SKRIPSI

IMPLEMENTASI ALGORITMA *CLUSTERING* DAN ASOSIASI PADA SISTEM INFORMASI SUMBER DAYA MANUSIA UNTUK MENILAI KINERJA KARYAWAN

Disusun dan Diajukan Oleh:

ROFIFAH NURUL ANNISA

D121181508

PROGRAM STUDI TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS HASANUDDIN

MAKASSAR

2022

LEMBAR PENGESAHAN SKRIPSI

IMPLEMENTASI ALGORITMA CLUSTERING DAN ASOSIASI PADA SISTEM INFORMASI SUMBER DAYA MANUSIA UNTUK MENILAI KINERJA KARYAWAN

Disusun dan diajukan oleh

ROFIFAH NURUL ANNISA

D121181508

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Informatika Fakultas Teknik Universitas Hasanuddin pada tanggal 19 Oktober 2022 dan dinyatakan telah memenuhi syarat kelulusan.

Menyetujui,

Pembimbing Utama,	Pembimbing Pendamping,
Dr. Amil Ahmad Ilham, S.T., MIT.	Iqra' Aswad, S.T., M.T
Nip. 197310101998021001	Nip. 199011282019043001

Ketua Program Studi,

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

PERNYATAAN KEASLIAN

Yang bertanda tangan dibawah ini :

Nama	: Rofifah Nurul Annisa
Nim	: D121181508
Program Studi	: Teknik Informatika
Jenjang	: S1

Menyatakan dengan ini karya tulisan saya berjudul:

IMPLEMENTASI ALGORITMA CLUSTERING DAN ASOSIASI PADA SISTEM INFORMASI SUMBER DAYA MANUSIA UNTUK MENILAI KINERJA KARYAWAN

Adalah karya tulisan saya sendiri dan bukan merupakan pengambilan alihan 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, 25 Oktober 2022

Yang Menyatakan,

Rofifah Nurul Annisa

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa karena atas bantuan dan limpahan karunia-Nya maka penulis dapat menyelesaikan tugas akhir ini.

Tentunya penulis menyadari bahwa tugas akhir ini 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 menyadari, bahwa tugas akhir ini tidak dapat diselesaikan tanpa adanya bantuan serta bimbingan dari berbagai pihak. Untuk itu, penulis menyampaikan ucapan terima kasih yang sedalam-dalamnya kepada:

1. Tuhan Yang Maha Esa, atas segala rahmat dan karunia-Nya yang diberikan kepada penulis hingga saat ini;
2. Kedua orang tua yang selalu memberikan doa, semangat dan dukungan kepada penulis;
3. Bapak Dr. Amil Ahmad Ilham, S.T., M.IT. selaku pembimbing I dan Bapak Iqra Aswad, S.T., M.T. selaku pembimbing II yang sudah menyediakan waktu, tenaga, pikiran dan perhatian dalam mengarahkan penulis dalam penyusunan tugas akhir;
4. Bapak Dr. Eng. Muhammad Niswar, S.T., M.IT. dan Ibu Anugrayani Bustamin, S.T., M.T. yang telah menyempatkan waktunya untuk memberikan saran kepada penulis;
5. Bapak dan Ibu Dosen Departemen Teknik Informatika yang sudah memberikan ilmu dan nasihatnya serta membantu selama perkuliahan;

6. Teman – teman di teknik informatika 2018 (Endang, Caca, Tami, Ayu, Pirda, Nana, Dee, Dea, Bela, Hikmah, Kamtina) yang sudah membantu dan menemani sejak maba;
7. Teman – teman di teknik informatika 2018 (Ayi, Iffat, Tamara, Ayu, Dei, Rahma) atas dukungan dan bantuannya khususnya saat proses pengerjaan skripsi ini;
8. Serta seluruh teman – teman Synchronous 18 yang belum sempat dituliskan namanya, terima kasih atas dukungan, bantuan dan semangatnya selama ini;
9. Teman seperjuangan penulis di SMA 5 Makassar, Delilah dan Salwa yang hingga saat ini senantiasa memberikan semangat dan menemani penulis;
10. Seluruh kakak – kakak volunteer KPAJ (Komunitas Peduli Anak Jalanan) Kota Makassar yang juga telah senantiasa memberikan dukungan dan menebarkan semangat kepedulian;
11. Seluruh adik binaan KPAJ, khususnya adik binaan area Manggala yang telah menyadarkan penulis untuk lebih memaknai hidup;
12. Seluruh pihak yang belum sempat saya tuliskan satu – persatu, terima kasih untuk waktu, tenaga dan pikirannya

Penulis berharap, adanya tugas akhir ini dapat memberikan manfaaat bagi seluruh masyarakat.

Makassar, Oktober 2022

Penulis

ABSTRAK

Di tengah persaingan bisnis yang semakin ketat, perusahaan perlu mengelola sumber daya manusia secara optimal. Evaluasi kinerja sumber daya manusia secara manual membutuhkan waktu yang lebih lama dan membutuhkan banyak sumber daya. Untuk itu sistem yang ada perlu dikembangkan sistem menjadi sistem prediksi analitik dengan menggunakan algoritma *clustering* dan asosiasi untuk mempermudah dalam proses pengambilan keputusan, sehingga penelitian ini dibuat untuk mengimplementasikan algoritma *clustering* dan asosiasi pada sistem informasi SDM untuk menilai kinerja pegawai. Data yang digunakan dalam penelitian ini berasal dari data penilaian kinerja karyawan. Penelitian ini menggunakan K-Means sebagai algoritma *clustering,* Apriori sebagai algoritma asosiasi, dan bahasa pemrograman *python*. Dari hasil *clustering* terdapat 4 klaster karyawan dengan klaster 2 sebagai klaster terbaik karena memiliki nilai variabel rata-rata tertinggi dan klaster 4 memiliki nilai rata-rata yang rendah dibandingkan dengan kelompok lain. Hasil klaster dengan penilaian terbaik dapat dihitung kembali untuk mendapatkan karyawan terbaik untuk direkomendasikan mendapatkan *reward.* Hasil *clustering* digunakan sebagai dataset untuk membuat aturan asosiasi dengan minimum *support* = 0,55 dan *confidence* = 0,9, tiap klaster akan memiliki aturan asosiasi akhir. 3 klaster memiliki *performance* dan *key performance indicator* sebagai variabel yang paling berpengaruh. Klaster lainnya memiliki keaktifan bersama dan apresiasi sebagai variabel yang paling berpengaruh. Kedua algoritma ini diimplementasikan pada sistem informasi. Berdasarkan penelitian yang dilakukan, sistem kini dapat memperkirakan hasil kinerja pegawai dan menentukan hubungan antar variabel pada setiap klaster dari penerapan algoritma *k-means* dan apriori pada sistem.

Kata kunci – *clustering, K-means,* asosiasi*, apriori,* sistem informasi

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1. Latar Belakang

Kompetisi bisnis dalam perusahaan saat ini semakin ketat. Untuk itu tiap perusahaan berusaha memaksimalkan potensi yang dimiliki agar terus berkembang. Salah satu faktor yang menopang keberhasilan suatu perusahaan adalah sumber daya manusia yang dimiliki. Perusahaan perlu mengelola sumber daya manusia dengan optimal agar dapat memudahkan dalam pengambilan keputusan yang berkaitan dengan sumber daya manusia dan penyesuaian dalam mencapai tujuan bisnis. Salah satu contoh pengelolaan sumber daya manusia adalah dengan mengevaluasi kinerja dengan memberikan penilaian bagi tiap sumber daya manusia (Zhao, 2020). Evaluasi kinerja sumber daya manusia secara manual ini membutuhkan waktu yang lebih lama dan sumber daya yang banyak. Terlebih lagi jika data karyawan yang dimiliki jumlahnya banyak.

Sistem informasi SDM yang ada saat ini mayoritas bersifat statistik deskriptif yaitu sistem yang menyajikan gambaran umum dari data atau hanya berfokus untuk menguraikan serta memberikan keterangan-keterangan mengenai suatu data. Untuk dapat melakukan pengembangan lebih lanjut dari data yang ada pada sistem, dibutuhkanlah sistem informasi yang bersifat analitik prediktif. Sistem analitik prediktif yaitu sistem informasi yang menggunakan penambangan data (*data mining*) untuk menganalisis data

historis yang ada (Kumar & L., 2018). *Data mining* akan mengolah data historis yang ada pada sistem kemudian menghasilkan suatu pengetahuan baru.

Penelitian dalam melakukan *clustering* dengan k-means pada data human resource telah dilakukan pada tahun 2020. Penelitian ini menyimpulkan bahwa manajemen sumber daya manusia memang harus ditingkatkan lagi. Aplikasi manajemen ini harus memanfaatkan teknologi agar dapat menganalisis, merancang, mengatur dan melakukan control terhadap sumber daya manusia (Zhao, 2020). Dalam studi lainnya, telah dilakukan penelitian untuk mengimplementasikan algoritma *clustering* dalam sistem informasi kepegawaian yang menghasilkan 3 klaster kelompok pegawai sesuai dengan masa pensiunnya (Johar dkk., 2019) . Pada penelitian lainnya, Gita Indah melakukan penelitian untuk mengevaluasi *performance* siswa menggunakan dua metode *data* mining yaitu asosiasi dan *clustering.* Penelitian ini menggunakan asosiasi untuk mengetahui variabel yang berhubungan dengan masa studi dan *clustering* untuk mengelompokkan kelompok siswa. Hasil penelitian ini menunjukkan bahwa kedua *data mining* ini memberi hasil yang saling mendukung (Marthasari, 2017).

Berdasarkan hasil observasi, sistem informasi SDM (sumber daya manusia) yang ada pada Kalla saat ini bersifat statistik deskriptif, yaitu hanya untuk menyimpan data dan menyajikan deskripsi dari data karyawan yang ada. Selain itu, pengambilan keputusan seperti penilaian kinerja dan pemberian r*eward*, masih menggunakan *excel.* Karena itu dibutuhkan

pengembangan sistem menjadi sistem analitik prediktif dengan menggunakan algoritma *clustering* dan asosiasi untuk memudahkan proses pengambilan keputusan, maka dibuatlah penelitian ini untuk mengimplementasikan algoritma *clustering* dan asosiasi pada sistem informasi SDM untuk menilai kinerja karyawan.

1.2. Rumusan Masalah

1. Bagaimana mengimplementasikan algoritma *data mining* pada sistem informasi SDM yang sifatnya statistik deskriptif sehingga menjadi sistem informasi yang sifatnya analitik prediktif?
2. Bagaimana memanfaatkan hasil analitik prediktif sistem informasi dalam pengambilan keputusan penilaian kinerja karyawan ?

1. 3. Tujuan

1. Untuk mengimplementasikan algoritma *data mining* pada sistem informasi yang sifatnya statistik deskriptif sehingga menjadi sistem informasi yang sifatnya analitik prediktif
2. Untuk memanfaatkan hasil analitik prediktif sistem informasi dalam pengambilan keputusan penilaian kinerja karyawan

1. 4. Manfaat

Penelitian ini bermanfaat bagi perusahaan untuk mengelola serta menentukan kinerja sumber daya manusia sehingga dapat mengefisienkan penggunaan waktu dan sumber daya.

1. 5. Batasan Masalah

Penilaian kinerja yang dimaksud dalam penelitian ini yaitu berdasarkan parameter ketercapaian KPI *(Key Performance Indicator), performance*, nilai *learning point*, nilai kompetensi, nilai *behaviour* tiap karyawan sesuai *value* perusahaan (kerja ibadah, lebih cepat lebih baik, aktif bersama dan apresiasi). Data yang digunakan adalah data penilaian karyawan pada tahun 2021.

1. 6. Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini menjelaskan mengenai latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah penelitian, serta sistematika penulisan

BAB II TINJAUAN PUSTAKA

Bab ini membahas landasan teori yang membangun kerangka berpikir serta membantu menganalisis dan menyelesaikan masalah yang diteliti serta metode dan variabel lain yang digunakan dalam penelitian.

BAB III METODOLOGI PENELITIAN

Bab ini membahas mengenai tahapan penelitian, instrumen penelitian, pengumpulan dan pengolahan data, perancangan sistem, implementasi dari data dengan algoritma *clustering* dan asosiasi serta sistem informasinya.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi tentang pembahasan hasil dari implementasi algoritma dan sistem yang berhasil dibangun

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 PUSTAKA

2.1. Data Mining

Data mining adalah proses menemukan pola menarik dari sejumlah besar data statistik sehingga didapatkan pengetahuan baru. *Data mining* dapat digunakan pada semua jenis statistik selama statistik tersebut memiliki tujuan yang mencakup statistik basis data, statistik *warehouse*, statistik transaksional dan lainnya (Pei dkk., 2012)

Data Mining sudah banyak diterapkan dalam berbagai aplikasi. Salah satu contoh penerapan data mining yang popular dan berhasil yaitu *business intelligence. Business intelligence* berfungsi untuk mengamati operasi bisnis yang lalu, saat ini, maupun memperediksi kegiatan bisnis yang akan datang. Banyak bisnis yang tidak bisa memberikan hasil yang optimal dalam analisis market, membandingkan *feedback* pelanggan, melihat kekuatan dan kelemahan perusahaan serta menentukan keputusan bisnis tanpa adanya *data mining* (Pei dkk., 2012)

Alur proses dalam *data mining* adalah sebagai berikut :

a. Data Selection

Proses *data selection* yaitu proses memilih data yang relevan dengan tujuan yang ingin dicapai.

b. *Preprocessing Data*

Preprocesing data seperti menghilangkan *noise* dalam data, mengecek nilai data yang kosong serta mengecek integritas data.

c. *Transformation*

Transformasi data yaitu normalisasi data yang bertujuan agar tiap variabel data memiliki skala yang sama.

d. *Data Mining*

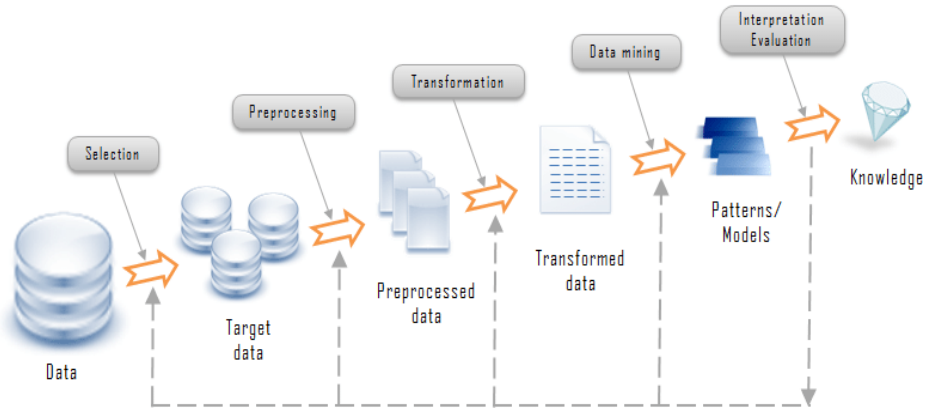
Proses pengaplikasian metode *data mining* untuk mendapatkan pola dalam data.

e. Evaluasi

Mengevaluasi hasil yang didapatkan untuk mendapatkan pola yang benar-benar sesuai dan menarik.

f. *Knowledge*

Representasi dari hasil pengetahuan yang didapatkan sehingga dapat dimengerti.



Gambar 2. 1 Alur proses dalam data mining

Seperti yang terlihat dalam Gambar 2. *1*. bahwa proses *data mining* akan menghasilkan suatu pola dalam data. Ada beberapa metode dalam *data*

mining yang dapat digunakan untuk mendapatkan pola ini, seperti klasifikasi, *clustering* dan asosiasi

2.2. Clustering

Analisis klaster adalah proses membagi sekumpulan objek data ke dalam himpunan bagian. Setiap *subset* adalah sebuah klaster, sehingga objek-objek dalam sebuah klaster memiliki kemiripan satu sama lain, namun berbeda dengan objek di klaster lain. Himpunan klaster yang dihasilkan dari analisis klaster dapat disebut sebagai *clustering.* Dalam konteks ini, metode pengelompokan yang berbeda dapat menghasilkan pengelompokan yang berbeda pada kumpulan data yang sama. Pembagian data ini dilakukan oleh algoritma *clustering*. Oleh karena itu, pengelompokan ini berguna karena dapat mengarah pada penemuan kelompok yang sebelumnya tidak diketahui dalam data (Pei dkk., 2012)

Clustering biasa juga disebut *data segmentation* karena *clustering* akan membagi sekumpulan data yang besar menjadi beberapa kelompok sesuai dengan kesamaannya. Data *clustering* berkontribusi dalam bidang penelitian *data mining*, statistik, *machine learning* dan area aplikasi lainnya (Pei dkk., 2012)

2.3. Algoritma Clustering

Dalam clustering terdapat beberapa algoritma yang dapat digunakan, salah satunya yaitu algoritma *k-means.* Algoritma *k-means* merupakan algoritma yang digunakan untuk mengelompokkan data berdasarkan centroid

yang sudah ditetapkan. Centroid adalah nilai rata – rata dari obyek yang ada di klaster (Pei dkk., 2012). *K-means* juga termasuk ke dalam metode *clustering* non-hirarki yang mengelompokkan data yang memiliki karakteristik sama (Johar dkk., 2019). *K-means* membagi data menjadi sejumlah 'k' klaster lalu pengelompokannya dilakukan dengan memperkecil kuadrat jarak. Perhitungan jarak kuadrat ini salah satunya bisa dengan menggunakan *Euclidean distance* (Bhatia, 2019).

Proses algoritma *k-means* diawali dengan menentukan jumlah klaster dari sekumpulan data. Setelah itu, akan dilakukan proses penentuan centroid dari tiap klaster. Penentuan centroid ini biasanya dilakukan dengan memberi nilai acak terlebih dahulu. Misalkan jumlah klaster (k) yang ditentukan adalah 4, maka di tiap klaster mulai klaster 1, 2, 3 dan 4 akan ditentukan titik *centroid* sebagai titik *centroid* awal. Langkah selanjutnya yaitu dengan menghitung jarak dari data sampel terhadap tiap centroid dengan menggunakan *Euclidean distance.* Kemudian, data akan dimasukkan ke dalam kelompok yang memiliki jarak *centroid* terkecil dengan data. Proses ini akan diulang dan dilakukan untuk setiap data yang ada (Bhatia, 2019)

Setelah tiap data sudah dihitung jaraknya dan dikelompokkan dalam *centroid* awal, maka Akan dilakukan perubahan *centroid*. Tiap klaster akan menghitung kembali *centroid*nya dengan menghitung rata-rata dari nilai data yang ada di kelompok tersebut.

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

Keterangan:

n = banyaknya data

x_i= nilai data x ke - i

Proses perhitungan jarak data terkecil dan *update centroid* ini akan dilakukan terus menerus hingga anggota tiap klaster tidak ada lagi yang mengalami perubahan. (Bhatia, 2019)

2.4. *Euclidean Distance*

Perhitungan jarak data ke centroid dalam k-means bisa dihitung menggunakan *euclidean distance*. Jarak dari 2 titik dalam koordinat (x,y) dan (a,b) dapat dituliskan melalui persamaan berikut :

$$euclidean\_dist\ _{((x,y),(a,b))} = \sqrt{(x-a)^2 + (y-b)^2} \qquad (2.\ 2)$$

2.5. *Elbow Method*

Metode Elbow ini memberikan ide atau wawasan dengan memilih nilai klaster kemudian menambahkan nilai klaster untuk digunakan sebagai model data untuk menentukan klaster terbaik. Dan juga, persentase dari perhitungan yang dihasilkan menjadi perbandingan antara jumlah klaster yang ditambahkan. Hasil persentase yang berbeda dari setiap nilai klaster dapat ditampilkan dengan menggunakan histogram sebagai sumber informasi. Jika nilai grup pertama sama dengan nilai grup kedua untuk sudut pada grafik, atau jika nilai tersebut paling banyak mengalami penurunan, maka nilai grup tersebut adalah yang terbaik. Untuk mendapatkan perbandingan tersebut harus menghitung SSE *(Sum of Squared Error)* dari setiap nilai klaster. Jika

jumlah klaster k semakin besar maka nilai SSE dari klaster k tersebut akan semakin kecil (Muningsih & Kiswati, 2018).

Sum of Squared Error (SSE) merupakan rumus yang digunakan untuk mengukur selisih antara data yang diperoleh dengan model prediksi yang telah dilakukan sebelumnya. SSE sering dijadikan acuan penelitian dalam menentukan klaster yang optimal. Rumus dari SSE dapat dilihat pada perhitungan berikut:

$$SSE = \sum_{i=1}^{n} (d)^2 \quad (2.\ 3)$$

dimana d adalah jarak data dengan centroid klaster (Nainggolan dkk., 2019).

2.6. *Silhouette Score*

Untuk mengukur kualitas suatu *clustering,* kita dapat menggunakan nilai koefisien siluet rata-rata dari semua objek dalam kumpulan data. Koefisien siluet dan ukuran intrinsik lainnya juga dapat digunakan dalam metode siku untuk secara heuristik menurunkan jumlah klaster dalam kumpulan data dengan mengganti jumlah varians dalam klaster. Nilai koefisien siluet adalah antara -1 dan 1. Ketika nilai koefisien siluet o mendekati 1, klaster yang mengandung o kompak dan o jauh dari klaster lain, yang merupakan kasus yang lebih disukai. Namun, ketika nilai koefisien siluet negatif (yaitu, b(o) < a(o)), ini berarti bahwa, dengan harapan, o lebih dekat ke objek di klaster lain daripada ke objek di klaster yang sama dengan o. Dalam banyak kasus, ini adalah situasi yang buruk dan harus dihindari (Pei dkk., 2012)

2.7. *Principal Component Analysis* (PCA)

Principal Component Analysis (PCA) digunakan untuk meletakkan dimensi rendah ruang ke data point pada ruang dimensi tinggi (Rene dkk., 2016). Dengan kata lain, PCA digunakan untuk mengurangi dimensi (jumlah variabel) dari sejumlah besar variabel yang saling terkait tetapi juga tetap berusaha mempertahankan sebanyak mungkin informasi *(variance).* PCA menghitung satu set variabel yang tidak berkorelasi (komponen atau pc). Perhitungan PCA direduksi menjadi masalah *eigenvalue* dan *eigenvector*. *(Hintze, 2007).* Misalkan terdapat vector X acak seperti pada persamaan

$$\left(X = \begin{matrix} X_1 \\ \vdots \\ X_p \end{matrix}\right) \tag{2. 4}$$

dengan matriks variance-covariance sebagai berikut :

$$var(X) = \sum = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix} \tag{2. 5}$$

Masing - masing dapat dianggap sebagai regresi linier, memprediksi Y_i dari $X_1, X_2, \dots, X_p$. Tidak ada intersep, tetapi $e_{i1}, e_{i2}, \dots, e_{ip}$ dapat dilihat sebagai koefisien regresi. Varians populasi dari Y_i (yang merupakan data acak) yaitu

$$var(Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} \sigma_{kl} = e_i' \sum e_i \tag{2. 6}$$

Selain itu, kovarians populasi dari Y_i dan Y_j dapat dihitung

$$cov(Y_i, Y_j) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{jl} \sigma_{kl} = e_i' \sum e_j \tag{2. 7}$$

Kemudian koefisien e_{ij} ini dapat ditulis dalam vector

$$e_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{ip} \end{pmatrix} \quad (2.\ 8)$$

Dari prosedur tersebut bisa didapatkan PCA1, PCA2 hingga PCAn. Dimulai dari PCA1 (Y_1), yang merupakan kombinasi linear dari variabel x yang memiliki varians maksimum. Sedangkan PCA2 (Y_2) adalah kombinasi linear dari variabel x yang menyumbang sebanyak mungkin variasi yang tersisa dengan batasan bahwa korelasi antara komponen pertama dan kedua adalah 0. Kemudian untuk komponen berikutnya, semua komponen utama tetap memiliki sifat yang sama yaitu kombinasi linear yang memperhitungkan sebanyak mungkin variasi yang tersisa dan tidak berkorelasi lagi dengan komponen utama lainnya (Hintze, 2007)

2.8. Aturan asosiasi

Aturan asosiasi dalam *data mining* adalah metode yang digunakan untuk melihat hubungan antar item. Aturan asosiasi terdiri dari mencari *frequent itemset* (set item, seperti A dan B, memenuhi ambang batas *minimum support threshold*, atau persentase tupel tugas yang relevan), dari aturan asosiasi yang kuat dalam bentuk A => B yang dihasilkan. Aturan-aturan ini juga memenuhi ambang batas *minimum confidence threshold.* Asosiasi dapat dianalisis lebih lanjut untuk mengungkap aturan korelasi, yang menyampaikan korelasi statistik antara himpunan item A dan B (Pei dkk., 2012)

Pengukuran kekuatan dan akurasi dari aturan asosiasi dapat menggunakan:

a. *Support*

Support menunjukkan seberapa dominan item ini muncul dalam suatu itemset. Jika N adalah total dari transaksi. Nilai support dari X adalah representasi dari jumlah X muncul dalam database dibagi N. (Bhatia, 2019). Perhitungan nilai support untuk X dapat dilihat pada persamaan 2. 9 berikut :

$$Support\ (X) = \frac{Jumlah\ kemunculan\ X}{Jumlah\ transaksi} = P(X) \quad (2.\ 9)$$

Sedangkan, perhitungan untuk nilai support X dan Y dapat dilihat pada persamaan 2. 10 berikut

$$Support(XY) = \frac{Jumlah\ kemunculan\ X\ dan\ Y\ bersamaan}{Jumlah\ transaksi} = P(X \cap Y) \quad (2.\ 10)$$

b. *Confidence*

Misalkan kita punya pasangan item $X \rightarrow Y$. Maka, *confidence* dapat dikatakan seberapa yakin kita bahwa Y akan muncul jika item X terjadi. Perhitungan nilai *confidence* dapat dilihat pada persamaan 2. 11 berikut:

$$Confidence(X \rightarrow Y) = \frac{Support\ (XY)}{Support\ (X)} = \frac{P(X \cap Y)}{P(X)} = P\ (Y|X) \quad (2.\ 11)$$

c. *Lift*

Lift merupakan rasio dari probabilitas kondisional Y ketika X diberi probabilitas yang tidak kondisional Y dalam dataset. Perhitungan nilai *lift* dalam aturan asosiasi dapat dilihat pada persamaan 2.12 berikut:

$$Lift = \frac{Confidence\ of\ (X \rightarrow Y)}{P(Y)} \quad (2.\ 12)$$

d. *Conviction*

Conviction merupakan salah satu perhitungan untuk melihat nilai akurasi minimum dari aturan asosiasi yang ada.

$$Conviction\ (X \rightarrow Y) = \frac{1-Support\ (Y)}{1-Confidence(X \rightarrow Y)} \quad (2.\ 13)$$

2.9. Algoritma Asosiasi

Algoritma apriori adalah salah satu metode dalam mencari aturan asosiasi. Apriori adalah algoritma seminal yang diusulkan oleh R. Agrawal dan R. Srikant pada tahun 1994 untuk menambang *frequent itemsets* untuk aturan asosiasi *Boolean* [AS94b] (Pei dkk., 2012).

Dalam apriori terdapat *apriori property* yaitu dimana semua sub set tidak kosong dari *frequent itemset* juga harus sering muncul. Hal ini dapat dilihat dari pengecekan tiap item set. Jika item set I tidak memenuhi batas *minimum support* ($P(I) < \min\ \_\sup$) maka I dianggap tidak sering muncul. Jika item A ditambahkan ke item set I, maka hasil dari item set tersebut ($I \cup A$) tidak akan muncul lebih sering daripada item I. Sehingga $I \cup A$ juga bukan *frequent* itemset karena $P\ (I \cup A) < \min\ \_sup$. Hal ini juga disebut *antimonotocity*, yaitu keadaan dimana jika satu set tidak bisa melewati tes maka semua super set dari set tersebut juga akan gagal melewati tes. (Pei dkk., 2012)

2.10. Sistem Informasi

Sebuah sistem informasi adalah seperangkat komponen yang saling terkait yang mengumpulkan, memanipulasi, menyimpan, dan menyebarkan

data dan informasi serta menyediakan mekanisme umpan balik untuk mencapai tujuan. Ini adalah umpan balik mekanisme yang membantu organisasi mencapai tujuan mereka. Bisnis dapat menggunakan sistem informasi untuk meningkatkan pendapatan dan mengurangi biaya (Stair & Reynolds, 2010).

Data terdiri dari fakta mentah, informasi adalah data yang diubah menjadi bentuk yang berarti. Proses mendefinisikan hubungan antar data membutuhkan pengetahuan. Pengetahuan adalah sebuah kesadaran dan pemahaman tentang sekumpulan informasi dan cara informasi dapat mendukung tugas tertentu. Untuk menjadi berharga, informasi harus memiliki beberapa karakteristik: akurat, lengkap, ekonomis untuk diproduksi, fleksibel, andal, relevan, mudah dipahami, tepat waktu, dapat diverifikasi, dapat diakses, dan aman. Nilai informasi terkait langsung bagaimana hal itu membantu orang dalam mencapai tujuan organisasi mereka (Stair & Reynolds, 2010).

2.11. Normalisasi

Satuan pengukuran yang digunakan dapat mempengaruhi analisis data. Untuk membantu menghindari ketergantungan pada pilihan unit pengukuran, data harus dinormalisasi atau distandarisasi. Normalisasi data digunakan untuk memberikan bobot yang sama dalam semua atribut (Pei dkk., 2012). Ada banyak metode dalam normalisasi, beberapa diantaranya yaitu:

a. Normalisasi *min-max*

Normalisasi *min-max* melakukan transformasi linier pada data asli. Misalkan $minA$ dan $maxA$ adalah nilai minimum dan maksimum dari suatu atribut A. Normalisasi min-max memetakan suatu nilai v_i dari A sampai v_0 dalam kisaran $[new\_min_A,\ new\_max_A]$ dengan menghitung

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \quad (2.\ 14)$$

Normalisasi *min-max* mempertahankan hubungan di antara nilai data asli. Ini akan menemukan kesalahan "di luar batas" jika kasus input di masa mendatang untuk normalisasi jatuh di luar rentang data asli untuk A (Pei dkk., 2012).

b. Normalisasi *Z-Score*

Dalam normalisasi nilai-z (atau normalisasi rata-rata nol), nilai untuk suatu atribut A, dinormalisasi berdasarkan $mean$ (rata-rata) dan standar deviasi dari nilai A. Nilai A, v_i, dari A dinormalisasi menjadi v_0 dengan menghitung,

$$v_i' = \frac{v_i - \overline{A}}{\sigma_A} \quad (2.\ 15)$$

di mana $\overline{A}$ dan A masing-masing adalah *mean* dan standar deviasi dari atribut. Metode normalisasi ini berguna ketika atribut minimum dan maksimum yang sebenarnya A tidak diketahui, atau ketika ada *outlier* yang mendominasi normalisasi *min-max* (Pei dkk., 2012).

c. Skala desimal

Normalisasi dengan skala desimal akan dinormalisasi dengan memindahkan titik desimal nilai atribut A. Jumlah titik desimal yang

dipindahkan tergantung pada absolut maksimum nilai A. Nilai A, v_i, dari A dinormalisasi menjadi v_0 dengan menghitung

$$vi' = \frac{vi}{10^j} \quad (2.\ 16)$$

dimana j adalah integer terkecil sehingga *max* ($|v_0$i|) < 1 (Pei dkk., 2012).

2.12. Penelitian Terkait

Pada tahun 2017, Gita Indah melakukan penelitian untuk mengevaluasi kinerja mahasiswa. Penelitian ini menggunakan dua metode *data mining*, yaitu asosiasi dan *clustering.* Penelitian ini menggunakan asosiasi untuk menemukan variabel yang berkaitan dengan masa studi dan *clustering* untuk mengelompokkan mahasiswa. Hasil dari penelitian ini yaitu dengan algoritma apriori dapat diperoleh pengetahuan hubungan jenis asal sekolah dan predikat lulusan dengan nilai *confidence* tertinggi 0,95. Gambaran karakteristik mahasiswa terbagi menjadi 6 *klaster*. Kedua teknik *data mining* memberi gambaran yang saling mendukung (Marthasari, 2017).

Penelitian selanjutnya dilakukan oleh Dwi Ardia dkk. Yang menggunakan algoritma *k-means* dalam mengevaluasi kualitas kinerja. Penelitian ini menggunakan 544 data dan 6 parameter terkait nilai *performance* karyawan. Hasil dari penelitian ini yaitu kinerja karyawan terbagi menjadi 5 klaster. Proses ini membutuhkan 10 kali iterasi (Ardia dkk., 2018).

Ashar Johar dkk. Pada tahun 2019 melakukan penelitian untuk mengimplementasikan algoritma *clustering* dalam sistem informasi

kepegawaian yang dibuat. Hasil penelitian ini yaitu sistem informasi dapat mengelompokkan 372 data pegawai menjadi 3 klaster sesuai masa pensiunnya. Parameter yang digunakan dalam *clustering* adalah usia dan eselon pegawai (Johar dkk., 2019).

Penelitian dalam melakukan *clustering* dengan *k-means* pada data *human resource* juga telah dilakukan pada tahun 2020. Penelitian ini menggunakan *clustering* pada platform spark dan menyimpulkan bahwa manajemen sumber daya manusia memang harus ditingkatkan lagi. Aplikasi manajemen ini harus memanfaatkan teknologi agar dapat menganalisis, merancang, mengatur dan melakukan kontrol terhadap sumber daya manusia (Zhao, 2020).

Pada 2021, Akhmad Upi Fitriyadi dkk. Melakukan penelitian untuk membandingkan kinerja algoritma *k-means* dan *k-medoids* dalam *clustering* penilaian kinerja karyawan. Dari penelitian yang dilakukan, didapatkan 4 klaster dan hasil algoritma *k-means* mempunyai akurasi 56%, presisi 25% dan *recall 60%.* Sedangkan algoritma *k-medoids* mempunyai tingkat akurasi 14%, presisi 25% dan *recall 25%.* Parameter yang digunakan selama proses penilaian kinerja yaitu strategi, *job desc,* pekerjaan, kehadiran, penampilan, aggresifitas, *problem solving* dan hasil kerja (Fitriyadi, 2021).