_time <- Sys.time()
2	• end_time <- Sys.time()

3	•	time <- end_time – start_time

Tabel 4-23 Waktu Komputasi K-Means, K-Medoids, dan HClust

Cluster	**K-Means**	**K-Medoids**	**HClust**
			(dalam millisecond)
2	5	37	929
3	8	61	1255
4	12	76	947
5	10	113	976
6	13	128	983
7	13	171	1264
8	12	159	927
9	14	355	1134
10	16	351	957



Gambar 4-9 Grafik Waktu Komputasi K-Means, K-Medoids, dan HClust

Dari tabel 4-21 dapat dilihat bahwa nilai *average silhouette* yang dihasilkan dari pengelompokan menggunakan metode K-Means mempunyai nilai lebih tinggi dibandingkan dua metode lainnya. Demikian dengan waktu komputasi yang diperlukan oleh K-Means untuk melakukan pengelompokan dengan jumlah *cluster* k=2 hingga k=10 cenderung stabil dan relatif lebih cepat seperti ditunjukkan pada gambar 4-9.

Hasil perbandingan dari ketiga metode *clustering* tersebut menjadi dasar pertimbangan penelitian ini menggunakan algoritma K-Means dalam proses pengelompokan WP badan per kelompok KLU.

4.2.5.1 *Clustering* Variabel x

Script untuk analisis *cluster* dapat dilihat pada tabel 4-24. Terdapat 36 buah baris program, dimana 11 barisnya merupakan keterangan penggunaan. Baris pertama sampai ketiga pada program digunakan untuk menginstall *packages* yang dibutuhkan dalam proses *clustering*. *Package* yang dipakai yakni ggplot2 dan factoextra. Program kemudian diarahkan ke direktori tempat penyimpanan data, mengambil data dan disimpan dengan nama 'data' (baris 5). Data yang diambil kemudian ditampilkan (baris 7 dan 8). Tabel 4-25 adalah tampilan enam baris pertama data awal variabel x, sedangkan tabel 4-26 adalah tampilan enam baris terakhir data awal variabel x. Jumlah enam baris data yang ditampilkan ini adalah jumlah *default*.

Setelah data berhasil disimpan, ditentukan kolom yang akan digunakan dalam proses *clustering*. Pada *clustering* variabel x, kolom yang digunakan adalah kolom 4 sampai dengan 8, maka kolom 1, 2, dan 3 dihapus dari data (baris 10). Kolom yang telah dipilih kemudian disimpan dengan nama 'z'. Setelah dipastikan data yang diambil sudah sesuai, dilakukan proses normalisasi terhadap data. Normalisasi dilakukan menggunakan metode *z-score*. Untuk mendapatkan nilai x baru dalam normalisasi, terlebih dahulu dicari nilai rata-rata atribut (baris 12) dan standar deviasi nya (baris 13). Setelah diketahui nilai rata-rata dan standar deviasinya, dilakukan normalisasi menggunakan fitur *scale* pada R seperti pada baris 14. Data yang sudah dilakukan normalisasi kemudian disimpan dengan nama 'nor'.

Langkah selanjutnya adalah *clustering*. *Clustering* dilakukan dengan menggunakan metode K-Means, dengan jumlah pusat *cluster* k = 3. Proses ini dijalankan oleh baris program 17. Pada tahap ini inisiasi atau titik pusat *cluster* dipilih secara acak oleh perangkat lunak R. Setelah melakukan *clustering*, baris program 18 menampilkan hasil *clustering* yang dapat dilihat pada tabel 4-27.

Berikutnya baris 21 dijalankan untuk memvisualisasikan hasil *clustering*. Hasil *cluster* yang telah diinterpretasikan ditampilkan pada Gambar 4-10. Baris 33-36 merupakan langkah terakhir, berfungsi untuk mengekspor hasil *clustering* ke dalam file csv. Tabel hasil *clustering* yang telah di ekspor dari pemrograman Bahasa R dapat dilihat pada Tabel 4-28.

Tabel 4-24 *Script Clustering* R

Baris	Script
1	install.packages("ggplot2")
2	install.packages("factoextra")
3	library(ggplot2)
4	#--- Mengambil Data ------------------------------------
5	data=read.csv(file.choose(), header=TRUE)
6	#--- Menampilkan Data ----------------------------------
7	head(data)
8	tail(data)
9	#--- Menghapus Kolom 1, 2, dan 3 ------------------------
10	z <- data[,-c(1,2,3)]
11	#--- Normalisasi menggunakan Z-score ----------------------
12	m <- apply(z, 2, mean)
13	sds <- apply(z, 2, sd)
14	nor <- scale(z,m,scale = sds)
15	#--- Melakukan proses clustering dengan jumlah k = 3 -----
16	set.seed(123)
17	km <- kmeans(nor, 3)
18	#--- Menampilkan hasil clustering ----------------------
19	km
20	#--- Memvisualkan hasil clustering ----------------------
21	clusplot(data, km$cluster, color=T, shade=T)
22	#--- Mengembalikan nilai normalisasi ke nilai sebenarnya -
23	data[,4:8]%>%
24	mutate(Cluster=km$cluster)%>%
25	group_by(Cluster)%>%
26	summarise_all("mean")
27	#--- Melihat grafik wss ----------------------------------
28	fviz_nbclust(data, kmeans, method="wss")
29	#--- Melihat nilai silhouette setiap cluster -------------
30	distance <- dist(nor)
31	print(distance, digits=3)
32	plot(silhouette(km$cluster, distance))
33	#--- Meng-export hasil clustering ----------------------
34	sp=data.frame(data,km$cluster)
35	View(sp)
36	write.csv(sp, "D:\\1_KULIAH\\SEMESTER 4\\sumbu_x_nor_cluster.csv")

Tabel 4-25 Enam Baris Pertama Data Variabel x

id_x	klu	kppadm	x1	x2	x3	x4	x5
1	1262	92	0	0	43.75	100	0
2	1262	51	0	0	100	100	0
3	2117	92	100	90	0	100	0
4	2117	91	0	0	0	50	0
5	10320	92	0	20	93.75	100	0
6	10330	92	0	65	25	100	0

Tabel 4-26 Enam Baris Terakhir Data Variabel x

id_x	klu	kppadm	x1	x2	x3	x4	x5
1008	85491	58	0	0	0	0	0
1009	85493	58	0	0	0	0	0
1010	85602	58	0	0	0	0	0
1011	86904	58	0	0	0	0	0
1012	95120	58	0	0	0	0	0
1013	47511	56	0	0	0	0	0

Tabel 4-27 Hasil *Clustering* Variabel x pada R

K-means clustering with 3 clusters of sizes 173, 684, 156

Clustering vector:

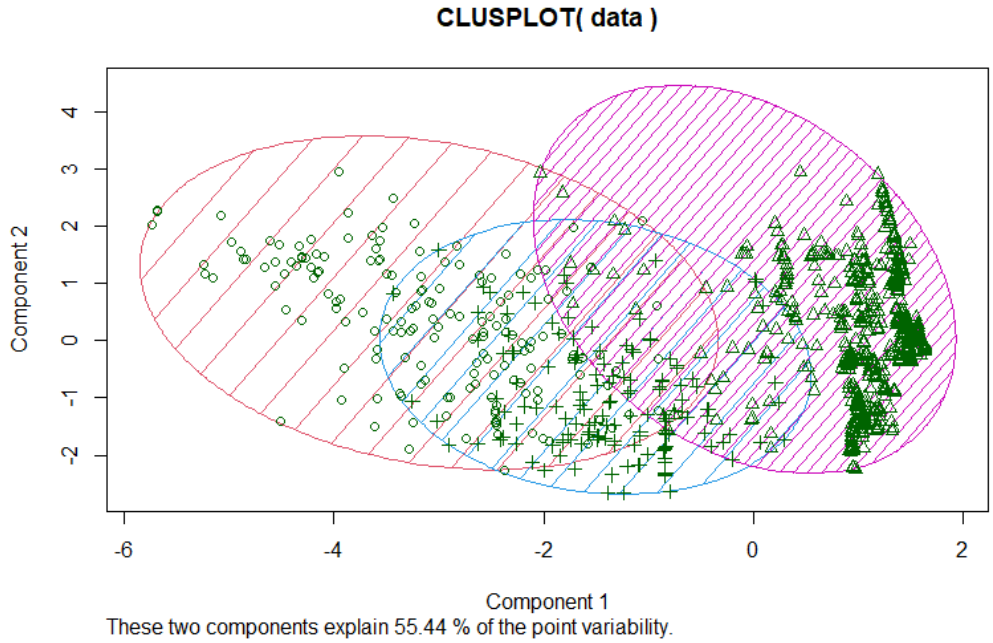
[1] 3 3 2 2 3 3 2 1 2 3 3 2 3 2 2 3 3 2 2 2 3 2 1 1 2 3 2 3 2 3 2 3 2 2 2 3 3 2 1 2
[41] 3 2 1 3 2 3 3 2 3 2 2 2 2 2 3 3 2 2 3 2 3 3 2 2 2 1 2 2 2 1 1 1 2 2 2 3 2 2 3 2
[81] 3 2 3 2 3 2 2 2 1 2 3 3 2 3 2 3 2 2 3 2 2 2 2 3 1 2 2 3 3 2 3 2 3 3 2 2 3 2 3 2
[121] 3 3 2 3 2 1 2 1 1 2 3 3 2 3 3 2 3 3 2 2 3 3 2 3 2 3 2 3 2 2 3 2 2 3 2 3 2 1 2 3
[161] 2 2 2 3 2 3 2 3 2 1 2 3 2 3 2 2 3 2 3 2 2 1 2 2 2 1 1 2 2 1 3 3 2 2 3 2 3 2 3 3
[201] 1 1 2 2 2 3 3 2 3 2 3 2 2 2 2 3 2 2 2 2 3 3 2 2 1 2 2 2 3 2 3 2 2 2 3 3 1 2 2 3
[241] 3 2 2 2 2 2 2 3 1 2 3 3 1 2 2 2 2 2 2 3 2 1 1 2 2 1 1 2 3 3 3 2 1 1 1 2 3 1 2 2 3
[281] 3 2 3 3 3 2 1 2 1 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 3 2 1 2 3 1 1 2 2 1 2
[321] 2 1 1 2 2 1 1 2 2 1 3 3 3 3 3 1 1 3 2 1 3 1 3 3 1 2 1 2 2 1 3 2 2 2 2 2 1 2 3 2
[361] 3 1 2 3 3 2 1 3 2 3 2 3 2 2 1 2 2 2 3 3 2 3 2 1 2 1 2 1 3 2 2 2 1 2 2 2 2 1 2 2
[401] 2 1 2 2 3 1 1 2 2 2 1 2 2 3 1 2 2 3 3 2 1 1 2 1 1 1 2 1 2 1 1 3 2 1 1 2 1 2 3 3
[441] 1 1 1 1 1 1 2 2 2 3 2 2 2 2 3 2 1 3 3 1 2 1 2 2 1 1 2 2 2 1 1 2 1 1 2 2 1 1 1 1
[481] 2 2 1 2 2 1 3 2 2 2 1 2 1 1 2 2 3 1 2 1 2 2 1 1 2 2 2 3 2 3 1 2 2 2 2 1 2 1 3 3
[521] 2 1 1 2 1 1 2 2 1 2 3 2 2 3 3 3 2 1 2 3 2 2 2 2 2 2 1 3 3 2 2 2 3 2 2 1 2 1 1 1
[561] 2 1 3 1 1 2 1 1 2 1 2 3 2 2 2 1 2 1 2 2 2 1 1 2 1 2 2 2 1 3 2 2 3 2 2 2 1 1 2 1 2
[601] 2 1 2 2 2 2 2 1 1 2 2 1 2 2 1 1 2 1 1 2 1 2 1 1 2 3 1 2 3 2 1 1 2 1 1 2 2 2 1 1 2
[641] 1 2 2 1 1 1 1 1 1 2 2 1 1 1 1 1 1 2 3 2 2 1 2 2 2 2 2 2 2 1 2 2 1 2 1 2 2 2 2 2
[681] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[721] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[761] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[801] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[841] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[881] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

```
 [921] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [961] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [ reached getOption("max.print") -- omitted 13 entries ]

Within cluster sum of squares by cluster:
[1] 1084075.5  423987.9  535048.0
 (between_SS / total_SS =  65.4 %)
```

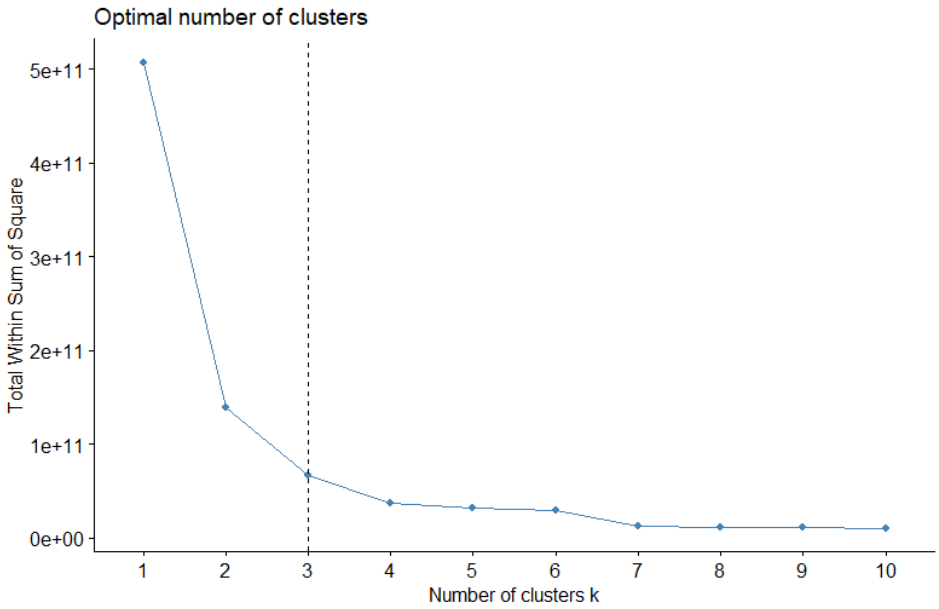
Tabel 4-28 Hasil *Clustering* Variabel x

id_x_all	klu	kppadm	x1	x2	x3	x4	x5	C
1	1262	92	0	0	43.75	100	0	3
2	1262	51	0	0	100	100	0	3
3	1262	57	0	0	0	0	0	2
4	1262	58	0	0	0	0	0	2
5	2117	92	100	90	0	100	0	3
6	2117	91	0	0	0	100	0	3
7	2117	58	0	0	0	0	0	2
8	10320	92	0	20	93.75	100	0	1
9	10320	57	0	0	0	0	0	2
10	10330	92	0	65	25	100	0	3
11	10431	92	0	0	37.5	100	0	3
12	10431	58	0	0	0	0	0	2
13	10432	92	0	0	0	100	0	3
14	10432	57	0	0	0	0	0	2
15	10432	58	0	0	0	0	0	2
16	10490	92	0	25	37.5	100	0	3
17	10520	92	0	0	0	100	25	3
18	10520	56	0	0	0	0	0	2
19	10520	57	0	0	0	0	0	2
20	10590	92	0	0	0	0	0	2
…	…	…	…	…	…	…	…	…
1011	86904	58	0	0	0	0	0	2
1012	95120	58	0	0	0	0	0	2
1013	47511	56	0	0	0	0	0	2



Gambar 4-10 Visualisasi *Clustering* Variabel x

Setiap data pada masing-masing kategori baik variabel x maupun variabel y di-*cluster* menjadi 3 kelompok yakni kelompok risiko tinggi, sedang dan rendah. Dapat dilihat pada tabel 4-27 dari keluaran hasil K-Means dengan k = 3, terbentuk *cluster* 1 sebanyak 173 KLU, *cluster* 2 sebanyak 684 KLU, dan *cluster* 3 sebanyak 156 KLU dengan nilai WSS 65.4%. Dengan perintah pada baris program 28, ditampilkan grafik *elbow* untuk variabel x seperti terlihat pada gambar 4-11. Pada plot tersebut didapatkan titik siku yang terbentuk diantara titik dua dan empat, setelah titik 3 sudah tidak lagi terjadi penurunan yang signifikan. Sebagai perbandingan dapat dilihat juga pada tabel 4-29 penurunan selisih nilai SSE yang paling signifikan terletak pada nilai k=3 dengan selisih SSE sebesar 665,088. Selanjutnya hasil SSE dari nilai k turun secara perlahan-lahan dan melandai. Sehingga dapat disimpulkan bahwa jumlah *cluster* k=3 menurut metode *elbow* adalah optimal.



Gambar 4-11 Grafik *Elbow* Variabel x

Tabel 4-29 Perbandingan SSE Variabel x

Nilai K	SSE	Selisih SSE
2	2,708,199	-
3	2,043,111	665,088
4	1,767,081	276,030
5	1,368,813	398,268
6	1,267,515	101,298
7	1,029,626	237,889
8	947,511	82,115
9	884,101	63,409
10	850,746	33,356

Untuk merepresentasikan karakteristik tiap *cluster* dapat menggunakan acuan nilai *means* untuk tiap kelompok yang terbentuk. *Script* program yang digunakan seperti pada tabel 4-30. *Script* ini juga digunakan untuk mengembalikan data normalisasi sesuai bentuk awal. Keluaran yang dihasilkan dapat dilihat pada tabel 4-31.

Tabel 4-30 *Script* R Menampilkan *Means*

Baris	Script
23	data[,4:8]%>%
24	mutate(Cluster=km$cluster)%>%
25	group_by(Cluster)%>%
26	summarise_all("mean")

Tabel 4-31 *Means Cluster* Variabel x

Cluster	**x1**	**x2**	**x3**	**x4**	**x5**
1	30.6	75.6	84.5	85.5	29.9
2	4.2	1.2	1.3	0.0	1.0
3	17.3	15.3	23.7	100.0	10.1

Berdasarkan hasil pada tabel 4-31, maka dapat dilakukan profilisasi tiap kelompok yang terbentuk. Dimana pada *cluster* 1 merupakan KLU yang memiliki tingkat kepatuhan paling rendah dari *cluster* yg lain dengan nilai rata-rata x1, x2, x3, dan x5 masing-masing sebesar 30.6, 75.6, 84.5, dan 29.9. Sedangkan *cluster* 2 menjadi kelompok dengan tingkat kepatuhan paling tinggi.

4.2.5.2 *Clustering* Variabel y

Setelah tahap *clustering* pada variabel x selesai, pengolahan yang sama juga diterapkan terhadap variabel y. Pada tahap transformasi data variabel y, semua nilai variabel telah dibagi 100.000 untuk memudahkan dalam pembacaan. Tabel 4-32 adalah hasil dari tampilan enam baris pertama dari data variabel y, sedangkan tabel 4-33 adalah tampilan enam baris terakhir dari data variabel y. Setelah data berhasil disimpan, ditentukan kolom yang akan digunakan dalam proses *clustering*, untuk variabel y kolom yang digunakan adalah kolom 4 sampai dengan 7. Setelah dipastikan data yang diambil sudah sesuai, dilakukan proses *clustering* terhadap data. Hasil *clustering* dapat dilihat pada tabel 4-34 atau tabel 4-35 untuk hasil *clustering* yang telah di ekspor. *Cluster* juga diinterpretasikan menggunakan clusplot yang ditampilkan pada Gambar 4-12.

Tabel 4-32 Enam Baris Pertama Data Variabel y

id_y	**klu**	**kppadm**	**y1**	**y2**	**y3**	**y4**
1	1262	92	15,553,200	15,189,600	-526,461	122,907
2	1262	51	210,386,250	201,806,000	-601,259	6,395,218
3	1262	57	0	0	0	181,224
4	1262	58	0	0	-4,570,267	11,572,470
5	2117	92	23,208,800	21,939,175	-45,460	979,177
6	2117	91	0	153,961	0	731

Tabel 4-33 Enam Baris Terakhir Data Variabel y

id_y	klu	kppadm	y1	y2	y3	y4
1008	85491	58	0	0	0	0
1009	85493	58	0	0	-11,588	97,857
1010	85602	58	0	0	0	1,013
1011	86904	58	0	0	0	162,427
1012	95120	58	0	0	0	29,762
1013	47511	56	0	0	0	1,269

Tabel 4-34 Hasil *Clustering* Variabel y pada R

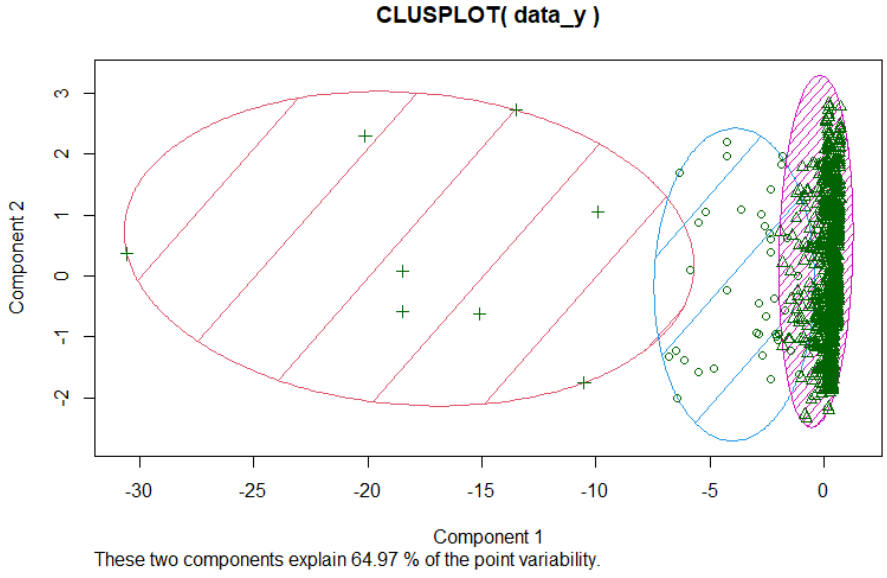
K-means clustering with 3 clusters of sizes 38, 967, 8

Clustering vector:
[1] 2 2 2 2 2 2 2 2 2 2 1 2 3 2 2 1 1 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 1 2 2 2
[41] 1 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[81] 2 2 2 2 2 2 2 2 2 2 2 1 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2
[121] 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2
[161] 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 1 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 3 2 2 2 2 2
[201] 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2
[241] 2 2 2 2 2 2 2 2 2 1 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2
[281] 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[321] 2 2 2 2 2 1 3 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2
[361] 2 2 2 2 3 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[401] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2
[441] 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[481] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[521] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 3 2 1 3 2
[561] 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2
[601] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[641] 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[681] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[721] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[761] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[801] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[841] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[881] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[921] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[961] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[reached getOption("max.print") -- omitted 13 entries]

Within cluster sum of squares by cluster:
[1] 341.9758 327.2899 399.9765
(between_SS / total_SS = 86.8 %)

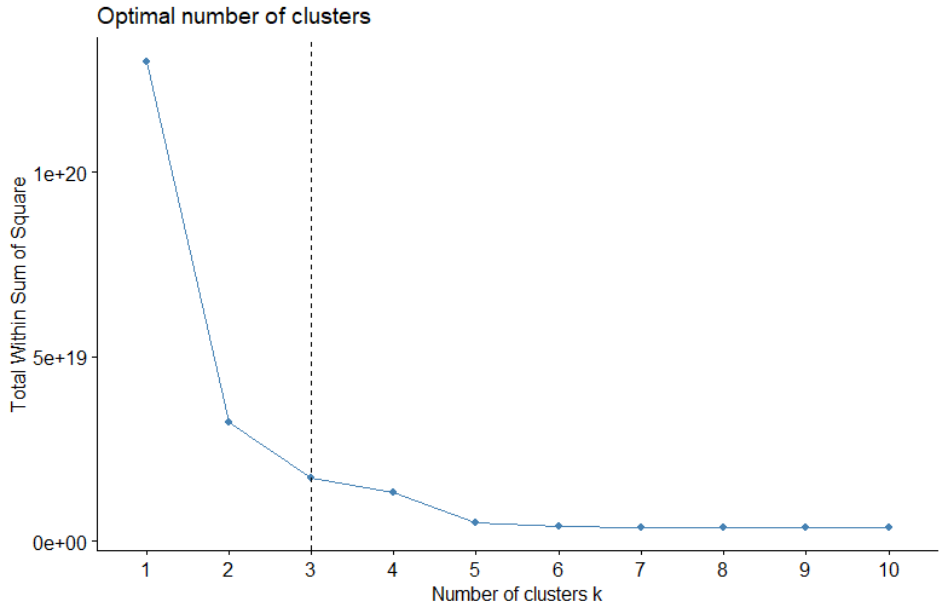
Tabel 4-35 Hasil *Clustering* Variabel y

id_ y_all	klu	kppadm	y1	y2	y3	y4	C
1	1262	92	1.56E+07	1.52E+07	-5.26E+05	1.23E+05	2
2	1262	51	2.10E+08	2.02E+08	-6.01E+05	6.40E+06	2
3	1262	57	0	0	0	1.81E+05	2
4	1262	58	0	0	-4.57E+06	1.16E+07	2
5	2117	92	2.32E+07	2.19E+07	-4.55E+04	9.79E+05	2
6	2117	91	0	1.54E+05	0	7.31E+02	2
7	2117	58	0	0	0	6.56E+04	2
8	10320	92	3.50E+07	3.11E+07	-1.59E+05	1.88E+06	2
9	10320	57	0	0	0	3.64E+03	2
10	10330	92	4.68E+07	4.05E+07	-3.39E+04	4.68E+06	2
11	10431	92	6.45E+08	6.24E+08	-2.42E+07	-1.03E+07	1
12	10431	58	0	0	-1.00E+06	1.34E+06	2
13	10432	92	1.57E+09	1.55E+09	-2.46E+07	-1.42E+06	3
14	10432	57	0	0	-4.64E+06	1.34E+06	2
15	10432	58	0	0	-3.73E+06	2.00E+05	2
16	10490	92	4.02E+08	3.92E+08	-6.02E+06	-1.91E+06	1
17	10520	92	4.36E+08	3.59E+08	-6.19E+05	5.60E+07	1
18	10520	56	0	0	-3.36E+03	1.54E+03	2
19	10520	57	0	0	-3.63E+03	3.60E+05	2
20	10590	92	0	0	0	0	2
…	…	…	…	…	…	…	…
1011	86904	58	0	0	0	1.62E+05	2
1012	95120	58	0	0	0	2.98E+04	2
1013	47511	56	0	0	0	1.27E+03	2



Gambar 4-12 Visualisasi *Clustering* Variabel y

Dapat dilihat pada tabel 4-34 dari keluaran hasil K-Means dengan k = 3, terbentuk *cluster* 1 sebanyak 38 KLU, *cluster* 2 sebanyak 967 KLU, dan *cluster* 3 sebanyak 8 KLU dengan nilai WSS 86.8%. Sama halnya seperti pada variabel x, ditampilkan pula grafik *elbow* variabel y seperti terlihat pada gambar 4-13. Pada plot tersebut didapatkan titik siku yang terbentuk diantara titik dua dan empat, setelah titik 3 sudah tidak lagi terjadi penurunan yang signifikan. Dapat dilihat pada tabel 4-36 penurunan selisih nilai SSE yang signifikan terletak pada nilai k=3 dengan selisih SSE sebesar 148,792,800. Sehingga dapat disimpulkan bahwa jumlah *cluster* k=3 pada variabel y menurut metode *elbow* adalah optimal.



Gambar 4-13 Grafik *Elbow* Variabel y

Tabel 4-36 Perbandingan SSE Variabel y

Nilai K	SSE	Selisih SSE
2	320,682,000	-
3	171,889,200	148,792,800
4	130,226,300	41,662,900
5	50,926,860	79,299,440
6	40,886,620	10,040,240
7	37,494,490	3,392,130
8	36,039,930	1,454,560
9	25,911,410	10,128,520
10	35,997,050	(10,085,640)

Untuk merepresentasikan karakteristik tiap *cluster*, acuan nilai means tiap kelompok yang terbentuk pada variabel y dapat dilihat pada tabel 4-37.

Tabel 4-37 Means *Cluster* Variabel y

Cluster	**y1**	**y2**	**y3**	**y4**
1	484,940,872	437,867,845	(5,721,710)	31,008,580
2	11,813,843	11,025,575	(219,675)	1,456,692
3	2,662,986,250	2,346,231,686	(17,194,026)	208,821,749

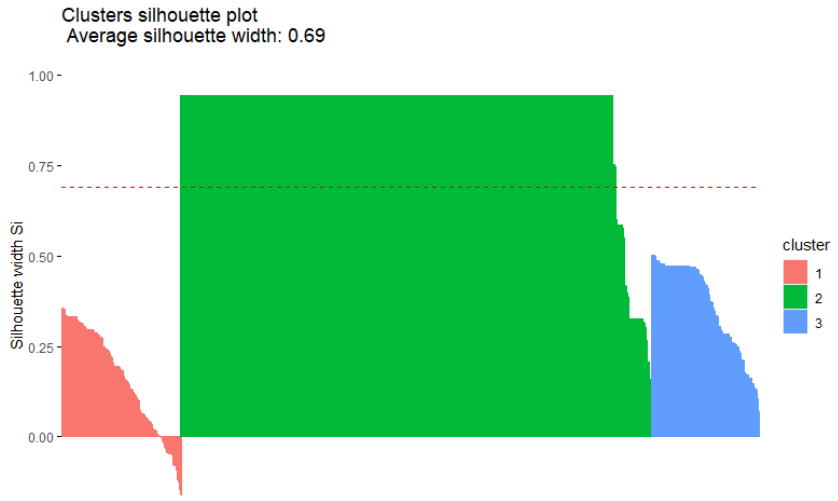
Berdasarkan hasil pada tabel 4-37, *cluster* 2 merupakan KLU yang memiliki dampak fiskal paling rendah dari *cluster* yg lain dengan nilai rata-rata y1, y2, y3, dan y4 masing-masing sebesar 11,813,843; 11,025,575; -219,675; dan 1,456,692. Sedangkan *cluster* 3 menjadi kelompok yang memiliki dampak fiskal paling tinggi.

4.2.5.3 *Silhouette Index*

Untuk mendapatkan nilai koefisien *silhouette*, dijalankan *script* pada program R seperti pada tabel 4-38. *Silhouette* adalah fitur R untuk melihat nilai koefisien *silhouette*. Grafik yang dihasilkan dapat dilihat pada gambar 4-14, dengan masing-masing nilai s*ilhouette* seperti pada tabel 4-39.

Tabel 4-38 Script R Mencari *Silhouette*

Baris	**Script**
30	• distance <- dist(nor)
31	• print(distance, digits=3)
32	• fviz_silhouette(silhouette(km$cluster, distance))



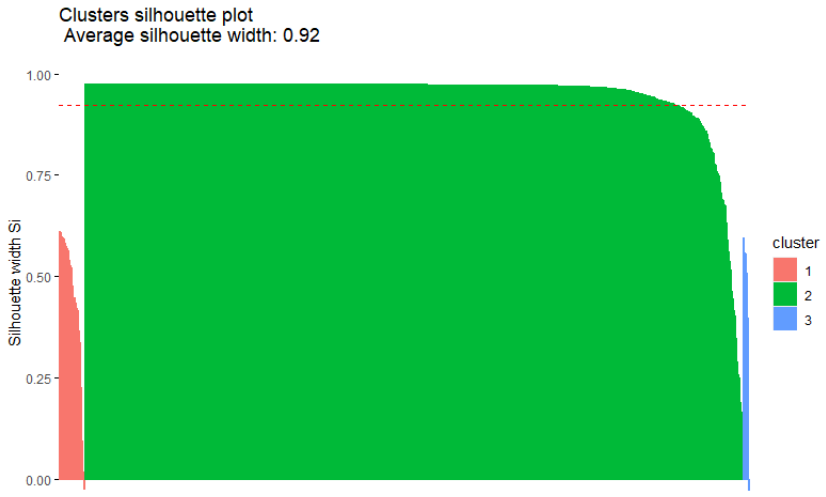
Gambar 4-14 Plot *Silhouette* Variabel x

Tabel 4-39 Nilai *Silhouette Cluster* Variabel x

Cluster	**Jumlah Anggota *Cluster***	**Nilai *Silhouette***
1	173	0.16
2	684	0.90
3	156	0.36
Average silhouette coefficient		0.69

Setelah didapat hasil dari nilai *silhouette* masing-masing *cluster*, diambil nilai rata-rata yang digunakan sebagai nilai *silhouette coeficient*. Dari hasil *clustering* menggunakan *euclidean distance* pada metode K-Means, nilai *average silhouette coefficient* variabel x untuk jumlah *cluster* k = 3 adalah 0.69.

Langkah yang sama dilakukan terhadap variabel y, grafik *silhouette* yang dihasilkan dapat dilihat pada gambar 4-15, dengan masing-masing nilai *silhouette* seperti pada tabel 4-40.



Gambar 4-15 Plot *Silhouette* Variabel y

Tabel 4-40 Nilai *Silhouette Cluster* Variabel y

Cluster	**Jumlah Anggota *Cluster***	**Nilai *Silhouette***
1	38	0.44
2	967	0.94
3	8	0.44
Average silhouette coefficient		0.92

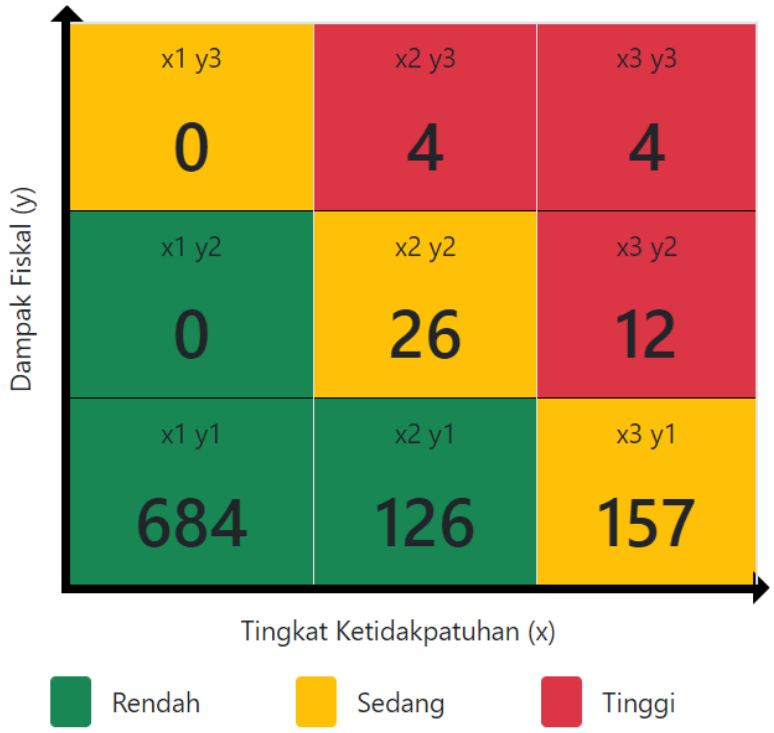
Setelah diketahui nilai *silhouette* dari masing-masing *cluster*, kemudian diambil rata-ratanya dan digunakan sebagai nilai *silhouette coeficient*. Maka, nilai *average silhouette coefficient* variabel y untuk jumlah *cluster* k = 3 pada metode K-Means *Clustering* adalah 0.92.

4.2.6 Visualisasi

Interpretasi penggabungan *cluster* kedua variabel x dan y dilakukan untuk mempermudah dalam memahami hasil *cluster*. Digunakan bahasa pemrograman php untuk membuat kuadran x dan y. Kuadran terlebih dahulu dibagi menjadi 9 bagian, kemudian hasil *cluster* yang telah digabungkan dimasukkan ke dalam kuadran yang sesuai, dengan variabel x sebagai koordinat horizontal, dan variabel y sebagai koordinat vertikal. Penjelasan dari masing-masing kuadran adalah sebagai berikut:

1. Kuadran x1−y1, berwarna hijau, terdapat 684 KLU dengan tingkat kepatuhan tinggi dan dampak fiskal rendah
2. Kuadran x1−y2, berwarna hijau, terdapat 0 KLU dengan tingkat kepatuhan tinggi dan dampak fiskal sedang
3. Kuadran x1−y3, berwarna kuning, terdapat 0 KLU dengan tingkat kepatuhan tinggi dan dampak fiskal tinggi
4. Kuadran x2−y1, berwarna hijau, terdapat 126 KLU dengan tingkat kepatuhan sedang dan dampak fiskal rendah
5. Kuadran x2−y2, berwarna kuning, terdapat 26 KLU dengan tingkat kepatuhan sedang dan dampak fiskal sedang
6. Kuadran x2−y3, berwarna merah, terdapat 4 KLU dengan tingkat kepatuhan sedang dan dampak fiskal tinggi
7. Kuadran x3−y1, berwarna kuning, terdapat 157 KLU dengan tingkat kepatuhan rendah dan dampak fiskal rendah
8. Kuadran x3−y2, berwarna merah, terdapat 12 KLU dengan tingkat kepatuhan rendah dan dampak fiskal sedang
9. Kuadran x3−y3, berwarna merah, terdapat 4 KLU dengan tingkat kepatuhan rendah dan dampak fiskal tinggi

Kemudian hasil *cluster* ditempatkan di kuadran yang sesuai. Maka didapatkan jumlah KLU pada masing-masing kuadran seperti ditunjukkan pada Gambar 4-16.



Gambar 4-16 Visualisasi Kuadran Variabel x dan y

4.2.7 Analisis

Berdasarkan hasil clustering variabel x dan y diketahui terdapat 20 sektor usaha penyumbang penerimaan terbesar yang terindikasi tidak patuh dalam menjalankan kewajiban perpajakannya (berada di kuadran merah x2-y3, x3-y3, dan x3-y2) dan mempunyai dampak fiskal tinggi terhadap penerimaan negara. Variabel x dapat memberi informasi KLU yang melakukan *underreporting* maupun *nonfilling*. Sedangkan variabel y digunakan sebagai acuan dalam melihat *underpayment* dari masing-masing KLU.

Underreporting adalah pelaporan pajak dalam SPT di bawah dari seharusnya yang dapat dilakukan oleh KLU dengan cara mengurangi jumlah perolehan PPN dan penyerahan PPN, atau dengan tidak melaporkan nilai peredaran

usaha yang sebenarnya. Indikasi ini direpresentasikan oleh variabel x2 dan x3. *Underreporting* dapat juga dilakukan dengan melaporkan biaya lain-lain yang tidak wajar yang direpresentasikan oleh variabel x5. Sedangkan *nonfilling* terjadi karena WP tidak menyampaikan laporan SPT nya. Indikasi ini dapat dilihat dari kepatuhan menyampaikan laporan SPT setiap tahunnya sebagai kepatuhan administrative yang direpresentasikan oleh variabel x1.

Dengan menyandingkan hasil *clustering* dan tabel skoring variabel x dan y, dapat dilakukan analisis sebagai berikut:

A. Kuadran x3-y3

1. Sektor Pembangkitan Tenaga Listrik (kode KLU: 35101)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN.
2. Sektor Pertambangan Batu Bara (kode KLU: 5101)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN.
3. Sektor Bank Pemerintah/Bumn/Persero (kode KLU: 64121)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Serta memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar, Selain itu, KLU tidak melaksanakan kepatuhan formalnya berupa pelaporan SPT Tahunan Badan selama 2 tahun berturut-turut.
4. Sektor Bank Umum Swasta Nasional Devisa (kode KLU: 64125)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Serta memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar, Selain itu, KLU tidak melaksanakan kepatuhan formalnya berupa pelaporan SPT Tahunan Badan selama 1 tahun.

B. Kuadran x2-y3

1. Sektor Industri Minyak Goreng Kelapa Sawit (kode KLU: 10432)
   Terdapat surat ketetapan terhadap KLU dari riwayat pemeriksaan sebelumnya.
2. Sektor Industri Kendaraan Bermotor Roda Empat atau Lebih (kode KLU: 29100)

Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN.

3. Sektor Perdagangan Besar Mobil Baru (kode KLU: 45101)
   Terdapat surat ketetapan terhadap KLU dari riwayat pemeriksaan sebelumnya. Selain itu, KLU tidak melaksanakan kepatuhan formalnya berupa pelaporan SPT Tahunan Badan selama 2 tahun berturut-turut.
4. Sektor Industri Pemurnian dan Pengilangan Minyak Bumi (kode KLU: 19211)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Serta memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar,

C. Kuadran x3-y2

1. Sektor Industri Semen (kode KLU: 23941)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Serta memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar,
2. Sektor Telekomunikasi Tanpa Kabel (kode KLU: 61200)
   Terindikasi melakukan *underreporting* pada perolehan PPN. Terdapat surat ketetapan terhadap KLU dari riwayat pemeriksaan sebelumnya. Selain itu, KLU tidak melaksanakan kepatuhan formalnya berupa pelaporan SPT Tahunan Badan selama 2 tahun berturut-turut.
3. Sektor Pertambangan Batu Bara (kode KLU: 5101)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN.
4. Sektor Industri Pupuk Buatan Tunggal Hara Makro Primer (kode KLU: 20122)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Serta memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar,
5. Sektor Distribusi Gas Alam dan Buatan (kode KLU: 35202)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN.
6. Sektor Perdagangan Besar Beras (kode KLU: 46311)

Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN.

7. Sektor Bank Campuran dan Asing (kode KLU: 64124)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Serta memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar,
8. Sektor Pembiayaan Konsumen (Consumers Credit) (kode KLU: 64922)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Selain itu, KLU tidak melaksanakan kepatuhan formalnya berupa pelaporan SPT Tahunan Badan selama 2 tahun berturut-turut.
9. Sektor Konstruksi Gedung Perkantoran (kode KLU: 41012)
   Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Selain itu, KLU tidak melaksanakan kepatuhan formalnya berupa pelaporan SPT Tahunan Badan selama 2 tahun berturut-turut.
10. Sektor Angkutan Udara Berjadwal Domestik Umum untuk Penumpang (kode KLU: 51101)
    Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Serta memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar.
11. Sektor Bank Sentral (kode KLU: 64110)
    Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Selain itu, KLU tidak melaksanakan kepatuhan formalnya berupa pelaporan SPT Tahunan Badan selama 2 tahun berturut-turut.
12. Sektor Asuransi Jiwa Konvensional (kode KLU: 65111)
    Terindikasi melakukan *underreporting* pada perolehan PPN dan penyerahan PPN. Serta memperkecil keuntungan dengan melaporkan biaya lain-lain yang tidak wajar. Selain itu, KLU tidak melaksanakan kepatuhan formalnya berupa pelaporan SPT Tahunan Badan selama 2 tahun berturut-turut.

Adapun penjelasan dari masing-masing sektor adalah sebagai berikut. Sektor pembangkitan tenaga listrik terdiri dari berbagai macam sub sektor yang meliputi usaha membangkitkan tenaga listrik dan pengoperasian fasilitas

pembangkit yang menghasilkan energi listrik, yang berasal dari berbagai sumber energi, seperti tenaga air (hidroelektrik), batu bara, gas (turbin gas), bahan bakar minyak, diesel dan energi yang dapat diperbarui, tenaga surya, angin, arus laut, panas bumi (energi termal), tenaga nuklir dan lain-lain.

Sektor pertambangan batu bara terdiri dari berbagai macam sub sektor yang meliputi usaha operasi penambangan, pengeboran berbagai kualitas batu bara seperti antrasit, bituminous dan subbitominous baik pertambangan di permukaan tanah atau bawah tanah, termasuk pertambangan dengan cara pencairan (liquefaction).

Sektor bank pemerintah/BUMN/persero terdiri dari berbagai macam sub sektor yang meliputi kegiatan bank yang seluruh atau sebagian besar modalnya dimiliki oleh negara sebagaimana tercantum dalam Undang-undang mengenai BUMN yang berlaku.

Sektor bank umum swasta nasional devisa terdiri dari berbagai macam sub sektor yang meliputi kegiatan bank yang dimiliki oleh swasta nasional yang memperoleh surat penunjukan dari Bank Indonesia untuk dapat melakukan kegiatan usaha perbankan dalam valuta asing dan atau melakukan transaksi perbankan dengan pihak luar negeri.

Sektor industri minyak goreng kelapa sawit terdiri dari berbagai macam sub sektor yang meliputi usaha pengolahan lebih lanjut (pemurnian, pemucatan dan penghilangan bau yang tidak dikehendaki) dari minyak mentah kelapa sawit (CPO) menjadi minyak goreng kelapa sawit.

Sektor kendaraan bermotor roda empat atau lebih terdiri dari berbagai macam sub sektor yang meliputi usaha pembuatan atau perakitan kendaraan bermotor untuk penumpang atau barang, seperti sedan, jeep, truck, pick up, bus dan stasion wagon. Termasuk pembuatan kendaraan untuk keperluan khusus, seperti mobil pemadam kebakaran, mobil toko, mobil penyapu jalan, ambulans, dll.

Sektor perdagangan besar mobil baru terdiri dari berbagai macam sub sektor yang meliputi usaha perdagangan besar mobil baru, termasuk mobil khusus (seperti ambulans, karavan, mikrobus, pemadam kebakaran, dan sebagainya), lori, trailer, semi-trailer dan berbagai kendaraan pengangkut bermotor lainnya.

Sektor industri pemurnian dan pengilangan minyak bumi terdiri dari berbagai macam sub sektor yang meliputi usaha pemurnian dan pengilangan minyak bumi yang menghasilkan gas atau LPG, Naphtha, Avigas, Avtur, Gasoline, Minyak Tanah atau Kerosin, Minyak Solar, Minyak Diesel, Minyak Bakar atau Bensin, Residu, Solvent/Pelarut, Wax, Lubricant dan Aspal.

Sektor industri semen terdiri dari berbagai macam sub sektor yang meliputi usaha pembuatan macam-macam semen (semen hidrolik dan arang atau kerak besi), seperti portland, natural, semen mengandung alumunium, semen terak dan semen superfosfat dan jenis semen lainnya.

Sektor telekomunikasi tanpa kabel terdiri dari berbagai macam sub sektor yang meliputi kegiatan penyelenggaraan jaringan yang melayani telekomunikasi bergerak dengan teknologi seluler di permukaan bumi.

Sektor industri pupuk buatan tunggal hara makro primer terdiri dari berbagai macam sub sektor yang meliputi usaha pembuatan pupuk hara makro primer jenis pupuk buatan tunggal seperti urea, ZA, TSP, DSP dan Kalsium Sulfat.

Sektor distribusi gas alam dan buatan terdiri dari berbagai macam sub sektor yang meliputi usaha penyaluran gas melalui jaringan yang bertekanan ekstra tinggi (lebih dari 10 bar); yang bertekanan tinggi (antara 4 bar s.d. 10 bar); dan yang bertekanan menengah ke bawah (di bawah 4 bar) baik berasal dari produksi sendiri maupun produksi pihak lain sampai ke konsumen atau pelanggan.

Sektor perdagangan besar beras terdiri dari berbagai macam sub sektor yang meliputi usaha perdagangan besar beras untuk digunakan sebagai konsumsi akhir.

Sektor bank campuran dan asing terdiri dari berbagai macam sub sektor yang meliputi kegiatan bank campuran dan bank asing yang termasuk kelompok bank devisa yang kegiatan utamanya menghimpun dana masyarakat dalam bentuk giro, deposito dan tabungan baik dalam bentuk rupiah maupun valuta asing serta menyalurkan kembali dananya dalam bentuk pemberian kredit, dan melayani transaksi luar negeri. Bank Campuran adalah bank yang didirikan dengan komposisi pemegang saham dimiliki oleh bank yang berkedudukan di luar negeri dan bank yang berkedudukan di Indonesia. Sedangkan Bank Asing adalah Kantor Cabang yang mempunyai alamat dan tempat kedudukan di Indonesia dari Bank

yang berkedudukan di luar negeri, yang didirikan berdasarkan hukum asing dan berkantor pusat di luar negeri, yang secara langsung maupun tidak langsung bertanggung jawab kepada kantor pusat Bank yang bersangkutan sebagaimana tercantum dalam ketentuan Bank Indonesia yang berlaku.

Sektor pembiayaan konsumen (*consumers credit*) terdiri dari berbagai macam sub sektor yang meliputi usaha yang kegiatan utamanya melakukan pembiayaan untuk pengadaan barang dan jasa berdasarkan kebutuhan konsumen dengan sistem pembayaran secara angsuran atau berkala.

Sektor konstruksi gedung perkantoran terdiri dari berbagai macam sub sektor yang meliputi usaha pembangunan gedung yang dipakai untuk perkantoran, seperti kantor dan rumah kantor (rukan).

Sektor angkutan udara berjadwal domestik umum untuk penumpang terdiri dari berbagai macam sub sektor yang meliputi usaha pengangkutan penumpang, kargo dan pos dengan pesawat udara berdasarkan pada rute dan jadwal tertentu dengan tujuan kota-kota atau provinsi di dalam negeri.

Sektor bank sentral terdiri dari berbagai macam sub sektor yang meliputi kegiatan perbankan yang mempunyai wewenang dan hak dari pemerintah untuk mengeluarkan dan mengedarkan alat pembayaran yang sah, merumuskan dan menjalankan kebijakan moneter, mengelola cadangan devisa, menjaga dan memelihara kestabilan nilai rupiah, mengatur dan mengawasi perbankan, menjalankan fungsi lender ofv the last resort dan bertindak sebagai bankir pemerintah. Kelompok ini mencakup kegiatan Bank Indonesia, lembaga negara yang berfungsi sebagai Bank Sentral.

Sektor asuransi jiwa konvensional terdiri dari berbagai macam sub sektor yang meliputi usaha perasuransian yang memberikan jasa dalam penanggulangan resiko yang dikaitkan dengan hidup atau meninggalnya seseorang yang dipertanggungkan.

Selanjutnya dilakukan penghitungan potensi penerimaan negara yang dapat digali dari 20 KLU tersebut sesuai indikasi ketidakpatuhannya. Penghitungan dari masing-masing KLU adalah sebagai berikut:

A. Kuadran x3-y3

1. Penghitungan potensi KLU Sektor Pembangkitan Tenaga Listrik (kode KLU: 35101)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	401,685,500,000,000
b	Penyerahan PPN	83,657,975,000,000
c	Perolehan PPN	76,473,625,000,000
	Selisih Ekualisasi Omset dan Penyerahan	281,875,830,000,000
	Selisih Equalisasi Omset dan Perolehan PPN	124,369,125,000,000
d	Selisih Terbesar	281,875,830,000,000
e	Penghasilan Bruto (a+d)	683,561,330,000,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	321,595,775,000,000
h	Penghasilan Neto (e-g)	361,965,555,000,000
i	Pajak (e x 25% Tarif Pajak)	90,491,388,750,000
j	Pembayaran Pajak	13,706,192,260,212
TAX GAP (i-j)		**76,785,196,489,788**

2. Penghitungan potensi KLU Sektor Pertambangan Batu Bara (kode KLU: 5101)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	232,712,250,000,000
b	Penyerahan PPN	27,712,675,000,000
c	Perolehan PPN	98,478,525,000,000
	Selisih Ekualisasi Omset dan Penyerahan	184,055,472,500,000
	Selisih Equalisasi Omset dan Perolehan PPN	17,877,600,000,000
d	Selisih Terbesar	184,055,472,500,000
e	Penghasilan Bruto	416,767,722,500,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	201,233,500,000,000
h	Penghasilan Neto (a-d)	215,534,222,500,000
i	Pajak (e x 25% Tarif Pajak)	53,883,555,625,000
j	Pembayaran Pajak	16,691,085,811,236
TAX GAP (i-j)		**37,192,469,813,764**

3. Penghitungan potensi Sektor Bank Pemerintah/Bumn/Persero (kode KLU: 64121)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	318,316,500,000,000
b	Penyerahan PPN	2,830,960,000,000
c	Perolehan PPN	1,392,979,750,000
	Selisih Ekualisasi Omset dan Penyerahan	286,837,055,000,000
	Selisih Equalisasi Omset dan Perolehan PPN	157,765,270,250,000
d	Selisih Terbesar	286,837,055,000,000
e	Penghasilan Bruto	605,153,555,000,000
f	Biaya Lain	124,469,250,000,000
	Selisih Rasio Biaya Lain-lain	54,232,050,000,000
g	Biaya Total	179,891,950,000,000
h	Penghasilan Neto (a-d)	425,261,605,000,000
i	Pajak (e x 25% Tarif Pajak)	106,315,401,250,000
j	Pembayaran Pajak	28,605,204,852,694
TAX GAP (i-j)		**77,710,196,397,307**
DENDA TIDAK LAPOR SPT		**2,000,000**

4. Penghitungan potensi Sektor Bank Umum Swasta Nasional Devisa (kode KLU: 64125)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	207,981,750,000,000
b	Penyerahan PPN	5,462,395,000,000
c	Perolehan PPN	9,135,337,500,000
	Selisih Ekualisasi Omset dan Penyerahan	183,800,997,500,000
	Selisih Equalisasi Omset dan Perolehan PPN	94,855,537,500,000
d	Selisih Terbesar	183,800,997,500,000
e	Penghasilan Bruto	391,782,747,500,000
f	Biaya Lain	52,964,675,000,000
	Selisih Rasio Biaya Lain-lain	2,792,975,000,000
g	Biaya Total	164,446,025,000,000

h	Penghasilan Neto (a-d)	227,336,722,500,000
i	Pajak (e x 25% Tarif Pajak)	56,834,180,625,000
j	Pembayaran Pajak	27,979,655,635,701
TAX GAP (i-j)		**28,854,524,989,300**
DENDA TIDAK LAPOR SPT		**2,000,000**

Sehingga diperoleh total potensi untuk kuadran x3-y3 sebesar 220,542,387,690,158.

KLU	KPP Adm	Nilai Potensi
35101	51	76,785,196,489,788
5101	91	37,192,469,813,764
64121	93	77,710,198,397,307
64125	91	28,854,526,989,300
TOTAL POTENSI		**220,542,391,690,159**

B. Kuadran x2-y3

1. Penghitungan potensi Sektor Industri Minyak Goreng Kelapa Sawit (kode KLU: 10432)

   KLU yang bersangkutan tidak ada temuan selisih ekualisasi namun terdapat data ketetapan dari pemeriksaan sebelumnya.

2. Penghitungan potensi Sektor Industri Kendaraan Bermotor Roda Empat atau Lebih (kode KLU: 29100)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	224,411,500,000,000
b	Penyerahan PPN	210,093,000,000,000
c	Perolehan PPN	179,748,500,000,000
	Selisih Ekualisasi Omset dan Penyerahan	5,878,535,000,000
	Selisih Equalisasi Omset dan Perolehan PPN	67,542,750,000,000
d	Selisih Terbesar	67,542,750,000,000
e	Penghasilan Bruto	291,954,250,000,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	216,553,250,000,000

h	Penghasilan Neto (a-d)	75,401,000,000,000
i	Pajak (e x 25% Tarif Pajak)	18,850,250,000,000
j	Pembayaran Pajak	18,731,445,902,387
TAX GAP (i-j)		**118,804,097,613**

3. Penghitungan potensi Sektor Perdagangan Besar Mobil Baru (kode KLU: 45101)

   KLU yang bersangkutan tidak ada temuan selisih ekualisasi namun terdapat data ketetapan dari pemeriksaan sebelumnya dan tidak melaksanakan kepatuhan formal berupa pelaporan SPT tahunan PPh badan selama 2 tahun berturut-turut.

Komponen STP	**Jumlah**
DENDA TIDAK LAPOR SPT	**2,000,000**

4. Penghitungan potensi Sektor Industri Pemurnian dan Pengilangan Minyak Bumi (kode KLU: 19211)

Ket	**Komponen STP**	**Jumlah**
a	Penghasilan Bruto (Omset SPT Tahunan)	530,337,000,000,000
b	Penyerahan PPN	387,484,950,000,000
c	Perolehan PPN	163,751,950,000,000
	Selisih Ekualisasi Omset dan Penyerahan	95,121,720,000,000
	Selisih Equalisasi Omset dan Perolehan PPN	101,416,550,000,000
d	Selisih Terbesar	101,416,550,000,000
e	Penghasilan Bruto	631,753,550,000,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	394,129,574,041,318
h	Penghasilan Neto (a-d)	237,623,975,958,682
i	Pajak (e x 25% Tarif Pajak)	59,405,993,989,670
j	Pembayaran Pajak	54,340,791,199,822
TAX GAP (i-j)		**5,065,202,789,848**

Sehingga diperoleh total potensi untuk kuadran x2-y3 sebesar 5,184,008,887,462.

KLU	KPP Adm	Nilai Potensi
35101	51	-
5101	91	118,804,097,613
64121	93	2,000,000
64125	91	5,065,202,789,848
TOTAL POTENSI		**5,184,008,887,462**

C. Kuadran x3-y2

1. Penghitungan potensi Sektor Industri Semen (kode KLU: 23941)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	26,894,450,000,000
b	Penyerahan PPN	27,864,475,000,000
c	Perolehan PPN	13,337,725,000,000
	Selisih Ekualisasi Omset dan Penyerahan	3,390,525,500,000
	Selisih Equalisasi Omset dan Perolehan PPN	109,500,000,000
d	Selisih Terbesar	3,390,525,500,000
e	Penghasilan Bruto	30,284,975,500,000
f	Biaya Lain	10,164,072,500,000
	Selisih Rasio Biaya Lain-lain	3,050,577,500,000
g	Biaya Total	20,661,072,500,000
h	Penghasilan Neto (a-d)	9,623,903,000,000
i	Pajak (e x 25% Tarif Pajak)	2,405,975,750,000
j	Pembayaran Pajak	2,037,288,268,352
TAX GAP (i-j)		**368,687,481,648**
DENDA TIDAK LAPOR SPT		**2,000,000**

2. Penghitungan potensi Sektor Telekomunikasi Tanpa Kabel (kode KLU: 61200)
   KLU yang bersangkutan tidak ada temuan selisih ekualisasi namun terdapat data ketetapan dari pemeriksaan sebelumnya dan tidak melaksanakan kepatuhan formal berupa pelaporan SPT tahunan PPh badan selama 2 tahun berturut-turut.

Komponen STP	Jumlah
DENDA TIDAK LAPOR SPT	**2,000,000**

3. Penghitungan potensi KLU Sektor Pertambangan Batu Bara (kode KLU: 5101)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	29,974,650,000,000
b	Penyerahan PPN	359,214,500,000
c	Perolehan PPN	8,589,162,500,000
	Selisih Ekualisasi Omset dan Penyerahan	26,917,717,000,000
	Selisih Equalisasi Omset dan Perolehan PPN	6,398,162,500,000
d	Selisih Terbesar	26,917,717,000,000
e	Penghasilan Bruto	56,892,367,000,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	25,186,500,000,000
h	Penghasilan Neto (a-d)	31,705,867,000,000
i	Pajak (e x 25% Tarif Pajak)	7,926,466,750,000
j	Pembayaran Pajak	1,281,812,923,607
TAX GAP (i-j)		**6,644,653,826,393**

4. Penghitungan potensi Sektor Industri Pupuk Buatan Tunggal Hara Makro Primer (kode KLU: 20122)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	34,686,075,000,000
b	Penyerahan PPN	35,845,275,000,000
c	Perolehan PPN	7,888,292,500,000
	Selisih Ekualisasi Omset dan Penyerahan	4,280,946,750,000
	Selisih Equalisasi Omset dan Perolehan PPN	9,454,745,000,000
d	Selisih Terbesar	9,454,745,000,000
e	Penghasilan Bruto	44,140,820,000,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	31,051,675,000,000
h	Penghasilan Neto (a-d)	13,089,145,000,000
i	Pajak (e x 25% Tarif Pajak)	3,272,286,250,000

j	Pembayaran Pajak	2,892,881,297,355
TAX GAP (i-j)		**379,404,952,645**

5. Penghitungan potensi Sektor Distribusi Gas Alam dan Buatan (kode KLU: 35202)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	44,914,525,000,000
b	Penyerahan PPN	5,425,207,500,000
c	Perolehan PPN	6,447,187,500,000
	Selisih Ekualisasi Omset dan Penyerahan	35,447,010,250,000
	Selisih Equalisasi Omset dan Perolehan PPN	16,010,075,000,000
d	Selisih Terbesar	35,447,010,250,000
e	Penghasilan Bruto	80,361,535,250,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	40,589,075,000,000
h	Penghasilan Neto (a-d)	39,772,460,250,000
i	Pajak (e x 25% Tarif Pajak)	9,943,115,062,500
j	Pembayaran Pajak	2,297,621,919,856
TAX GAP (i-j)		**7,645,493,142,644**

6. Penghitungan potensi Sektor Perdagangan Besar Beras (kode KLU: 46311)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	30,438,000,000,000
b	Penyerahan PPN	1,793,995,500,000
c	Perolehan PPN	2,740,983,750,000
	Selisih Ekualisasi Omset dan Penyerahan	25,904,584,500,000
	Selisih Equalisasi Omset dan Perolehan PPN	12,478,016,250,000
d	Selisih Terbesar	25,904,584,500,000
e	Penghasilan Bruto	56,342,584,500,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	31,690,225,000,000
h	Penghasilan Neto (a-d)	24,652,359,500,000

i	Pajak (e x 25% Tarif Pajak)	6,163,089,875,000
j	Pembayaran Pajak	220,887,305,005
TAX GAP (i-j)		**5,942,202,569,995**

7. Penghitungan potensi Sektor Bank Campuran dan Asing (kode KLU: 64124)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	41,222,075,000,000
b	Penyerahan PPN	2,667,062,500,000
c	Perolehan PPN	2,022,837,500,000
	Selisih Ekualisasi Omset dan Penyerahan	34,845,025,750,000
	Selisih Equalisasi Omset dan Perolehan PPN	18,588,200,000,000
d	Selisih Terbesar	34,845,025,750,000
e	Penghasilan Bruto	76,067,100,750,000
f	Biaya Lain	17,793,950,000,000
	Selisih Rasio Biaya Lain-lain	9,110,202,500,000
g	Biaya Total	19,835,622,500,000
h	Penghasilan Neto (a-d)	56,231,478,250,000
i	Pajak (e x 25% Tarif Pajak)	14,057,869,562,500
j	Pembayaran Pajak	13,407,857,103,254
TAX GAP (i-j)		**650,012,459,246**

8. Penghitungan potensi Sektor Pembiayaan Konsumen (Consumers Credit) (kode KLU: 64922)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	36,517,925,000,000
b	Penyerahan PPN	2,632,277,500,000
c	Perolehan PPN	2,733,255,000,000
	Selisih Ekualisasi Omset dan Penyerahan	30,599,034,250,000
	Selisih Equalisasi Omset dan Perolehan PPN	15,525,707,500,000
d	Selisih Terbesar	30,599,034,250,000
e	Penghasilan Bruto	67,116,959,250,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	27,687,150,000,000

h	Penghasilan Neto (a-d)	39,429,809,250,000
i	Pajak (e x 25% Tarif Pajak)	9,857,452,312,500
j	Pembayaran Pajak	3,652,117,845,601
TAX GAP (i-j)		**6,205,334,466,900**
DENDA TIDAK LAPOR SPT		**2,000,000**

9. Penghitungan potensi Sektor Konstruksi Gedung Perkantoran (kode KLU: 41012)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	40,903,950,000,000
b	Penyerahan PPN	31,878,975,000,000
c	Perolehan PPN	20,618,975,000,000
	Selisih Ekualisasi Omset dan Penyerahan	5,343,619,500,000
	Selisih Equalisasi Omset dan Perolehan PPN	167,000,000,000
d	Selisih Terbesar	5,343,619,500,000
e	Penghasilan Bruto	46,247,569,500,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	37,735,325,000,000
h	Penghasilan Neto (a-d)	8,512,244,500,000
i	Pajak (e x 25% Tarif Pajak)	2,128,061,125,000
j	Pembayaran Pajak	1,908,600,957,826
TAX GAP (i-j)		**219,460,167,174**
DENDA TIDAK LAPOR SPT		**2,000,000**

10. Penghitungan potensi Sektor Angkutan Udara Berjadwal Domestik Umum untuk Penumpang (kode KLU: 51101)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	54,014,350,000,000
b	Penyerahan PPN	27,682,475,000,000
c	Perolehan PPN	25,540,375,000,000
	Selisih Ekualisasi Omset dan Penyerahan	21,470,583,500,000
	Selisih Equalisasi Omset dan Perolehan PPN	1,466,800,000,000
d	Selisih Terbesar	21,470,583,500,000

e	Penghasilan Bruto	75,484,933,500,000
f	Biaya Lain	19,966,725,000,000
	Selisih Rasio Biaya Lain-lain	2,861,940,000,000
g	Biaya Total	54,154,010,000,000
h	Penghasilan Neto (a-d)	21,330,923,500,000
i	Pajak (e x 25% Tarif Pajak)	5,332,730,875,000
j	Pembayaran Pajak	770,184,631,499
TAX GAP (i-j)		**4,562,546,243,501**

11. Penghitungan potensi Sektor Bank Sentral (kode KLU: 64110)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	77,658,750,000,000
b	Penyerahan PPN	358,539,000,000
c	Perolehan PPN	1,790,277,500,000
	Selisih Ekualisasi Omset dan Penyerahan	70,310,923,500,000
	Selisih Equalisasi Omset dan Perolehan PPN	37,039,097,500,000
d	Selisih Terbesar	70,310,923,500,000
e	Penghasilan Bruto	147,969,673,500,000
f	Biaya Lain	-
	Selisih Rasio Biaya Lain-lain	-
g	Biaya Total	42,364,975,000,000
h	Penghasilan Neto (a-d)	105,604,698,500,000
i	Pajak (e x 25% Tarif Pajak)	26,401,174,625,000
j	Pembayaran Pajak	10,326,403,047,938
TAX GAP (i-j)		**16,074,771,577,062**
DENDA TIDAK LAPOR SPT		**2,000,000**

12. Penghitungan potensi Sektor Asuransi Jiwa Konvensional (kode KLU: 65111)

Ket	Komponen STP	Jumlah
a	Penghasilan Bruto (Omset SPT Tahunan)	43,184,100,000,000
b	Penyerahan PPN	64,726,110,360
c	Perolehan PPN	306,776,250,000
	Selisih Ekualisasi Omset dan Penyerahan	39,232,804,889,640
	Selisih Equalisasi Omset dan Perolehan PPN	21,285,273,750,000

d	Selisih Terbesar	39,232,804,889,640
e	Penghasilan Bruto	82,416,904,889,640
f	Biaya Lain	34,143,525,000,000
	Selisih Rasio Biaya Lain-lain	20,833,410,000,000
g	Biaya Total	23,533,640,000,000
h	Penghasilan Neto (a-d)	58,883,264,889,640
i	Pajak (e x 25% Tarif Pajak)	14,720,816,222,410
j	Pembayaran Pajak	1,003,411,881,452
TAX GAP (i-j)		**13,717,404,340,958**
DENDA TIDAK LAPOR SPT		**2,000,000**

Sehingga diperoleh total potensi untuk kuadran x2-y3 sebesar 62,409,983,228,166.

KLU	KPP Adm	Nilai Potensi
23941	51	368,689,481,648
61200	93	2,000,000
5101	51	6,644,653,826,393
20122	51	379,404,952,645
35202	51	7,645,493,142,644
46311	51	5,942,202,569,995
64124	91	650,012,459,246
64922	91	6,205,336,466,900
41012	93	219,462,167,174
51101	93	4,562,546,243,501
64110	93	16,074,773,577,062
65111	93	13,717,406,340,958
TOTAL POTENSI		**62,409,983,228,166**

BAB 5
KESIMPULAN

5.1 Kesimpulan

Berdasarkan pengujian terhadap pengelompokkan KLU menggunakan metode c*lustering* K-Means dengan jumlah 3 pusat *cluster*, kesimpulan yang dapat diambil adalah sebagai berikut:

1. Data yang digunakan adalah data KLU untuk sembilan KPP dalam rentang tahun 2016 s.d tahun 2019. Data tersebut dibagi menjadi tiga kelompok, di mana tiap kelompok memiliki profil dan karakteristik yang berbeda. Profilisasi dilakukan terhadap variabel x dan variabel y, dan didapatkan kesimpulan bahwa kelompok yang memiliki nilai titik pusat pada variabel x tinggi dianggap memiliki tingkat kepatuhan yang rendah. Dan kelompok yang memiliki nilai titik pusat pada variabel y tinggi, maka kelompok tersebut dianggap berpotensi. Kelompok yang potensial dinilai mampu memberikan kontribusi yang baik untuk meningkatkan penerimaan pajak. Pengawasan dan pemeriksaan dapat dilakukan dengan memprioritaskan 20 KLU yang masuk ke dalam kuadran berwarna merah dengan tingkat ketidakpatuhan tinggi (variabel X) dan memiliki dampak fiskal yang tinggi (variabel Y). Dengan estimasi potensi penerimaan yang dapat digali sebesar 288,136,383,805,787.
2. Hasil pengujian didapatkan bahwa pada variabel x, 173 KLU memiliki tingkat kepatuhan rendah, 156 KLU memiliki tingkat kepatuhan sedang, 684 KLU memiliki tingkat kepatuhan tinggi, sedangkan pada variabel y, 967 KLU memiliki dampak fiskal rendah, 38 KLU memiliki dampak fiskal sedang, dan 8 KLU memiliki dampak fiskal tinggi.
3. Nilai rata-rata koefisien *silhouette* yang didapatkan untuk variabel x dan variabel y masing-masing adalah 0.69 dan 0.92. Semakin besar nilai *silhoutte coefficient* dan mendekati nilai 1 maka semakin baik *clustering* yang dihasilkan. Dengan demikian kualitas *clustering* yang terbentuk pada penelitian ini dapat dikatakan baik.

5.2 Saran

Adapun saran yang bisa diberikan adalah sebagai berikut:

1. Pemilihan metode *preprocessing* pada *data mining* mempengaruhi akurasi dari hasil *clustering*, maka perlu diterapkan metode *preprocessing* yang lebih baik agar hasil *cluster* yang terbentuk lebih akurat.
2. Untuk mendapatkan analisis yang lebih valid, agar selain menggunakan data internal dari DJP sebaiknya juga dilengkapi dengan data eksternal yang berkaitan dengan perpajakan seperti dari instansi, lembaga, asosiasi, dan pihak lain (ILAP).

DAFTAR PUSTAKA

[1] U. Fayyad and R. Uthurusamy, “Data mining and knowledge discovery in databases,” Communications of the ACM, vol. 39, no. 11, pp. 24–26, 1996.

[2] M. J. Mazur and A. H. Plumley, “Understanding the Tax Gap,” National Tax Journal, vol. 60, no. 3, pp. 569–576, 2007.

[3] C. Vercellis, Business intelligence: data mining and optimization for decision making. Chichester: John Wiley and Sons, 2009.

[4] T. Wuryanto, “ANALISIS PENGGALIAN POTENSI PAJAK PENGHASILAN DAN PAJAK PERTAMBAHAN NILAI DI KPP PRATAMA SUKOHARJO,” Thesis, Dept. Economics and Development Study, Universitas Sebelas Maret, Surakarta, Indonesia, 2011.

[5] T. S. Madhulatha, “AN OVERVIEW ON CLUSTERING METHODS,” IOSR Journal of Engineering, vol. 02, no. 04, pp. 719–725, 2012.

[6] R.-S. Wu, C. S. Ou, H.-ying Lin, S.-I. Chang, and D. C. Yen, “Using data mining technique to enhance tax evasion detection performance,” Expert Systems with Applications, vol. 39, no. 10, pp. 8769–8777, 2012.

[7] Dewi, Olivia, and Retnaningtyas Widuri. "Faktor-faktor Yang Mempengaruhi Keberhasilan Penerimaan Pajak Daerah Kota Tarakan." Petra Christian University Tax and Accounting Review, vol. 3, no. 2, 2013.

[8] M. Robani and A. Widodo, “Algoritma K-Means Clustering Untuk Pengelompokan Ayat Al Quran Pada Terjemahan Bahasa Indonesia,” JURNAL SISTEM INFORMASI BISNIS, vol. 6, no. 2, p. 164, 2016.

[9] Y. D. Darmi and A. Setiawan, “Penerapan Metode Clustering K-Means dalam Pengelompokan Penjualan Produk,” JURNAL MEDIA INFOTAMA, vol. 12, no. 2, 2017.

[10] L. Maulida, “Penerapan Datamining dalam Mengelompokkan Kunjungan Wisatawan ke Objek Wisata Unggulan di Prov. DKI Jakarta dengan K-Means,” JISKA (Jurnal Informatika Sunan Kalijaga), vol. 2, no. 3, p. 167, 2018.

[11] D. A. Nasution, H. H. Khotimah, and N. Chamidah, “Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN,” Computer Engineering, Science and System Journal, vol. 4, no. 1, p. 78, 2019.

[12] A. Rahadian, T. M. Barusman, and H. Haninun, “Analisis Hasil Pemeriksaan Pajak untuk Memetakan (Mapping) Klasifikasi Lapangan Usaha (KLU) Wajib Pajak Badan yang Potensial di Kantor Wilayah DJP Bengkulu dan Lampung Periode Tahun 2016-2019,” Jurnal Manajemen VISIONIST, vol. 9, no. 2, pp. 1–15, 2020.

[13] G. Subroto, Artikel - MEMAHAMI TAX GAP. [Online]. Available: https://bppk.kemenkeu.go.id/content/artikel/balai-diklat-keuangan-denpasar-memahami-tax-gap-2020-01-09-6bfb976f/. [Accessed: 21-Jan-2021].

[14] M. M. R. Sari, and N. N. Afriyanti. "Pengaruh Kepatuhan Wajib Pajak Dan Pemeriksaan Pajak Terhadap Penerimaan PPh Pasal 25/29 Wajib Pajak Badan Pada KPP Pratama Denpasar Timur". Jurnal Ilmiah Akuntansi dan Bisnis, [S.l.], vol. 7, no. 1, jan. 2012.

[15] Keputusan Direktur Jenderal Pajak. "Nomor : KEP - 233/PJ/2012 tentang Klasifikasi Lapangan Usaha Wajib Pajak". 2012.

[16] Surat Edaran. "Nomor SE-15 /PJ/2018 tentang Kebijakan Pemeriksaan". 2018.

[17] M.-C. Hung, J. Wu, and J. H. Chang, “An Efficient k-Means Clustering Algorithm Using Simple Partitioning,” *Journal of Information Science and Engineering*, vol. 21, pp. 1157–1177, 2005.

[18] D. J. Strauss and J. A. Hartigan, “Clustering Algorithms,” *Biometrics*, vol. 31, no. 3, p. 793, 1975.

[19] A. C. Rencher, “Methods of Multivariate Analysis,” *Wiley Series in Probability and Statistics*, 2002.

[20] J. Han and M. Kamber, *Data mining: concepts and techniques*. Burlington, MA: Elsevier, 2012.

[21] F. Nur Aini, S. Palgunadi, and R. Anggrainingsih, “Clustering Business Process Model Petri Net dengan Complete Linkage,” *Jurnal Teknologi & Informasi ITSmart*, vol. 3, no. 2, p. 47, 2016.

[22] A. Barakbah, and A. Helen. "Optimized K-Means: An Algorithm of Initial Centroids Optimization for K-means". 2005

[23] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to data mining*. Pearson, 2020.

[24] Velmurugan, “Computational complexity between k-means and K-Medoids clustering algorithms for normal and uniform distributions of data points,” *Journal of Computer Science*, vol. 6, no. 3, pp. 363–368, 2010.

[25] A. Ilham, “Penggabungan metode U-CONTROL Chart DAN Metode automatic Clustering Differential Evolution UNTUK Penentuan Jumlah klaster pada METODE K-MEANS,” 2019.

Halaman ini sengaja dikosongkan

BIOGRAFI PENULIS

JESSICA RAHMAWATI NUGROHO, lahir pada tanggal 17 Juni 1990 di Malang, Jawa Timur. Anak ketiga dari empat bersaudara ini memulai Pendidikan formal di SD Al Hikmah Surabaya tamat tahun 2002, masuk SMP Negeri 12 Surabaya tamat tahun 2005, meneruskan ke SMA Negeri 15 Surabaya tamat tahun 2008. Kemudian melanjutkan Pendidikan di Institut Teknologi Sepuluh Nopember pada jurusan Sistem Informasi, lulus tahun 2012.

Saat ini penulis bekerja sebagai Pegawai Negeri Sipil di Direktorat Jenderal Pajak. Sebelumnya penulis pernah bekerja di perusahaan swasta PT Sigma Metrasys Solution.

Alhamdulillah, pada tahun 2019 dengan beasiswa dari Kementrian Komunikasi dan Informatika penulis berkesempatan melanjutkan studi Pascasarjana di Institut Teknologi Sepuluh Nopember, pada Fakultas Teknologi Elektro dan Informatika Cerdas, Departemen Teknik Elektro, Bidang Keahlian Telematika dan lulus tahun 2021.

Untuk bisa berkomunikasi dengan penulis terkait dengan penelitian yang telah dikerjakan, dapat dihubungi melalui Email: nugrohojessica@gmail.com.