SKRIPSI

IMPLEMENTASI ALGORITMA APRIORI PADA ANALISIS POLA USAHA DAN ALGORITMA K-MEANS PADA PENYEBARAN USAHA UMKM DI KOTA MAKASSAR

Disusun dan diajukan oleh :

JUMRAINI.J.JAMALUDDIN

D121171013

DEPARTEMEN TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS HASANUDDIN

MAKASSAR

2022

LEMBAR PENGESAHAN SKRIPSI

IMPLEMENTASI ALGORITMA APRIORI PADA ANALISIS POLA USAHA DAN ALGORITMA K-MEANS PADA PENYEBARAN USAHA UMKM DI KOTA MAKASSAR

Disusun dan diajukan oleh

JUMRAINI. J. JAMALUDDIN

D121171013

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 27 Januari 2023 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Dr. Amil Ahmad Ilham, S.T., MIT.	Dr. Ir. Ingrid Nurtanio, MT
Nip. 197310101998021001	Nip. 196108131988112001

Ketua Program Studi,

Prof. Dr. Indrabayu, S.T., M.T.,M.Bus.Sys
Nip. 19750716 200212 1 004

PERNYATAAN KEASLIAN

Yang bertanda tangan di bawah ini:

Nama : Jumraini. J. Jamaluddin

NIM : D121171013

Departemen : Teknik Informatika

Jenjang : S1

Menyatakan dengan ini karya tulisan saya berjudul:

IMPLEMENTASI ALGORITMA APRIORI PADA ANALISIS POLA USAHA DAN ALGORITMA K-MEANS PADA PENYEBARAN USAHA UMKM DI KOTA MAKASSAR

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilalihan tulisan orang lain bahwa skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini hasil karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Makassar, 27 Januari 2022

Yang menyatakan,

Jumraini. J. Jamaluddin

ABSTRAK

Usaha Mikro, Kecil dan Menengah (UMKM) merupakan salah satu pilar utama dalam perekonomian nasional yang berwawasan kemandirian, yang memiliki potensi besar untuk meningkatkan kesejahteraan masyarakat. Kota Makassar sendiri mencatat pertumbuhan yang cukup tinggi. Merujuk pada data dinas koperasi dan UKM kota Makassar pada tahun 2017-2018 jumlah pelaku UMKM berjumlah 2.683. Tahun 2019-2020 Dinas koperasi dan UKM mencatat terdapat 13.277 UMKM. Di tengah pertumbuhan UMKM yang berkembang pesat masih banyak kendala (permasalahan) yang terjadi, contohnya kurangnya permodalan, kesulitan dalam pemasaran, masih banyak pelaku UMKM yang masih gagap teknologi dan permasalahan lainnya. Untuk itu diperlukannya analisis pola usaha dan penyebaran UMKM di kota Makassar menggunakan data mining. Penelitian ini dilakukan dengan mencari keterkaitan antar variabel untuk melihat pola usaha menggunakan metode asosiasi, dan melihat pengelompokan penyebaran UMKM dengan metode *Clustering*. Pada metode asosiasi menggunakan algoritma *apriori* dengan 2 parameter yaitu *minimum support* 35% dan *minimum confidence* 80%, menghasilkan 8 *rules* dengan keterkaitan antara variabel cukup tinggi. Sedangkan pada metode *clustering* nilai k yang merupakan jumlah klaster ditentukan menggunakan metode *Elbow* dengan SSE (144.58) dan nilai *Silhouette Score* (0.44) menghasilkan 6 klaster. pada hasil klaster Bentuk Usaha, pendapatan bersih perbulan, jumlah karyawan dan pendidikan terakhir pemilik usaha mempengaruhi setiap klasternya.

Kata kunci: UMKM, Asosiasi, *Clustering*, *K-Means*, *Minimum Support*, *Minimum Confidence*, SSE, *Silhouette Score*.

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah *subhanahu wa ta'ala*, karena atas bantuan dan limpahan karunia-Nya maka penulis dapat menyelesaikan tugas akhir dengan judul "Implementasi Algoritma Apriori pada Analisis Pola Usaha dan Algoritma K-Means pada Penyebaran Usaha UMKM di Kota Makassar".

Penulis menyadari bahwa penyusunan dan penulisan tugas akhir ini tidak dapat terselesaikan dengan baik tanpa adanya bantuan, bimbingan, serta dukungan dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih banyak kepada

1. Allah *subhanahu wa ta'ala*, atas segala rahmat bantuan serta karunianya diberikan kepada penulis hingga saat ini.
2. Kedua orang tua penulis, Bapak Jamaluddin dan Ibu Jastia yang selalu memberikan dukungan, doa, dan semangat yang tiada henti-hentinya.
3. Bapak Dr. Amil Ahmad Ilham, S.T., M.IT. selaku pembimbing I dan Ibu Dr. Ir. Ingrid Nurtanio, M.T. selaku pembimbing II, yang senantiasa menyediakan waktu, tenaga, pikiran, dan perhatian yang luar biasa dalam mengarahkan penulis untuk menyelesaikan tugas akhir.
4. Segenap staf dan dosen Departemen Teknik Informatika, Fakultas Teknik, Universitas Hasanuddin yang telah membantu kelancaran penyelesaian tugas akhir.

5. Fitriani Nasir yang selalu membantu dan mendukung penulis dari awal hingga menyelesaikan tugas akhir ini.
6. Six Girl, Suci, Ilmi, Priska, Rieka, Rini yang sejak maba hingga saat ini selalu mendukung, menemani dan membantu penulis.
7. Penghuni Lab UBICON yang senantiasa membantu dan mendengar keluhan penulis selama menyelesaikan Tugas Akhir ini.
8. Irma, devy dan Seluruh anggota RECOGNIZER yang belum sempat dituliskan namanya, terimakasih atas dukungan, bantuan, dan semangatnya selama ini.
9. Serta pihak-pihak lain yang tidak disebutkan dan tanpa sadar telah menjadi inspirasi dan membantu penulis dalam menyelesaikan tugas akhir.

Penulis berharap semoga Allah *subhanahu wa ta'ala* berkenan membalas segala kebaikan serta jasa dari semua pihak yang telah banyak membantu penulis dalam menyelesaikan tugas akhir ini. Penulis menyadari bahwa tugas akhir ini masih jauh dari kata sempurna dan masih memiliki kekurangan dan keterbatasan. Oleh karena itu, dengan segenap hati, penulis memohon maaf atas segala kesalahan dan kekurangan yang ada pada tugas akhir ini. Penulis juga berharap diberi segala bentuk saran serta masukan yang membangun dari berbagai pihak. Semoga tugas akhir ini dapat memberi manfaat bagi para pembaca dan semua pihak. Aamiin.

Gowa, 20 November 2022

Penulis

Jumraini.J. Jamaluddin

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1. Latar Belakang

Usaha Mikro, Kecil dan Menengah (UMKM) merupakan salah satu pilar utama dalam perekonomian nasional yang berwawasan kemandirian, yang memiliki potensi besar untuk meningkatkan kesejahteraan masyarakat. Perekonomian di Indonesia secara nasional menunjuk bahwa kegiatan UMKM merupakan usaha yang konsisten dan mampu terus berkembang. Fakta menunjukkan bahwa kesempatan kerja yang diciptakan oleh kelompok UMKM tersebut jauh lebih banyak dibandingkan dengan tenaga kerja yang bisa diserap oleh usaha besar. Menurut badan pusat statistik (BPS) pada tahun 2018 jumlah Usaha Mikro Kecil dan Menengah (UMKM) mencapai 64,2 Juta. Dimana 60 persen dari perannya menyumbangkan angka besar bagi produk domestik bruto (PBD). Tak hanya itu UMKM juga membuka peluang tenaga kerja mencapai 98 persen keseluruhan usaha yang beroperasi di Indonesia.

Kota Makassar sendiri, jumlah usaha mikro kecil dan menengah (UMKM) yang bergerak dalam berbagai industri mencatat pertumbuhan yang cukup tinggi dalam beberapa tahun terakhir. Merujuk pada data dinas koperasi dan UKM kota Makassar pada tahun 2017-2018 jumlah pelaku UMKM berjumlah 2.683. Tahun 2019-2020 Dinas koperasi dan UKM mencatat terdapat 13.277 UMKM yang terdiri dari usaha rumah tangga 5.311, usaha mikro sebanyak 4.647, serta usaha menengah sebanyak 3.319. Bahkan di masa krisis pandemi UMKM terus berkembang dan berinovasi melalui platform-platform digital.

Sejauh ini UMKM telah diketahui sebagai sektor usaha yang sangat penting dalam peningkatan perekonomian di masyarakat, sehingga memiliki prospek yang baik untuk terus dikembangkan. Maka perlunya pemberdayaan UMKM yang lebih diarahkan pada peningkatan proses panjang pengusaha kecil menjadi pengusaha menengah dan pengusaha mikro menjadi usaha kecil. Oleh karena itu, harus selalu diupayakan strategi yang tepat untuk memberdayakan UMKM agar kesejahteraan masyarakat semakin terangkat.

Akan tetapi dalam penerapan, masih banyak kendala atau permasalahan yang terjadi seperti kurangnya permodalan, hal ini sesuai dengan survei yang dilakukan oleh *pricewaterhouse Coopers* tahun 2019, yang mana 78% UMKM di Indonesia belum mendapatkan akses pembiayaan. Selanjutnya kurangnya inovasi, keterampilan dan sumber daya pengelolaan dari pelaku UMKM dalam menjalankan usahanya, sehingga banyak usaha yang hanya bertahan selama 1-2 tahun kemudian bangkrut karena produk atau jasa yang ditawarkan tidak kuat atau kalah bersaing. Iklim usaha seperti perizinan, aturan undang-undang yang tidak kondusif juga sangat mempengaruhi perkembangan UMKM.

Kendala lainnya yang sangat berpengaruh dalam perkembangan UMKM ialah kesulitan dalam pemasaran atau promosi UMKM serta masih banyak pelaku UMKM yang masih gagap teknologi. Meskipun pemerintah sekarang merancang gerakan nasional bangga buatan Indonesia (BBI) yang mendukung usaha UMKM di Indonesia secara digital. Dimana pada akhir Desember 2020 UMKM yang go digital melalui gerakan BBI sudah mencapai 3.8 juta dan pada maret 2021 jumlah UMKM yang

memasuki ekosistem digital melonjak menjadi 4.8 juta. Akan tetapi ini masih termasuk sedikit jika dibandingkan dengan jumlah keseluruhan UMKM saat ini. Bahkan di masa pandemi UMKM di seluruh kota di Indonesia mengalami kemunduran produktivitas. Menurut *United Nations Development Programme* (UNDO) dan Institut Penelitian Ekonomi dan Sosial (LPEM) Universitas Indonesia pada januari 2021, sembilan dari sepuluh UMKM di Indonesia mengalami penurunan permintaan, dan bahkan 80% memiliki keuntungan yang lebih rendah. Tak terkecuali juga untuk kota Makassar, dan kendala lainnya yang yang membuat keterbatasan kemampuan usaha kecil untuk terus dikembangkan.

Dari masalah-masalah tersebut maka diperlukannya analisis untuk mengetahui bagaimana perkembangan UMKM berdasarkan pola usaha dan penyebaran UMKM di kota Makassar, sehingga bisa memberikan sebuah gambaran pemberdayaan atau perlakuan yang cocok untuk UMKM-UMKM yang ada di kota makassar.

Berdasarkan uraian diatas untuk mengetahui bagaimana pola usaha dan penyebaran UMKM di Kota Makassar maka penulis mengusulkan judul "Implementasi Algoritma Apriori pada Analisis Pola Usaha dan Algoritma K-Means pada Penyebaran Usaha UMKM di Kota Makassar". Pemanfaatan *data mining* dengan teknik *association rules* dapat menjadi solusi dalam membantu analisa dasar dalam pengambilan keputusan untuk menentukan bagaimana pola usaha UMKM di kota Makassar. Selain itu analisis ini juga akan menggunakan teknik data mining clustering untuk mengelompokkan UMKM untuk melihat penyebaran UMKM di kota Makassar. Dari hasil analisis ini diharapkan dapat membantu pemerintah atau komunitas terkait

untuk memberikan pemberdayaan atau perlakuan yang sesuai agar pelaku usaha UMKM bisa mengembangkan usahanya lebih baik dan bisa meningkatkan kesejahteraan pelaku usahanya.

1.2. Rumusan Masalah

Berdasarkan latar belakang, maka rumusan masalah pada penelitian ini antara lain:

1. Bagaimana menganalisis keterkaitan antar variabel UMKM untuk menghasilkan pola usaha dengan *data mining*?
2. Bagaimana menganalisis pengelompokan penyebaran UMKM di kota Makassar dengan *data mining*?

1.3. Tujuan Penelitian

Tujuan dari penelitian ini antara lain:

1. Untuk menetapkan metode *data mining* dalam mengidentifikasi keterkaitan antar variabel UMKM untuk mengetahui bagaimana pola usaha yang ada pada UMKM di Kota Makassar
2. Untuk menetapkan metode *data mining* dalam mengelompokkan penyebaran usaha UMKM di kota Makassar

1.4. Manfaat Penelitian

Dengan dilakukannya penelitian ini, diharapkan manfaat yang didapatkan antara lain:

1. Membantu dalam melakukan analisis untuk mengetahui bagaimana pola usaha dan penyebaran usaha UMKM menggunakan metode *data mining.*
2. Hasil analisis dapat dimanfaatkan oleh pemerintah atau komunitas sebagai dasar untuk pemberdayaan atau memberikan tindakan yang diperlukan demi kepentingan perkembangan UMKM

1.5. Batasan Masalah

Batasan masalah yang ditentukan dalam sistem ini adalah :

1. Pengambilan data dilakukan secara Kuesioner, dimana data dikumpulkan dengan membagikan kuesioner *online* maupun *offline* ke pelaku usaha UMKM dengan bantuan komunitas UMKM dan data pelaku usaha dari Dinas Koperasi.
2. Data yang di analisis pada penelitian ini adalah data-data pelaku usaha umkm mulai dari jenis usaha, jumlah karyawan, pendapatan bersih, pendidikan terakhir, penggunaan *E-Commerce*, pernah melakukan pemberdayaan, adanya pemberdayaan lanjutan, serata 20 kendala selama menjalankan usaha UMKM.

1.6. Metode Penelitian

1. Metode pengumpulan data

   Pengambilan data dilakukan melalui kuesioner baik membagikannya secara *online* maupun secara *offline* ke pelaku usaha UMKM
2. Studi Literatur

Studi literatur dilakukan dengan mengumpulkan informasi yang dapat menjadi landasan teori melalui berbagai sumber, mulai dari buku, jurnal, skripsi, paper, internet dan sumber lainnya.

3. Diskusi dan Konsultasi

   Diskusi dan konsultasi dilakukan dengan melakukan tanya jawab dengan dosen pembimbing, pihak pemerintah melalui dinas koperasi, pihak komunitas UMKM serta pihak lainnya yang dapat mendukung penyelesaian penelitian ini.

Untuk memberikan gambaran singkat mengenai isi tulisan secara keseluruhan, maka akan diuraikan beberapa tahapan dari penulisan secara sistematis, yaitu:

BAB I PENDAHULUAN

Bab ini menguraikan secara umum mengenai hal yang menyangkut Latar Belakang, Perumusan Masalah, Batasan Masalah, Tujuan Penelitian, Manfaat Penelitian, dan Sistematika Penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini membahas landasan teori yang digunakan untuk menganalisis masalah yang akan diteliti serta hal-hal lain yang berhubungan dengan variabel data yang akan digunakan, *Data Mining, Asociacion Rules, Clustering, Algoritma K-Means*.

BAB III METODOLOGI PENELITIAN

Pada bab ini berisi tentang tahapan penelitian, waktu dan lokasi penelitian, instrumen penelitian, pengumpulan data, penerapan metode penelitian, penerapan

algoritma, teknik pengolahan data, serta informasi akhir berupa sebuah pengetahuan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi tentang sistem yang telah berhasil dibangun serta pembahasannya.

BAB V PENUTUP

Bab ini berisi tentang kesimpulan yang didapatkan berdasarkan hasil penelitian yang telah dilakukan serta saran-saran untuk pengembangan lebih lanjut.

BAB II

TINJAUAN PUSTA KA

2.1. UMKM

UMKM adalah unit usaha produktif yang berdiri sendiri yang dilakukan oleh orang atau perorangan atau badan usaha di semua sektor ekonomi. Pada prinsipnya, perbedaan antara Usaha Mikro (UMI), Usaha Kecil(UK), Usaha Menengah(UM), dan Usaha Besar (UB) umumnya didasari pada nilai aset awal (tidak termasuk tanah dan bangunan), omset rata-rata per tahun, atau jumlah pekerjaan tetap (Tambunan, 2012)

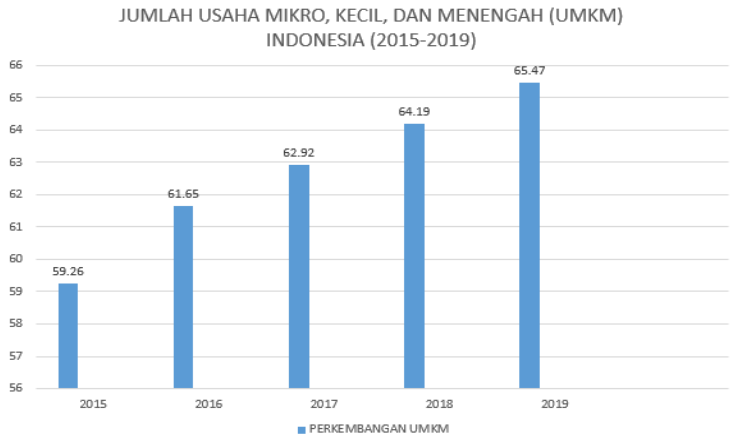
Menurut undang-undang nomor 20 tahun 2008 UMKM memiliki kriteria dimana Usaha Mikro, memiliki kekayaan bersih paling banyak Rp. 50.000.000 tidak termasuk tanah dan bangunan tempat usaha, dengan usaha penjualan tahunan paling banyak Rp 300.000.000. Usaha Kecil, Memiliki kekayaan bersih lebih dari Rp50.000.000,00 sampai dengan paling banyak Rp500.000.000,00 dan memiliki hasil penjualan tahunan lebih dari Rp300.000.000,00 sampai dengan paling banyak Rp2.500.000.000,00. Usaha Menengah Memiliki kekayaan bersih lebih dari Rp500.000.000,00 sampai dengan paling banyak Rp10.000.000.000,00 dan memiliki hasil penjualan tahunan lebih dari Rp2.500.000.000,00 sampai dengan paling banyak Rp50.000.000.000,00.

Usaha Mikro, Kecil, dan Menengah (UMKM) merupakan bagian terbesar dalam sistem perekonomian nasional, dimana merupakan indikator tinggi partisipasi masyarakat dalam sektor kegiatan ekonomi. UMKM menjadi salah satu tulang punggung perekonomian Indonesia. Pada tahun 2017 total keseluruhan pelaku usaha

di indonesia, kementerian koperasi dan UKM RI melaporkan bahwa secara unit, UMKM memiliki pangsa sekitar 99,99% (62,9 juta unit), sementara usaha besar hanya sebanyak 0,01 % atau sekitar 5400 unit. Usaha Mikro menyerap sekitar 107,2 juta tenaga kerja (89,2%), Usaha Kecil 5,7 juta (4,74%), dan Usaha Menengah 3,73 juta (3,11%); sementara Usaha Besar menyerap sekitar 3,58 juta jiwa. Artinya secara gabungan UMKM menyerap sekitar 97% tenaga kerja nasional, sementara Usaha Besar hanya menyerap sekitar 3% dari total tenaga kerja nasional (Haryanti, 2018)

UMKM tercatat terus berkembang di setiap tahunnya, Pada tahun 2019 Kementerian Koperasi dan UKM mencatat, jumlah usaha mikro, kecil, dan menengah (UMKM) mencapai 65.47. Jumlah tersebut naik 1,98% jika dibandingkan pada tahun sebelumnya yang sebesar 64,19 juta unit. Jumlah tersebut mencapai 99.99% dari total usaha yang ada di Indonesia. Sementara, usaha berskala besar hanya sebanyak 5.637 unit atau setara dengan 0.01%. secara rinci, sebanyak 64,6 juta unit merupakan usaha mikro. Jumlahnya setara dengan 98,67% dari total UMKM di seluruh Indonesia. Sebanyak 798.679 unit merupakan usaha kecil, proporsinya sebesar 1,22% dari total UMKM di dalam negeri. Sementara usaha menengah hanya sebanyak 65.465 unit. Jumlah itu memberi adil sebanyak 0,1% dari total UMKM (Mahdi, 2022).

Pada tahun 2020 UMKM memiliki kontribusi besar terhadap PDB (Produk Domestik Bruto) yaitu 61,97% dari total PDB nasional atau setara dengan 8.500 triliun. Dengan total penyerapan tenaga kerja dalam jumlah yang besar sebesar 97% dari daya serap dunia usaha (*Upaya Pemerintah Memajukan UMKM Indonesia*, 2020)



Gambar 2.1 Perkembangan UMKM 2015-2019 (KEMENKOP UKM)

Dari perkembangan UMKM yang terus meningkat di setiap tahunnya maka bisa menghasilkan pengetahuan bagaimana pola usaha dan penyebaran umkm yang akan meningkatkan kualitas dari umkm-umkm tersebut.

Pola Usaha dan Penyebaran UMKM

UMKM sendiri merupakan usaha produktif yang telah terbukti memberikan lapangan kerja dalam menjadi penggerak roda perekonomian di Indonesia (Puntoriza & Fibriani, 2020)

Keanekaragaman bidang usaha maupun lokasi usaha pada industri UMKM dapat mengembangkan berbagai pola usaha dari UMKM. Untuk membentuk sebuah UMKM dibutuhkan berbagai syarat sehingga usaha tersebut bisa dikategorikan sebagai UMKM. Badan Pusat Statistik (BPS) memberikan definisi UMKM berdasarkan jumlah tenaga kerja (BPS, 2013). Sementara UU 20/2008 UMKM mendefinisikan UMKM berdasarkan kekayaan bersih dan hasil penjualan tahunan (Firmansyah, 2018)

Keberhasilan usaha UMKM tidak terlepas dari peranan seorang pemilik. Pengaruh tingkat pendidikan secara keseluruhan berpengaruh positif dan signifikan terhadap pendapatan UMKM (Utari & Dewi, 2014). Sedangkan dalam menjalankan UMKM tersebut terdapat banyak kondisi yang bisa meningkatkan produktivitas dan pendapatan bagi pemilik usaha umkm, mulai dari UMKM yang *Go digital* sehingga bisa menjadi sarana untuk menjangkau pasar yang lebih luas, serta dilakukannya strategi pemberdayaan yang mampu meningkatkan daya saing UMKM (Indonesiabaik.id, 2021)

Tetapi dalam menjalankan usaha tersebut, terdapat banyak permasalahan dan hambatan yang dihadapi. Mulai dari kurangnya permodalan, kesulitan dalam pemasaran, persaingan usaha yang ketat, kesulitan bahan baku dan kendala lainnya yang bisa menghambat produktifitas dari UMKM (Bahri et al., 2019)

Terkait dengan keanekaragaman yang dapat meningkatkan penghasilan atau pendapatan sebuah UMKM. Dengan melakukan pengelompokan UMKM dapat membantu pemerintah menetapkan strategi pemasaran yang tepat sebagai prioritas utama untuk mengembangkan pasar (Puntoriza & Fibriani, 2020). Sedangkan pola usaha pada UMKM di gunakan untuk menemukan hubungan antara variabel yang dimiliki oleh sebuah UMKM.

2.2. Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual (Siburian, 2014)

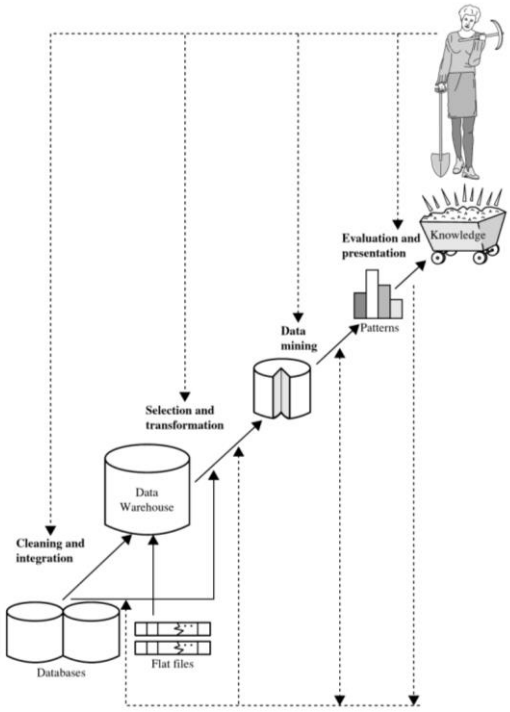
Data Mining digunakan sebagai proses penentuan struktur data yang menarik menggunakan satu atau beberapa algoritma untuk mengidentifikasikan *trend* atau pola menarik dalam satu data. *Knowledge* yang diperoleh dari model *data mining* akan digenerasikan yang tujuannya dapat diterapkan pada situasi yang baru (Roger, 2017).

Data mining juga disebut sebagai *Data Discovery* atau *Knowledge Discovery from Database* (KDD). *Knowledge Discovery from Database* (KDD) didefinisikan sebagai *ekstraksi* informasi potensial, implisit dan tidak dikenal dari sekumpulan data. Proses *Knowledge Discovery in Database* melibatkan hasil proses *Data Mining* (proses pengekstrak kecenderungan suatu pola data), kemudian mengubah hasilnya secara akurat menjadi informasi yang mudah dipahami (Mulya, 2019).

Proses *data mining* atau *Knowledge Discovery from Database* (KDD) terbagi atas 7 tahap, sebagai berikut (Han et al., 2012):

1. *Data Cleaning:* tahap menghapus atau menghilangkan data *noise* dan data tidak konsisten
2. *Data Integration*: tahap dimana beberapa sumber data dapat digabungkan
3. *Data Selection*: tahap dimana data yang relevan dengan analisis diambil dari *database.*
4. *Data Transformation:* Tahap dimana data ditransformasikan dan dikonsolidasikan ke dalam bentuk yang sesuai untuk penambangan dengan melakukan operasi *summary* atau *aggregation.*
5. *Data Mining:* Tahap penting dimana metode cerdas diterapkan untuk mengekstrak pola data.

6. *Data Evaluation*: Tahap untuk mengidentifikasikan pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan ukuran ketertarikannya (*distance/interestingness measure)*.
7. *Knowledge Presentation*: Tahap dimana teknik gambar visualisasi dan pengetahuan digunakan untuk menyajikan pengetahuan kepada pengguna.



Gambar 2.2 Tahap *Data Mining* (Jiawei Han et al, 2012)

2.3. Asosiasi

Analisis Asosiasi berguna untuk mengungkapkan hubungan yang menarik yang tersembunyi dalam dataset besar. Hubungan terungkap itu dapat dipresentasikan dalam bentuk aturan asosiasi (*association rules)* untuk himpunan item yang sering muncul

(*sets of frequent items).* Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam suatu waktu, dan berusaha untuk mengungkapkan aturan untuk mengukur hubungan antara dua atau lebih atribut (Sianturi et al., 2019)

Aturan asosiasi sering dinamakan *market basket analysis* (analisis keranjang belanja), atau bisa dinamakan aturan asosiasi dalam bentuk *"if-then"* atau "jika-maka". Aturan ini dihitung dari sekumpulan data yang sifatnya *probabilitas.* Aturan asosiasi merupakan suatu proses pada data mining untuk menentukan semua aturan *asosiatif* yang memenuhi syarat minimum untuk *support* (dukungan) dan *confidence* (kepercayaan) pada database (Musthafa & Wibowo, 2020).

Aturan asosiasi dalam bentuk "jika pendahuluan maka konsekuen," *(if antecedent, then consequent)* dengan ukuran dukungan dan kepercayaan yang berhubungan dengan aturan, sebagai contoh supermarket tertentu mungkin menemukan bahwa dari 1000 pelanggan yang berbelanja, 200 membeli popok dan 50 membeli bir, dengan demikian aturan asosiasi menjadi "Jika membeli popok maka membeli bir" dengan *support* 200/1000 = 20% dan *confidence* 50/200 = 25% (Sianturi et al., 2019).

Aturan asosiasi di atas juga dapat diartikan menjadi 25% dari transaksi di *database* yang memuat pembelian item popok juga memuat pembelian item bir. Sedangkan 20% dari seluruh transaksi yang ada di database memuat kedua item itu. Dapat juga dibahasakan menjadi "seorang konsumen yang membeli popok kemungkinan 25% untuk membeli bir. Aturan ini mewakili 20% dari seluruh catatan transaksi" (Munzir, 2021). Dengan pengetahuan seperti ini pemilik supermarket dapat

mengatur penempatan barang-barangnya dan merancang strategi pemasaran dengan pengetahuan yang diperoleh dari hasil asosiasi tersebut.

Setiap aturan asosiasi disusun oleh dua *itemset* yang berbeda, dimana item sisi kiri disebut *antecedent* dan sisi kanan disebut *consequent*, kedua item tersebut tidak tergantung satu sama lain, misalnya jika membeli popok dan mentega

Kuat tidaknya sebuah aturan asosiasi ditentukan oleh dua parameter yaitu nilai dukungan (*support)* dan nilai kepercayaan (*confidence)* yang memiliki nilai antara 0% - 100% (Rabbany, 2015).

Menurut Daniel T, Larose (2005), aturan asosiasi dari database besar terdiri dari dua langkah, yaitu (Mandala, 2017):

1. Temukan semua *frequent itemset*, yaitu menemukan semua *itemset* dengan *frekuensi* ≥ *minimum support.*
2. Dengan *frequent itemset,* buat aturan asosiasi yang memenuhi kondisi *minimum support* dan *minimum confidence*

Metode dasar aturan asosiasi dibagi menjadi dua tahap, yaitu (Purnama, 2021) :

1. Analisa pola frekuensi tertinggi

   pada tahap ini akan dicari kombinasi item yang memenuhi syarat *minimum* dari nilai *support* dalam *database*. Rumus untuk mendapatkan nilai *support* adalah sebagai berikut:

$$\text{Support (A)} = \frac{jumlah\ kejadian\ mengandung\ A}{total\ kejadian} \text{ x } 100\%$$

Persamaan 2.1

Sementara nilai *support* untuk 2 itemset diperoleh dengan menggunakan rumus sebagai berikut:

$$\text{Support (A,B)} = \frac{jumlah\ kejadian\ mengandung\ A\ dan\ B}{total\ kejadian} \text{ x } 100\%$$

Persamaan 2.2

Sementara nilai *support* untuk 3 itemset diperoleh dengan menggunakan rumus sebagai berikut:

$$\text{Support (A,B, C)} = \frac{jumlah\ kejadian\ mengandung\ A\ ,B\ dan\ C}{total\ kejadian} \text{ x } 100\%$$

Persamaan 2.3

2. Pembentukan Aturan Asosiasi

Akurasi dari suatu aturan asosiasi disebut *confidence*, *confidence* disebut dengan nilai kepastian, dimana kuatnya hubungan antara item dalam aturan asosiasi. Setelah semua pola frekuensi tertinggi telah ditemukan, baru dicari aturan asosiasi A → B "Jika A maka B. untuk memperoleh nilai *confidence*, dilakukan dengan rumus:

$$\text{Confidence} = (A \rightarrow B) = \frac{jumlah\ kejadian\ mengandung\ A\ dan\ B}{total\ kejadian\ mengandung\ A}$$

Persamaan 2.4

A → B dimana A dikatakan dengan *Antecedent* (pernyataan) dengan B merupakan *Consequences* (kesimpulan). *Anteseden* merupakan sebab yang menjadi item *consequences.* sedangkan *consequence* adalah akibat atau juga item yang akan terjadi setelah terjadi *anteseden*.

2.3.1 Algoritma Apriori

Algoritma apriori adalah suatu algoritma dasar yang diusulkan oleh Agrawal dan srikant pada tahun 1994 untuk menentukan *frequent itemset* untuk aturan *asosiasi Boolean*. Algoritma apriori adalah salah satu algoritma yang melakukan pencarian *frequent itemset* dengan menggunakan teknik *association rule* (Musthafa & Wibowo, 2020).

Algoritma Apriori merupakan algoritma yang paling populer dikenal sebagai dengan paradigma *general and test*, yaitu pembuatan kandidat kombinasi *item* yang mungkin berdasarkan aturan tertentu lalu diuji apakah kombinasi item tersebut memenuhi syarat support minimum (M. Malik et al., 2019)

Beberapa istilah yang sering digunakan dalam algoritma apriori antara lain (Fitriyanto, 2017):

1. *Support* (dukungan): probabilitas (kemungkinan) pelanggan membeli beberapa produk secara bersamaan dari seluruh transaksi. *Support* untuk aturan "X, Y" adalah probabilitas atribut atau kumpulan atribut X dan Y yang terjadi bersamaan.
2. *Confidence* (tingkat kepercayaan): probabilitas kejadian beberapa produk dibeli bersamaan dimana salah satu produk sudah pasti di beli.
3. *Minimum Support*: parameter yang digunakan sebagai batasan frekuensi kejadian atau *support count* yang harus dipenuhi suatu kelompok data untuk dapat dijadikan aturan.
4. *Minimum Confidence*: parameter yang mendefinisikan minimum *level* dari *confidence* yang harus dipenuhi oleh aturan yang berkualitas.

5. *Itemset*: kelompok produk
6. Kandidat *Itemset (C*k*): itemset-itemset* yang akan dihitung *support count*-nya.
7. *Frequent itemset*(Fk): *itemset* yang sering terjadi, atau *itemset-itemset* yang sudah melewati batas *minimum support* yang telah ditentukan.

Penggunaan Algoritma Apriori membantu dalam proses pengambilan keputusan, dengan Algoritma ini dapat membantu dalam membentuk kandidat kombinasi item yang mungkin, kemudian dilakukan pengujian apakah kombinasi tersebut memenuhi parameter *support* dan *confidence minimum* yang merupakan nilai minimum yang diberikan oleh pengguna (Nurajizah, 2019).

Terdapat dua proses utama yang dilakukan untuk mendapatkan *itemset* yang sering terjadi (*frequent itemset*), yaitu (Nurajizah, 2019):

1. *Join* (penggabungan)

   Dalam proses ini, setiap item dikombinasikan dengan item lain sampai tidak dapat terbentuk kombinasi lagi.
2. *Pruning* (pemangkasan).

   Pada proses ini, hasil kombinasi item akan dipangkas berdasarkan minimum *support* yang telah ditemukan.

Cara kerja Algoritma Apriori dapat dijelaskan sebagai berikut (Haris, 2016):

1. Tentukan *minimum support*
2. Iterasi 1: hitung *item-item* dari *support* (transaksi yang memuat seluruh item) dengan memindai database untuk 1-*itemset*, setelah *itemset* didapatkan, apabila

telah memenuhi *minimum support*, 1-*itemset* tersebut akan menjadi pola frekuensi tinggi

3. Iterasi 2: untuk mendapatkan 2-*itemset*, perlu dilakukan kombinasi dari k-*itemset* sebelumnya. Kemudian pindai database sekali lagi untuk menghitung *item-item* yang memuat *support*. *Itemset* yang memenuhi *minimum support* akan dipilih sebagai pola frekuensi tertinggi dari kandidat.
4. Tetapkan nilai k-*itemset* dari *support* yang telah memenuhi nilai *minimum support* dari k-*itemset*.
5. Lakukan proses untuk *iterasi* selanjutnya hingga tidak ada lagi k-*itemset* yang memenuhi nilai *minimum support*.

2.3.2 Lift Ratio

Lift ratio adalah suatu ukuran (parameter) untuk mengetahui kekuatan aturan asosiasi yang telah dibentuk dari nilai *support* dan *confidence*. Nilai *lift ratio* biasanya digunakan sebagai penentu apalah aturan asosiasi valid atau tidak valid (Simanjuntak & Windarto, 2020).

Lift ratio digunakan untuk mengukur seberapa penting *rule* yang telah terbentuk berdasarkan nilai *support* dan *confidence*. *Lift Ratio* merupakan nilai yang menunjukkan kevalidan proses transaksi dan memberikan informasi apakah benar produk A dibeli bersamaan dengan produk B. Lift ratio dihitung berdasarkan rumus pada persamaan 2.5 (Wicaksono, 2015) berikut:

$$\text{Lift Ratio} = \frac{Support\ (A\ U\ B)}{Support\ (A)\ x\ Support\ (B)}$$

Persamaan 2.6

Lift ratio biasanya digunakan sebagai penentu apakah aturan keakuratan suatu asosiasi, nilai *lift* yang baik yaitu jika nilai *lift* > 1 maka pada nilai ini proses transaksi dikatakan valid, karena keterkaitan item bergantung positif (Rahardjo et al., 2021).

2.3.3 Conviction

Conviction adalah perhitungan untuk menentukan nilai akurasi minimum pada metode *association rules*. Pada proses ini dihitung performansi yaitu akurasi untuk *rule* yang dihasilkan oleh sistem. Mengukur akurasi dari metode yang digunakan (Ikbal, Indwiarti & Yuliant, 2015).

Conviction digunakan untuk mengukur tingkat implikasi karena bersifat terarah dimana angka maksimal untuk implikasi sempurna dan menghitung kedua item dalam itemset secara benar. *Conviction* berbeda dengan *lift ratio* dikarenakan nilai lift ratio merupakan pengukuran yang tidak searah (lift (A→B) = lift (B→A). Sementara *conviction* sangat sensitif dimana arah *rules*-nya *conviction* (A → B) ≠ *conviction* (B → A) (Risyad & Ramantoko, 2020). Berikut merupakan rumus conviction (Ardhi, 2013)

$$\text{Conviction (A} \rightarrow \text{B)} = \frac{1-Support\ B}{1-confidence\ (A \rightarrow B)}$$

Persamaan 2.7

Nilai *range* pada *conviction* ini, berada pada nilai 0.5,…,1…∞ (*inf*) dengan ketentuan *conviction* dianggap memiliki nilai tak terhingga (*infinite*) apabila nilai dari *conf*(A → B) sama dengan 1. Sama seperti *lift*, apabila conviction menghasilkan nilai rule yang menjauh dari 1 bahkan sampai tak terhingga, maka akan dianggap semakin akurat atau *rule* tersebut memiliki tingkat kekuatan yang baik (Ardhi, 2013).

2.4. Clustering

Salah satu teknik yang dikenal dalam *data mining* yaitu *clustering*. Pengertian *clustering* keilmuan dalam data mining adalah pengelompokan sejumlah data atau objek ke dalam *cluster (group)* sehingga setiap dalam *cluster* tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam cluster yang lainnya. Tujuannya adalah menemukan *cluster* yang berkualitas dalam waktu yang layak (Alfina et al., 2012).

Clustering atau pengklasteran merupakan pengelompokan data dengan kemiripan nilai (*homogen*). Bentuk data yang bisa dikelompokan ke pengklasteran adalah hasil pengamatan, *record* data, ataupun kelas serta objek dengan kemiripan, kemiripan objek diperoleh dari kedekatan nilai-nilai *atribut* yang menjelaskan objek-objek data sedangkan objek-objek data biasanya dipresentasikan sebagai titik dalam ruangan *multidimensi* (Handayani, 2022).

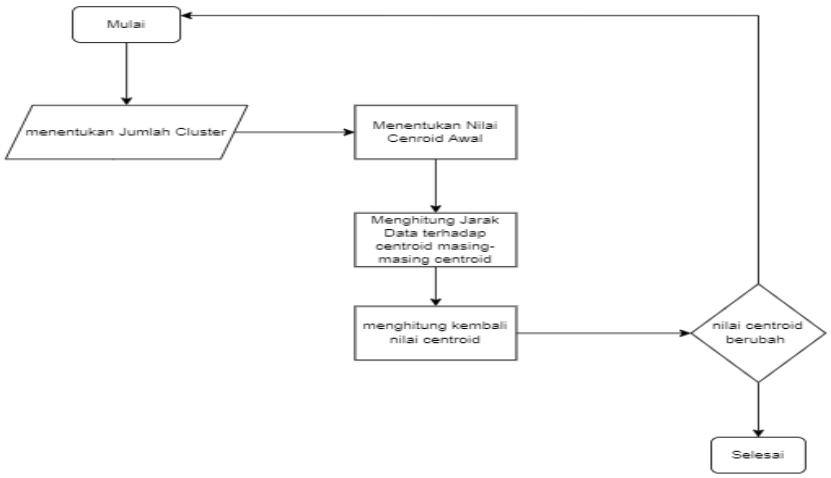
Pengelompokan benda serupa kedalam kelompok yang berbeda, atau lebih tepatnya partisi dari sebuah dataset ke dalam *subnet*, sehingga data dalam setiap *subnet* memiliki arti yang bermanfaat merupakan proses dari *clustering*. Dimana dalam *cluster*

terdiri dari kumpulan benda-benda yang mirip antara satu dengan yang lainnya dan berbeda dengan benda yang terdapat pada *cluster* lainnya. *Clustering* berusaha untuk mengidentifikasikan kelompok objek yang mirip-mirip dan membantu menemukan pola penyebaran dan pola hubungan dalam sekumpulan data yang besar (Sonang et al., 2019).

2.4.1. K-Means Clustering

K-Means clustering sendiri bisa didefinisikan sebagai suatu metode pemodelan *data mining* dan metode dalam penganalisaan data yang mengelompokkan data dengan sistem partisi atau yang melakukan proses pemetaan tanpa *supervise (unsupervised)*, dimana *k-means* berusaha mengelompokkan atau memetakan data yang dalam tiap kelompok, dimana data dari satu kelompok memiliki karakteristik yang sama dengan data lainnya, serta memiliki karakteristik yang berbeda dengan data yang terdapat didalam kelompok yang lainnya (R. A. Malik et al., 2018).

Pada penerapan metode *K-Means cluster*, data yang bisa diolah dalam perhitungan adalah data *numerik* yang berbentuk angka. Setiap data dihitung berdekatan dengan nilai *centroid* yang sudah ditentukan sebelumnya, jarak terkecil antara data dengan masing-masing *centroid* merupakan anggota *cluster* yang terdekat. Terdapat langkah-langkah perhitungan Algoritma *K-Means* dapat dilihat dari *flowchart* (Puntoriza & Fibriani, 2020) berikut:



Gambar 2.3 Tahap Data Mining (Han et al., 2012)

Tahapan dari proses analisis *dataset* menggunakan pemodelan algoritma *K-Means* ini dengan menggambarkan langkah-langkah pemodelan *k-means clustering* dari awal algoritma dimulai hingga menghasilkan pengelompokan data akhir (pada saat iterasi ke-n saat tidak terjadi perubahan pusat (*cluster/centroid*) dengan taksiran bahwa tolak ukur terhadap *input* adalah jumlah *dataset* sebanyak n data dan jumlah *inisialisasi centroid* (pusat klaster) hingga pada saat *iterasi* ke-n saat tidak terjadi perubahan pusat *cluster/centroid* menggunakan persamaan *Eucledian Distance.* Algoritma *K-means Clustering* (R. A. Malik et al., 2018) sebagai berikut:

1. Masukkan data yang akan diklaster atau dikelompokan
2. Tentukan nilai K sebagai jumlah klaster yang akan dibentuk
3. Inisialisasi k dari data sebanyak jumlah klaster secara acak sebagai pusat klaster (*Centroid*).

4. Hitung jarak antar masing-masing data dengan pusat klaster (*Centroid*), dengan menggunakan persamaan Euclidean Distance

$$d(x+y) = \sum_{k=0}^{n} \sqrt{|xn - yn|^2}$$

$$atau$$

$$d(x,y) = \sum_{k=0}^{n} \sqrt{(x1i - y1i)^2 + (x2i - y2i)^2 + \ldots + (x2n - y2n)^2}$$

Persamaan 2.8

Dimana:

d (x, y) = jarak data x ke pusat klaster j
xn = data ke n pada atribut klaster x
yn = titik pusat ke n pada atribut y
n = banyaknya objek

5. Kelompok setiap data berdasarkan jarak terdekat antara data dengan *centroid*-nya.
6. Tentukan posisi pusat klaster (*centroid*) baru (k)

   Jika pusat *cluster* tidak berubah maka proses klaster telah selesai, jika belum maka ulangi langkah ke-4 sampai pusat *cluster* (*centroid*) tidak berubah lagi.

2.4.2. Sum of Squared Error

Menurut Tippaya Thinsungnoen (2015), SSE (*sum of Squared Error*) merupakan salah satu pengukuran yang dapat digunakan untuk memvalidasi hasil pengelompokan. Saat nilai SSE dari suatu klaster semakin mendekati 0, maka titik-

titik data dalam suatu set data yang telah dipetakan ke dalam satu klaster memiliki kedekatan yang sangat signifikan. Sebaliknya jika nilai SSE semakin besar maka titik-titik data dalam suatu set data yang digabungkan ke dalam suatu klaster sangat tersebar (Nasir, 2021).

Jika nilai klaster pertama dengan klaster kedua memberikan sudut dalam grafik atau nilainya mengalami penurunan paling besar maka jumlah nilai klaster tersebut yang tepat. Untuk mendapatkan perbandingannya maka menghitung SSE (*sum of square error*) dari masing-masing klaster. Karena semakin besar jumlah nilai klaster K, maka nilai SSE akan semakin kecil (Dewi & Pramita, 2019).

$$SSE = \sum_{K=0}^{K} \sum_{Xi} |xi - ck|^2$$

Persamaan 2.9

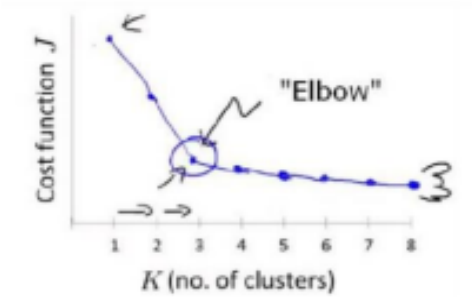
Dimana:

K = *cluster* k

xi = jarak data objek ke-i

ck = pusat *cluster* ke-i

setelah dilihat ada beberapa nilai K yang mengalami penurunan besar selanjutnya hasil dari nilai K akan turun secara perlahan-perlahan sampai hasil dari nilai K tersebut stabil. Misalnya nilai *cluster* K=2 ke K=3, kemudian dari K=3 ke K=4, terlihat penurunan drastis membentuk siku pada titik K=3. artinya K = 3 merupakan jumlah cluster yang paling baik atau optimal (Izzadin, 2020)



Gambar 2.4 Identifikasi titik elbow berdasarkan SSE (Izzadin, 2020)

Algoritma metode *Elbow* pada SSE dalam menentukan nilai K pada *K-means* (Izzadin, 2020):

1 Mulai

2 Inisialisasi awal nilai K

3 Naikkan nilai K

4 Hitung hasil *sum of square error* dari tiap nilai K

5 Melihat hasil *sum of square error* dari nilai K yang turun secara drastis

6 Tetapkan nilai K yang berbentuk siku.

2.4.3. Silhouette Score

Menurut Kaufman dan Rousseeuw (1990), salah satu metode evaluasi yang dapat digunakan untuk melihat kualitas dan kekuatan *cluster* adalah metode *Silhouette Coefficient.* Metode ini merupakan metode validasi *cluster* yang menggabungkan metode *cohesion* dan *separation* (Athifaturrofifah et al., 2019). Fungsi lainnya *Silhouette coefficient* yaitu untuk mengidentifikasikan derajat

kepemilikan dari setiap objek dalam sebuah klaster. Metode *cohesion* yang berfungsi untuk mengetahui seberapa dekat keterkaitan antar objek dalam klaster, dan *separation* yang berfungsi untuk mengetahui seberapa jauh jarak terpisahnya antar klaster (Fahmi et al., 2021).

Menghitung nilai *Silhouette Coefficient* untuk setiap data ke-i (Hidayati et al., 2021).

$$SC = \frac{b_i - a_i}{max\{a_i, b_i\}}$$

Persamaan 2.10

a_i : rata-rata jarak data ke-i dengan semua data pada satu *cluster* yang sama

b_i : rata-rata jarak data ke-I dengan semua data pada *cluster* yang berbeda

SC_1: nilai *Silhouette Coefficient* pada data ke-i

Nilai hasil *silhouette coefficient* terletak pada kisaran nilai -1 hingga 1. Jika nilai *silhouette coefficient* mendekati 1 maka semakin baik pengelompokan data dalam satu *cluster*. Sebaliknya, jika nilai *silhouette coefficient* mendekati -1 maka maka kualitas pengelompokkan data dalam suatu cluster tidak cukup baik.

Table 2.1 Interpretasi Nilai *Silhouette Coefficient*

Silhouette Coefficient	Interpretasi
0.71 – 1.00	Struktur yang dihasilkan kuat

Silhouette Coefficient	Interpretasi
0.51 – 0.70	Struktur yang dihasilkan baik
0.26 – 0.50	Stuktur yang dihasilkan lemah
≤ 0.25	Tidak terstruktur

2.5. Preprocessing

Data *Preprocessing* adalah fase awal pada proses *data mining* yang bertujuan untuk mengubah data yang tidak ideal bagi proses *data mining* menjadi ideal untuk diproses. Data pada dunia nyata sangat rentan pada beberapa noda seperti, ketidak konsistenan data, kesalahan *input* data, kesalahan pengetikan, dan lain-lain. Data pada dunia nyata berukuran terlalu besar sehingga proses *data mining* yang dilakukan kepada data tersebut akan menjadi efektif. Selain itu seringkali data pada dunia nyata tidak cukup akurat untuk menggambarkan kondisi yang terwakilkan oleh data, karena itu dalam beberapa sisi perlu adanya sistem yang menimbulkan nilai-nilai tertentu untuk meningkatkan akurasi data (Hermawan et al., 2011).

Mempersiapkan data adalah tahap penting *preprocessing* yang sangat penting pada *data mining*, alasan utamanya adalah karena kualitas dari *input* data sangat mempengaruhi kualitas *output* analisis yang dihasilkan. Ada 7 (tujuh) tahapan proses *data mining*, dimana 4 (empat) tahapan pertama disebut juga dengan data *preprocessing* (yang terdiri dari data *cleaning*, data *integration*, data *selection*, dan data

transformation), yang dalam implementasinya membutuhkan sekitar 60% dari keseluruhan proses (Junaedi et al., 2011).

Hal mendasar yang harus disiapkan dalam penyiapan data, dapat dijelaskan (García et al., 2015) sebagai berikut:

- Membersihkan data → Data *Cleaning*

  Data *cleaning* atau data *cleansing* termasuk operasi data yang buruk, menyaring data yang salah dari kumpulan data dan mengurangi detail data yang tidak diperlukan. Perlakuan terhadap *missing* data (data hilang) dan *noise* data (data yang menyimpang / kebisingan data) terdapat dalam proses ini. Tugas data *cleaning* lainnya adalah mendeteksi perbedaan dan data kotor (fragmen dari data asli yang tidak masuk akal).

- Menggabungkan dan menyesuaikan data → Data *Integration*

  Dalam proses mengintegrasi data, ini merupakan penggabungan data dari beberapa penyimpanan data. Proses ini harus dilakukan dengan hati-hati agar menghindari redudansi dan inkonsistensi dalam kumpulan data yang dihasilkan. Penggabungan data terutama dilakukan pada data yang memiliki representasi berbeda karena berbagai hal pada sebuah data.

- Memberikan data yang akurat → Data *Transformation*

  Pada tahap *preprocessing* ini data yang telah diintegrasikan, kemudian dinormalisasi dan dilakukan proses generalisasi data, proses ini memastikan tidak adanya data yang berlebihan. Sehingga data akan dihimpun dalam sebuah tempat

penyimpanan, yang dependensinya harus masuk akal, data juga akan ditransformasikan dalam bentuk yang sesuai.

Pada tahap ini pula data akan dikonversikan atau diubah bentuknya. contohnya seperti mengubah bentuk data angka menjadi suatu bentuk kategori yang berbeda. Tahapan transformasi data ini berguna untuk mengurangi jumlah data. Cara-cara transformasi antara lain: *aggregate* data, *smoothing*, *attribute construction*, normalisasi data dan *discretization* data.

- Menyatukan dan menskalakan data → Data *Normalization*

  Satuan pengukuran yang digunakan dapat mempengaruhi analisis data. Semua atribut data harus dinyatakan dalam satuan pengukuran yang sama dan harus menggunakan skala atau rentang yang sama. Normalisasi data mencoba untuk memberikan semua bobot atribut yang sama.

- Menangani data yang hilang → *Missing Data Imputation*

  Proses ini berisi pembersihan data, dimana tujuannya adalah mengisi variabel yang berisi *missing values* dengan beberapa data intuitif. Dalam sebagian besar kasus, menambahkan perkiraan yang masuk akal dari nilai data yang sesuai lebih baik daripada membiarkannya kosong.

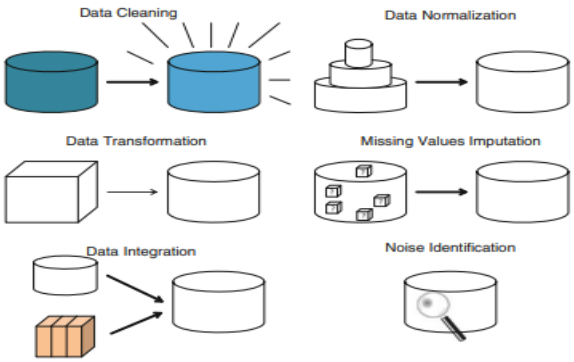
- Mendeteksi dan mengelolah *noise* (kebisingan) → *Noise Identification*

  Pada tahap ini termasuk dalam langkah data *cleaning* dan juga dikenal sebagai pemulusan dalam transformasi data. Tujuan utamanya adalah mendeteksi

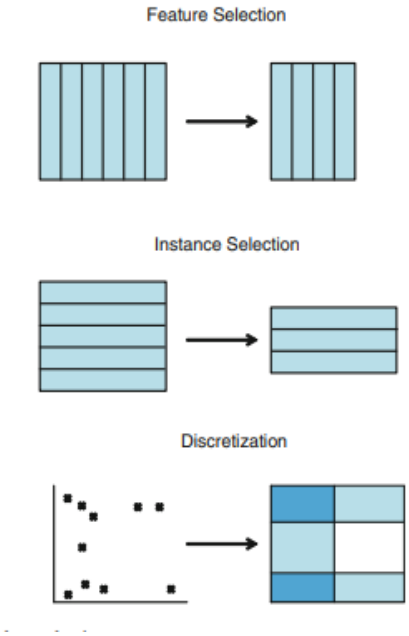
kesalahan acak atau varians dalam variabel yang diukur. Pada proses ini akan dideteksi noise daripada penghilangan noise.

- Data *Reduction*

  Pada tahapan ini, data reduksi atau mereduksi data. karena jumlah data yang besar, maka tingkat akurasinya pun dikhawatirkan akan rendah atau bahkan tidak akurat. Sebab itulah diperlukannya reduksi data, reduksi merupakan suatu proses untuk mengurangi jumlah data namun dengan mempengaruhi proses analisis data. Dengan menggunakan pengurangan data, maka akan menjadikan proses penyimpanan menjadi lebih efisien. Dalam kasus reduksi data, data yang dihasilkan biasanya mempertahankan struktur dan integrasi penting dari data asli, tetapi jumlah data dirampingkan. Proses pada reduksi data dapat dijelaskan sebagai berikut, dimana *Feature Selection* (FS) adalah mengurangi dimensi data. *Instance Selection* (IS) adalah menghapus contoh yang berlebihan dan bertentangan. *Discretization* adalah bagaimana cara menyederhanakan domain dari sebuah atribut, lalu yang terakhir ialah *Feature Extraction* dan/atau *Instance Generation* dimana bagaimana cara mengisi kekosongan data.



Gambar 2.5 Bentuk persiapan data (García et al., 2015)



Gambar 2.6 Bentuk Data Reduksi (García et al., 2015)

2.6. Principal Component Analysis

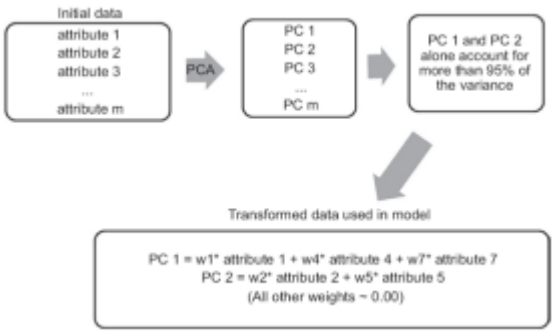
Principal component analysis (PCA) atau disebut juga transformasi Karhunen-Loeve adalah teknik yang digunakan untuk menyederhanakan suatu data, dengan cara

mentransformasikan *linear* sehingga terbentuk sistem koordinat baru dengan variasi maksimum. PCA dapat digunakan untuk mereduksi dimensi suatu data tanpa mengurangi karakteristik data tersebut secara signifikan. Metode ini mengubah dari sebagian besar variabel asli yang saling berkorelasi menjadi satu himpunan variabel baru yang lebih kecil dan saling bebas (tidak berkorelasi lagi). Prinsip dasar dari algoritma *Principal Component Analysis* adalah mengurangi satu set data namun tetap mempertahankan sebanyak mungkin variasi dalam set data tersebut. Secara matematis PCA mentransformasikan sebuah variabel yang berkorelasi ke dalam bentuk yang bebas tidak berkorelasi (Firliana et al., 2015).

Metode PCA sangat berguna digunakan jika data yang ada memiliki jumlah variabel yang besar dan memiliki korelasi antar variabelnya. Perhitungan dari *principal component analysis* didasarkan pada perhitungan nilai *eigen* dan *vector eigen* yang menyatakan penyebaran data dari suatu *dataset*. Dengan menggunakan PCA, variabel yang tadinya sebanyak n variabel akan diseleksi menjadi k variabel baru yang disebut *principal component*, dengan jumlah k lebih sedikit dari n. dengan menggunakan k *principal component* akan menghasilkan nilai yang sama dengan menggunakan n variabel. Variabel hasil seleksi disebut *principal component* atau PC (Nasution et al., 2019).

Menurut kotu dan despande pada buku Predictive Analytics and Data Mining, 2015 PCA digunakan untuk menjelaskan struktur matriks *varians-kovarians* dari suatu set variabel melalui kombinasi *linier* dari variabel-variabel tersebut. Secara umum *principal component* (PC) dapat berguna untuk seleksi fitur dan interpretasi variabel-

variabel skema konseptual yang diilustrasikan bagaimana PCA dapat membantu untuk yang diilustrasikan bagaimana PCA dapat membantu untuk menyederhanakan dimensi dari data melalui hipotesis *dataset* berjumlah m variabel dapat ditunjukkan (Nasution, 2019) sebagai berikut:



Gambar 2.7 Model konseptual PCA untuk tahap seleksi fitur (Kotu & Deshpande, 2015)

Menurut jolliffe (2002), Prosedur pengerjaan *Principal Component Analysis* bertujuan untuk menyederhanakan dan menghilang *factor* yang kurang dominan dan kurang relevan tanpa mengurangi maksud dan tujuan dari data asli dari variabel acak x (matrik berukuran n x n, dimana baris-baris yang berisi observasi sebanyak n dari variabel acak x) adalah (Nasution et al., 2019) sebagai berikut:

1. Menghitung matrik *covariance matrix* dari pada observasi.

   Varians (Var(x)) dihitung untuk menemukan penyebaran data dalam set data untuk menentukan penyimpangan data dalam set data sampel. Matriks **kovarian** Cov (x, y) ialah matriks yang nilai-nilai *kovariansi* pada tiap *cell*-nya diperoleh dari sampel. Misalkan x dan y adalah variabel acak.

$$\text{Var(x)} = \frac{1}{2}\sum_{i=1}^{n}(Z_{ij} - \mu_{yj})$$

$$\text{Cov (x,y)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_{xj})(y_{ij} - \mu_{yj})$$

Persamaan 2.11

Dimana μ merupakan rata-rata dari variabel yang bersangkutan, dan dapat ditemukan menggunakan persamaan dibawah ini:

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

Persamaan 2.12

Dengan μ_x dan μ_y merupakan rata-rata (*mean*) yang dari variabel x dan y, dimana variabel x_i dan y_i merupakan nilai observasi ke-i dan variabel x dan y. dari data nilai yang digunakan, maka diperoleh matriks kovarian berukuran n x n. Matriks kovarians yang ditemukan akan memiliki nilai positif dan negatif, nilai positif berarti korelasi positif antara kedua variabel.

2. Mencari nilai *eigenvalues* dan *eigenvector* dari matriks *kovarian* yang telah diperoleh, yaitu: nilai eigen dan vektor eigen untuk matriks *kovarians* dihitung. Nilai eigen yang dikomputasi kemudian ditransformasikan (*rotasi orthogonal varimax*) menggunakan persamaan di bawah ini:

$$\text{Det } (A - \lambda I) = 0$$

Dimana:

A = matrix n x n
λ = matrix eigenvalue
I = matriks identitas (matriks persegi dengan elemen diagonal utamanya bernilai 1 sedangkan element lain bernilai 0

3. Menentukan nilai proporsi *principal component* (proporsi Principal Component (%) dengan persamaan:

$$\text{PC}(\%) = \frac{NilaiEigen}{VarianeCivarian}\ x\ 100\%$$

4. Menghitung bobot faktor (*factor loading*) berdasarkan *eigenvector* dengan persamaan:

$$Ax = \lambda x$$

Sehingga diperoleh kombinasi linear yaitu:

a. $\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_n$ adalah *eigenvalue* matrix A
b. $x_1, x_2, x_3 \ldots x_n$ adalah *eigenvector* sesuai dengan *eigenvalue*-nya (λ_n)

Persamaan *eigenvalue* & *eigenvector* merupakan *Eigen Value Decomposition* (EVD), dengan persamaan sebagai berikut:

$$AX = XD$$

$$A = X\ D\ X^{-1}$$

Dimana:

A = matrix n x n yang memiliki b *eigenvalue* (λ_n)

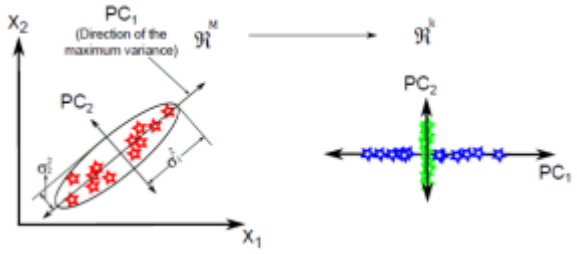
D = *eigenvalue* dari *eigenvector*-nya

X = *eigenvector* dari matriks A

$X^{-1} =$ invers dari *eigenvector* X

Menurut Alaa Tharwat (2017), Saat membangun ruang dimensi yang baru dengan PCA, pada tahap akhir, titik-titik data dipusatkan pada nol yang merupakan

titik pusat baru pada ruang PCA berdasarkan rata-rata setiap variabel. Seringkali terjadi kondisi meski data matriks masukan semuanya bernilai positif, hasil akhir PCA menunjukkan adanya nilai yang negatif. Nilai negatif pada hasil akhir PCA memiliki 2 penyebab yang pertama menandakan bahwa nilai data asli lebih kecil atau lebih rendah dari rata-rata variabelnya sehingga matriks *mean-centering* data yang diproyeksi terhadap *principal components* menjadi negatif. Yang kedua hanya menandakan posisi titik data pada *principal components* di ruang PCA berdasarkan hasil proyeksi *mean-centering* data terhadap vektor eigen (*principal components*) (Nasir, 2021). Gambar 2.6 menunjukkan ilustrasi dari pembentukan PCA (Nasir, 2021) sebagai berikut:



Gambar 2.8 Ilustrasi pembentukan ruangan PCA (Alla Tharwat, 2017)