Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

PENDEKATAN DATA-DRIVEN: CLUSTERING UNTUK SEGMENTASI KARYAWAN YANG CENDERUNG MENGALAMI ATRISI

Fanindia Purnamasari[1], Umaya Ramadhani Putri Nasution[2]
1,2) Teknologi Informasi, Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Indonesia

Article Info

Article history:

Received: 30 Mei 2025
Revised: 04 Juni 2025
Accepted: 20 Juni 2025

ABSTRACT

Abstrak

Analisis sumber daya manusia dapat dilakukan melalui pendekatan berbasis data untuk mendukung pengambilan keputusan yang lebih efektif. Pengambilan keputusan dapat diperoleh dengan beberapa cara yaitu melakukan pengelompokan karyawan. Clustering, yaitu metode pengelompokan data berdasarkan kesamaan fitur tanpa label sebelumnya, merupakan salah satu teknik yang banyak digunakan. Penelitian ini bertujuan untuk menemukan pola dan mengklasifikasikan kelompok karyawan yang berpotensi mengalami atrisi berdasarkan karakteristik seperti performa kerja, masa kerja, dan data demografis. Penelitian ini melewati beberapa tahapan yaitu praproses data, identifikasi outlier, mengidentifikasi korelasi antar fitur, penerapan K-Means dan DBSCAN, serta melakukan evaluasi terhadap dua algoritma clustering tersebut. Davies-Bouldin dan Silhouette Score score untuk DBSCAN masing-masing sebesar 0.805843 dan 0.517547, sedangkan K-Means sebesar 2.535194 dan 0.114955. Clustering menunjukkan terdapat empat kelompok karyawan dengan karakteristik yang serupa yaitu pekerja senior dengan performa tinggi, karyawan dengan pengalaman baru, tenaga teknis berpengalaman, dan kelompok produktif dengan latar belakang pengalaman yang beragam. Hal ini diharapkan dapat membantu manajemen untuk membuat strategi pengembangan SDM yang lebih sesuai dengan tujuan mereka.

Kata Kunci: clustering, atrisi, karyawan, pengembangan SDM

Abstract

Human resource analysis can be done through a data-driven approach to support more effective decision making. Decision making can be obtained in several ways, namely clustering employees. Clustering, which is a method of grouping data based on similarity of features without prior labeling, is one of the widely used techniques. This research aims to find patterns and classify groups of potentially employees have attrition based on characteristics such as work performance, tenure, and demographic data. This research goes through several stages, namely data preprocessing, identifying outliers, identifying correlations between features, applying K-Means and DBSCAN, and evaluating the two clustering algorithms. Davies-Bouldin and Silhouette Score scores for DBSCAN are 0.805843 and 0.517547, respectively, while K-Means is 2.535194 and 0.114955. Clustering shows that there are four groups of employees with similar characteristics, namely senior workers with high performance, employees with new experience, experienced technical personnel, and productive groups with diverse experience backgrounds. This is expected to help management to create HR development strategies that are more in line with their goals.

Keywords: clustering, attrition, employee, Human Resource Development

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Corresponding Author:
E-mail : fanindia@usu.ac.id

1. PENDAHULUAN

Perkembangan era digital telah mendorong organisasi dan perusahaan untuk memanfaatkan teknologi serta sistem informasi dalam mengelola data sumber daya manusia (SDM). Sejumlah penelitian sebelumnya telah menunjukkan pentingnya integrasi teknologi informasi dalam manajemen SDM. Penerapan sistem digital dalam pengelolaan SDM mampu meningkatkan efisiensi pengambilan keputusan dan optimalisasi proses bisnis [1]. Hal ini menjadi isu penting dalam manajemen sumber daya manusia karena tingkat atrisi yang tinggi dapat membawa dampak negatif yang signifikan bagi perusahaan, seperti penurunan performa karyawan, gangguan operasional, bahkan proses rekrutmen yang berulang dilakukan [2]. Salah satu pemanfaatan data adalah analisis terhadap isu yang terjadi seperti atrisi karyawan, karyawan yang mengalami *turnover* dan penilaian terhadap kinerja karyawan [2], [3]. Visualisasi data seperti penyajian data dalam bentuk grafik, diagram, atau dasbor interaktif membantu memudahkan pemantauan performa pegawai serta mendukung pengambilan keputusan yang tepat sasaran[4]. Penggunaan data historis yang meliputi kepuasan kerja, usia, pendapatan, dan pengembangan karier berpengaruh dalam atrisi karyawan yang diteliti dan dapat memprediksi karyawan yang akan keluar dari pekerjaan. Selain itu faktor seperti gaji, usia, jarak tempuh dari rumah ke kantor, status pernikahan dan jenis kelamin juga mempengaruhi karyawan untuk meninggalkan pekerjaanya. Hal ini diteliti oleh [5] dengan membandingkan algoritma KNN dan Random Forest, model prediksi kemungkinan karyawan *turnover* sebelum keputusan dalam rekrutmen.

Dalam penelitian sebelumnya, atrisi telah diprediksi dengan menggunakan pendekatan yang berbeda. Algoritma Random Forest menunjukkan kinerja yang lebih unggul dalam memprediksi atrisi karyawan, dengan tingkat akurasi sebesar 85,12% menggunakan lima faktor utama yakni sosial, budaya, finansial, professional dan faktor

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

relasional[6]. *Logistic Regression* telah digunakan untuk mengidentifikasi karyawan yang berpotensi meninggalkan organisasi berdasarkan sejumlah atribut personal dan profesional [7]. Faktor lainnya seperti kompensasi, *work-life balance*, dan budaya organisasi telah diteliti, namun kurang pendekatan untuk menganalisis interaksi antar faktor yang dapat memengaruhi atrisi di berbagai industri dan organisasi. Pendekatan berbasis data juga digunakan untuk proses segmentasi yang membantu perusahaan mengidentifikasi prediktor utama dari *turnover* serta merancang intervensi yang lebih tepat guna [8]. Metode clustering merupakan salah satu teknik dalam *unsupervised learning* yang digunakan untuk mengelompokkan objek berdasarkan kemiripan karakteristik, dengan tujuan memaksimalkan homogenitas dalam satu kelompok dan meminimalkan kesamaan antar kelompok [9]. Sedangkan dari konteks manajemen SDM, clustering telah diterapkan untuk membedakan karyawan berdasarkan keahlian di bidang teknologi informasi [10] serta untuk mengukur dan mengevaluasi performa kerja. Berdasarkan penelitian sebelumnya, yang berbasis klasifikasi dan prediksi terhadap karyawan yang cenderung mengalami atrisi sehingga juga dapat mengidentifikasi korelasi antar variabel dengan kategori pengelompokan tertentu.

Penelitian ini membandingkan dua algoritma clustering yang umum digunakan yaitu K-Means dan DBSCAN. K-Means dikenal dengan efisiensinya dalam mengelompokkan data berdimensi rendah hingga sedang, namun kurang optimal untuk data yang tidak memiliki distribusi berbentuk bulat atau mengandung noise [11] Di sisi lain, DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) mampu membentuk klaster berdasarkan kepadatan data dan mengidentifikasi titik-titik yang tidak termasuk dalam klaster sebagai *noise* [12]. Perbandingan performa K-Means dan DBSCAN pada data berdimensi tinggi memberikan hasil berbeda tergantung pada karakteristik distribusi data[13], [14]. Melalui pendekatan ini, diharapkan sebagai referensi bagi perusahaan untuk mengidentifikasi kelompok karyawan dengan risiko tinggi mengalami atrisi, serta menyusun kebijakan retensi yang lebih terarah dan berbasis data.

2. METODE PENELITIAN

Penelitian ini melakukan beberapa tahapan antara lain data preprocessing, mencari korelasi antar variabel, pemilihan fitur, implementasi clustering dan evaluasi hasil

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

clustering. Data yang digunakan adalah data karyawan attrition yang diambil dari kaggle sejumlah 1470 data dan terdiri dari 35 fitur.

2.1 Data Preprocessing

Pada tahap ini, peneliti melakukan pembersihan data (data cleaning), peneliti membersihkan data dengan cara mengidentikasi data yang tidak konsisten, nilai yang hilang, dan *noise*. Selanjutnya, dilakukan tahap encoding pada data kategorikal untuk memudahkan proses pemrosesan data oleh model.

2.2 Identifikasi Outlier

Pada penelitian ini menggunakan metode rentang quartile untuk mengidentifikasi outlier [15]. IQR memberikan informasi mengenai sebaran data di sekitar nilai median. Semakin besar nilai IQR, maka semakin besar keragaman data. IQR digunakan untuk mendeteksi outlier yaitu nilai yang berada di luar rentang batas bawah dan atas[16]

FOR setiap kolom dalam data:
- Urutkan nilai-nilai dalam kolom tersebut
- Hitung kuartil pertama (Q1), yaitu persentil ke-25
- Hitung kuartil ketiga (Q3), yaitu persentil ke-75
- Hitung IQR sebagai selisih antara Q3 dan Q1
- Hitung nilai minimum sebagai Q1 dikurangi 1.5 dikali IQR
- Hitung nilai maksimum sebagai Q3 ditambah 1.5 dikali IQR

END

2.3 Korelasi spearman

Korelasi antara fungsi distribusi kumulatif (CDF) dari dua variabel kontinu dengan peringkat Spearman yang termasuk dalam korelasi tingkat. Umumnya korelasi ini digunakan pada data yang bersifat ordinal [17].

2.4 Pemilihan Fitur

Univariate Feature Selection (SelectKBest) merupakan metode filter dalam proses seleksi fitur. SelectKBest[18] digunakan untuk meningkatkan akurasi atau untuk meningkatkan kinerja pada dataset yang memiliki dimensi tinggi. SelectKBest

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

merupakan bagian *Univariate Feature selection* yang dapat memilih fitur terbaik berdasarkan uji statistik univariat atau uji ANOVA.

2.5 Implementasi K-Means

Pada tahap ini, akan dilakukan inisialisasi centroid (titik tengah) awal yang dipilih secara acak dari sekumpulan data. Selanjutnya setiap data point yang terbentuk akan dihitung jarak ke semua centroid menggunakan Euclidean distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Keterangan:

d(x, y) : Jarak antara titik *x* dan titik *y*

x_i, y_i: Nilai atribut ke-*i* dari titik *x* dan *y*

n: Jumlah dimensi (fitur) dari data.

Lakukan secara berulang inisialisasi centroid sebanyak 10 kali dan memilih hasil terbaik berdasarkan *inertia.* **Update** langkah untuk setiap centroid yang telah dipernarui menjadi rata-rata (mean) titik yang ditugaskan kepadanya hingga mencapai konvergensi dalam hal ini max iterasi sebesar 300.

2.6 Implementasi DBSCAN

DBSCAN mengelompokkan data berdasarkan kepadatan, melihat berapa banyak poin minimum yang ada dalam jangkauan tertentu.. Parameter seperti epsilon dan MinPts (minimum number of points) ditentukan berdasarkan kriteria minimum sebuah kelompok dapat dijadikan titik pusat. untuk mendapatkan hasil clustering yang akurat. Variasi nilai epsilon yang digunakan pada penelitian ini berdasarkan visualisasi dari clustering dengan melakukan percobaan dari beberapa nilai yaitu 0.5 hingga 5.0 dengan kenaikan 0.5. Berdasarkan pengujian diperoleh bahwa hasil terbaik terjadi pada ε = 0.5, yang menghasilkan klaster dengan komposisi dan pemisahan yang paling optimal. Selanjutnya menghitung nilai radius epsilon setiap tetangga. Jika jumlah tetangga ≥ min_samples ditandai sebagai core point, sedangkan titik dalam neighborhood core dianggap sebagai border point dan titik lainnya dianggap sebagai noise. Lakukan evaluasi

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

dengan menghitung metrik per konfigurasi, lalu memilih ε yang menghasilkan klaster yang tepat. Setelah ditemukan epsilon terbaik, berikan label dan skor akhir dari cluster yang dihasilkan.

2.7 Evaluasi Clustering

Silhouette Score: Mengukur seberapa mirip objek dengan cluster terbentuk dibandingkan dengan cluster lain. Nilai tersebut berkisar antara -1 dan 1, di mana nilai lebih tinggi menunjukkan klasterisasi yang lebih baik [19]

$$s(i) = (b(i)-a(i))/(\max(a(i),b(i))). \quad (2)$$

Keterangan:

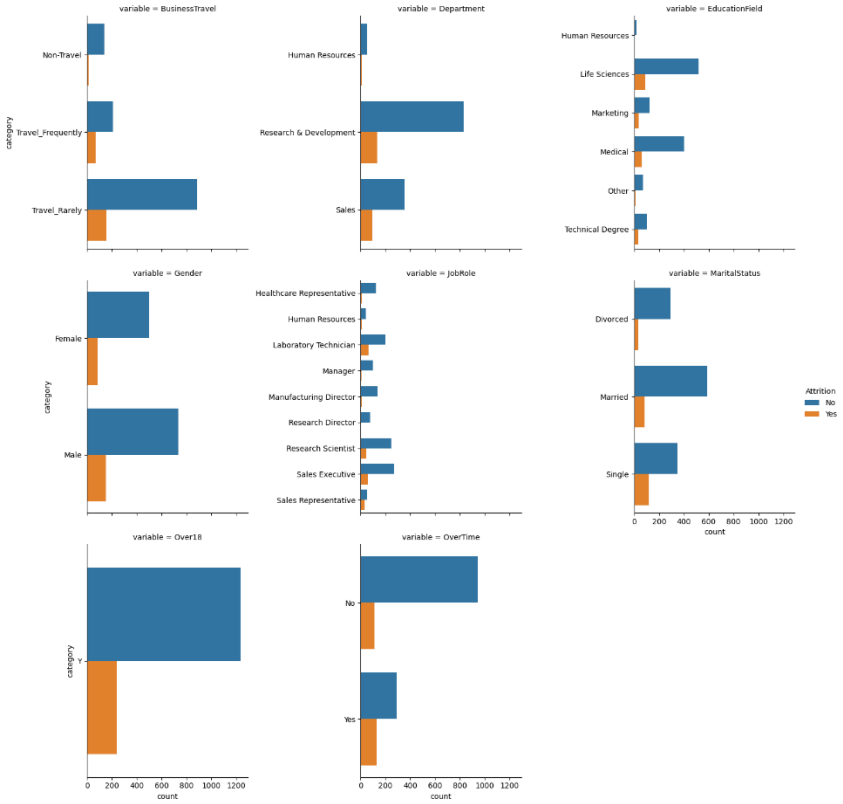
a(i): jarak rata-rata antara data point i dan semua titik lain dalam cluster yang sama
b(i): jarak rata-rata antara data point I dan semua titik dalam cluster terdekat tetangga

Davies-Bouldin Score adalah metrik evaluasi yang digunakan untuk menilai kualitas hasil clustering, seberapa baik cluster yang dipisahkan satu sama lain. Semakin rendah nilainya menunjukkan hasil clustering yang baik [20]. Davies-Bouldin Score menghitung rata-rata dari nilai maksium rasio similaritas antar cluster untuk setiap cluster

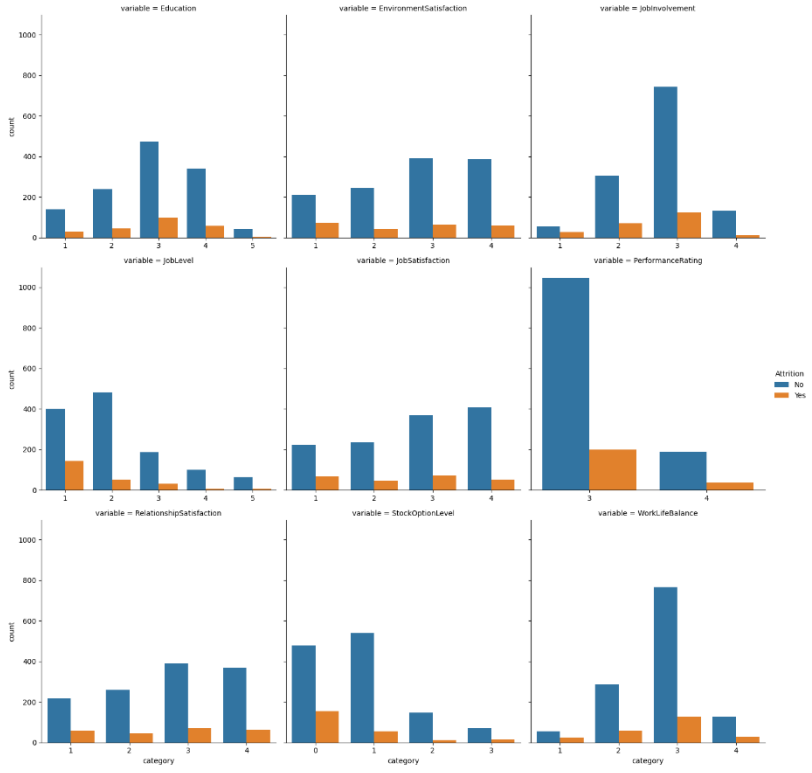
3. HASIL DAN PEMBAHASAN

Pada bagian ini akan dijelaskan hasil penerapan metode yang telah dijelaskan pada bagian sebelumnuya. Visualisasi data menjadi langkah awal yang penting dalam memahami distribusi variabel serta hubungan antar fitur dalam dataset karyawan. Dengan bantuan teknik eksplorasi data, pola-pola awal dapat diidentifikasi sebelum diterapkan metode clustering untuk pengelompokan lebih lanjut. Clustering memungkinkan pembagian karyawan ke dalam beberapa segmen berdasarkan parameter tertentu, seperti produktivitas, tingkat kepuasan, atau pola kehadiran.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno



Gambar 4(a) Persebaran Data per Kategori Variabel

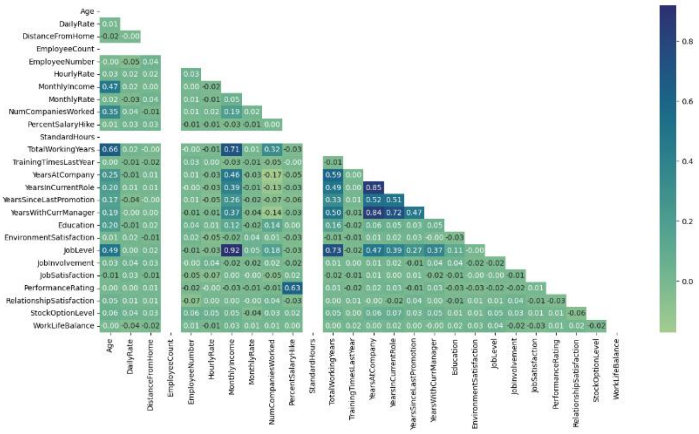


Gambar 4(b) Persebaran Data per Kategori Variabel

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Gambar 4(a) dan (b) merupakan visualisasi data sebaran. Data tiap variabel yang menginterpretasikan attrition atau tidak. Visualisasi data menunjukkan beberapa hasil berikut:

1. Secara keseluruhan, terdapa sejumlah 237 karyawan yang mengundurkan diri dari total 1.470 karyawan, dimana karyawan laki-laki memiliki tingkat pengunduran diri *(attrition)* yang lebih tinggi dibandingkan perempuan.
2. Karyawan yang mengalami tingkat pengunduran diri tertinggi berada pada posisi dan departement *Laboratory Technician*, *Research and Development* dan bidang pendidikan *Life Sciences*, perjalanan dinas juga cenderung mempengaruhi tingkat pengunduran diri. Karyawan yang jarang melakukan perjalanan dinas *(Business Travel: Rarely)* lebih cenderung mengundurkan diri dibandingkan mereka yang sering bepergian.



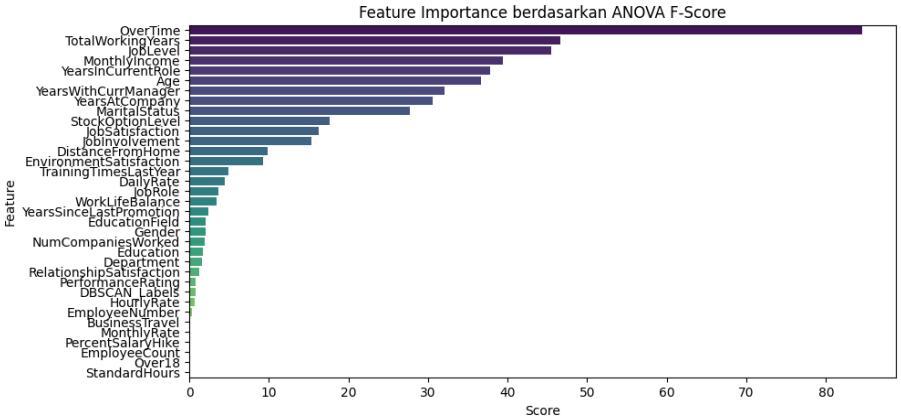
Gambar 5. Korelasi antar Variabel

Gambar 5 menunjukkan *heatmap* korelasi yang merupakan representasi visual dari hubungan antar variabel dalam suatu dataset. Warna tersebut menunjukkan tingkat korelasi, warna yang lebih gelap atau lebih terang menunjukkan seberapa kuat hubungan antar variabel antara lain:

1. TotalWorkingYears vs Age (0.66) menunjukkan lama bekerja memiliki korelasi dengan usia
2. TotalWorkingYears vs MonthlyIncome (0.71) menunjukkan pengalaman kerja yang lebih lama berkorelasi dengan pendapatan bulanan yang lebih tinggi.
3. YearsAtCompany vs YearsInCurrentRole (0.85) menunjukkan lama bekerja di perusahaan berkorelasi erat dengan lama berada di *Role* saat ini.

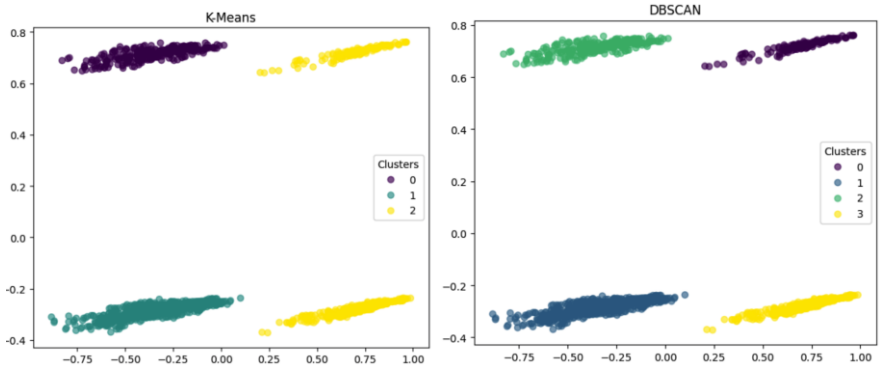
Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

4. YearsAtCompany vs YearsWithCurrManager (0.72) menunjukkan semakin lama seseorang bekerja di perusahaan, semakin lama juga mereka bekerja dengan manajer yang sama.
5. Beberapa variabel seperti DistanceFromHome, PerformanceRating, JobSatisfaction, dan WorkLifeBalance memiliki korelasi mendekati nol dengan variabel lain, yang berarti variabel tersebut tidak memiliki hubungan yang kuat satu sama lain.
6. Education memiliki korelasi yang sangat kecil dengan MonthlyIncome dan JobLevel, yang menunjukkan bahwa tingkat pendidikan tidak selalu menentukan posisi pekerjaan atau gaji.



Gambar 6. Pemilihan Fitur

Gambar 6 menunjukkan tingkat kepentingan dari berbagai variabel dalam dataset berdasarkan ANOVA F-Score. Dalam clustering, pemilihan fitur sangat penting karena mempengaruhi hasil segmentasi data. Memilih fitur yang tepat dapat meningkatkan kualitas pemisahan antar kelompok. OverTim**e** memiliki F-Score tertinggi menunjukkan bahwa karyawan yang bekerja lembur memiliki hubungan yang sangat kuat dengan variabel target.



Gambar 7. Klaster yang dihasilkan

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Pada Gambar 7, algoritma K-Means menunjukkan hasil clustering data menjadi tiga klaster yang jelas, masing-masing diwakili oleh warna ungu, hijau, dan kuning. Klaster ungu mendominasi area kiri atas dan kiri bawah grafik, sedangkan klaster kuning tersebar di bagian kanan, baik di atas maupun di bawah. Namun, algoritma tetap menganggap data ini sebagai satu kelompok secara spasial karena jarak yang dekat di antara keduanya. Terdapat korelasi antara fitur yang diproyeksikan dalam dua dimensi melalui PCA yang ditunjukkan oleh distribusi data untuk setiap klaster. Tidak ada tumpang tindih di antara klaster, yang menunjukkan bahwa metode K-Means cukup efektif dalam mengelompokkan data ini. Sedangkan DBSCAN Berdasarkan visualisasi hasil clustering DBSCAN, dapat diinterpretasikan bahwa algoritma ini membagi data menjadi empat klaster utama, masing-masing diberi label 0 (ungu), 1 (biru), 2 (hijau), dan 3 (kuning). Distribusi klaster tampak cukup terpisah yang menunjukkan bahwa parameter eps dan min_samples yang digunakan cukup ideal.

Tabel 1 Representasi Clustering

Klaster yang dihasilkan	Keterangan
Cluster 0	Klaster dengan karakteristik khusus, misalnya kelompok senior atau pekerja lama dengan performa tinggi.
Cluster 1	Klaster merupakan karyawan baru atau junior dengan pengalaman rendah dan performa kerja yang masih dalam tahap pengembangan.
Cluster 2	Klaster dengan karakteristik karyawan yang sudah cukup berpengalaman namun berada pada peran pekerjaan yang bersifat teknis
Cluster 3	Klaster yang mewakili kelompok produktif dengan performa baik dan pengalaman kerja lebih bervariasi.

Selanjutnya untuk mengevaluasi performa model K-Means dalam mengelompokkan data, dilakukan evaluasi performa dari metode clustering K-Means dan DBSCAN, digunakan dua metrik evaluasi utama, yaitu Silhouette Score dan Davies-Bouldin Index (DBI)**.** Silhouette Score digunakan untuk mengukur keberadaan data pada klaster yang tepat, dengan rentang nilai antara -1 hingga 1. Dalam hal ini semakin tinggi nilai silhouette, semakin baik pemisahan antar klaster. Sedangkan Davies-Bouldin Index digunakan untuk menilai seberapa besar jarak antar klaster relatif terhadap sebaran dalam klaster; semakin rendah nilai DBI, maka semakin baik kualitas klasterisasi.

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

Tabel 2. Evaluasi Cluster

	K-Means	DBSCAN
Silhouette Score	0.114955	0.517547
Davies-Bouldin Index (DBI)	2.535194	0.805843

Tabel 2, menunjukkan hasil evaluasi algoritma DBSCAN yang mengindikasikan bahwa performa klasterisasi yang lebih unggul dibandingkan K-Means. Hal ini dapat dilihat dari nilai Silhouette Score DBSCAN sebesar 0.517547, yang mengindikasikan bahwa objek-objek dalam masing-masing klaster memiliki pemisahan yang baik terhadap klaster lainnya. Nilai ini tergolong cukup baik dalam konteks klasterisasi, karena mendekati 1, yang berarti mayoritas titik data dikelompokkan ke dalam klaster yang sesuai. Sedangkan nilai Davies-Bouldin Index (DBI) yang dihasilkan DBSCAN sebesar 0.805843 lebih rendah dibandingkan K-Means. Nilai DBI yang rendah mengindikasikan bahwa klaster yang terbentuk oleh DBSCAN memiliki sebaran internal yang dengan jarak antar klaster yang terpisah dengan baik sehingga menghasilkan pemisahan klaster yang optimal. Hal ini dapat dibandingkan dengan Silhouette Score pada K-Means yang jauh lebih rendah, yaitu hanya sebesar 0.114955**,** dan nilai DBI yang tinggi, yaitu 2.535194**.** Nilai yang dihasilkan menunjukkan bahwa klaster yang dihasilkan K-Means cenderung kurang kecil **.**

4. SIMPULAN

Pendekatan data-driven dengan menggunakan teknik clustering dapat mengelompokkan karyawan yang akan mengundurkan diri (attrition) berdasarkan karakteristik dan pola perilaku karyawan. DBSCAN dapat membentuk cluster berdasarkan kepadatan data, sedangkan K-means memerlukan penentukan jumlah cluster di awal. Namun performa DBSCAN sangat bergantung pada pemilihan epsilon (radius tetangga) yang dapat mengakibatkan densitas antar cluster bervariasi. Berdasarkan cluster yang terbentuk sebanyak empat cluster menunjukkan DBSCAN lebih adaptif terhadap variasi pola perilaku karyawan Pengelompokan yang terbentuk diharapkan dapat membentuk strategi dalam manajemen SDM yaitu untuk memaksimalkan retensi, meingkatkan produktivitas dalam berinovasi dan pengembangan talenta dengan memberikan pelatihan bagi karyawan

Djtechno : Jurnal Teknologi Informasi
Vol. 6, No. 2, Agustus 2025
E-ISSN: 2745-3758, P-ISSN : 2776-8546 DOI: 10.46576/djtechno

REFERENSI

[1] Linda Daniati Melinda, B. Harto, Sulistianingsih, Hery Syaerul Homan, and Dwi Puryati, "Integrasi Teknologi Informasi dalam Manajemen Sumber Daya Manusia: Sebuah Studi Kualitatif tentang Dampaknya pada Kinerja Keuangan Perusahaan," *ATRABIS Jurnal Administrasi Bisnis (e-Journal)*, vol. 9, no. 2, pp. 321–335, Dec. 2023, doi: 10.38204/atrabis.v9i2.1820.

[2] A. R. B. J. Jamroni, Wahyu Hadikristanto, and Muhamad Fatchan, "Analisis Faktor dan Prediksi Atrisi untuk Optimalisasi Retensi Karyawan Menggunakan Machine Learning," *bit-Tech*, vol. 7, no. 3, pp. 1057–1067, Apr. 2025, doi: 10.32877/bt.v7i3.2301.

[3] S. F. Sari and K. M. Lhaksmana, "Employee Attrition Prediction Using Feature Selection with Information Gain and Random Forest Classification," *Journal of Computer System and Informatics (JoSYC)*, vol. 3, no. 4, pp. 410–419, Sep. 2022, doi: 10.47065/josyc.v3i4.2099.

[4] S. Majumder, "Visualizing Insights to Empower HR Decision-Making: A Data-Driven Approach," 2024, pp. 127–138. doi: 10.1007/978-981-97-6992-6_10.

[5] M. Atef, D. S. Elzanfaly, and S. Ouf, "Early Prediction of Employee Turnover Using Machine Learning Algorithms 135 Original Scientific Paper."

[6] M. Pratt, M. Boudhane, and S. Cakula, "Employee attrition estimation using random forest algorithm," *Baltic Journal of Modern Computing*, vol. 9, no. 1, pp. 49–66, 2021, doi: 10.22364/BJMC.2021.9.1.04.

[7] G. Pratibha and N. P. Hegde, "HR Analytics : Early Prediction of Employee Attrition using KPCA and Adaptive K-means based Logistic Regression," in *2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS)*, IEEE, May 2022, pp. 11–16. doi: 10.1109/ICPS55917.2022.00010.

[8] M. R. Shafie, H. Khosravi, S. Farhadpour, S. Das, and I. Ahmed, "A cluster-based human resources analytics for predicting employee turnover using optimized Artificial Neural Networks and data augmentation," *Decision Analytics Journal*, vol. 11, p. 100461, Jun. 2024, doi: 10.1016/j.dajour.2024.100461.

[9] A. Serra and R. Tagliaferri, "Unsupervised Learning: Clustering," in *Encyclopedia of Bioinformatics and Computational Biology*, Elsevier, 2019, pp. 350–357. doi: 10.1016/B978-0-12-809633-8.20487-1.

[10] D. Zakiyah, N. Merlina, and N. A. Mayangky, "Penerapan Algoritma K-Means Clustering Untuk Mengetahui Kemampuan Karyawan IT," *Computer Science (CO-SCIENCE)*, vol. 2, no. 1, pp. 59–67, Jan. 2022, doi: 10.31294/coscience.v2i1.623.

[11] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation," *Electronics (Basel)*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/electronics9081295.

[12] Y. Zhang, "Application of nonlinear clustering optimization algorithm in web data mining of cloud computing," *Nonlinear Engineering*, vol. 12, no. 1, Jan. 2023, doi: 10.1515/nleng-2022-0239.

[13] L. Febrianti and N. Rahmadi, "Pengaruh Reward dan Motivasi Kerja terhadap Kinerja Karyawan pada PT Aquafarm Nusantara," *Reslaj: Religion Education Social Laa Roiba Journal*, vol. 5, no. 6, p. 3699, 2023, doi: 10.47476/reslaj.v5i6.4518.

[14] F. R. Ferdiansyah, R. W. Nugraha, R. Sofian, H. Purwanto, D. Saepudin, and E. Andriansyah, "Implementation of K-Means and DBSCAN algorithms: A Bibliometric Review," 2024, pp. 192–202. doi: 10.2991/978-94-6463-618-5_21.

[15] R. Efendi, A. Junaidi, and A. M. Rizki, "Penentuan Pusat Klaster Secara Otomatis Pada Algoritma Density Peaks Clustering Berbasis Metode Inter Quartile Range," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 3, Aug. 2024, doi: 10.23960/jitet.v12i3.4997.

[16] Ch. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, p. 100164, Mar. 2023, doi: 10.1016/j.dajour.2023.100164.

[17] P. Warner, "Quantifying association in ordinal data," *Journal of Family Planning and Reproductive Health Care*, vol. 36, no. 2, pp. 83–85, Apr. 2010, doi: 10.1783/147118910791069187.

[18] Md. Raihan-Al-Masud and M. R. H. Mondal, "Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms," *PLoS One*, vol. 15, no. 2, p. e0228422, Feb. 2020, doi: 10.1371/journal.pone.0228422.

[19] Y. Januzaj, E. Beqiri, and A. Luma, "Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 19, no. 04, pp. 174–182, Apr. 2023, doi: 10.3991/ijoe.v19i04.37059.

[20] D. A. Tarigan, "Optimization of the K-Means Clustering Algorithm Using Davies Bouldin Index in Iris Data Classification," *Media Online)*, vol. 4, no. 1, pp. 545–552, 2023, doi: 10.30865/klik.v4i1.964.