Progresif: Jurnal Ilmiah Komputer
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0877
p-ISSN: 0216-3284

Analisis *Employee Satisfaction* Menggunakan Teknik *Clustering* Dan *Classification Machine Learning*

I Ketut Adi Wirayasa[1*], Handri Santoso[2]
Program MagisterTeknik Informasi, Universitas Pradita
Scientia Business Park Tower I, Jl. Boulevard Gading Serpong Blok O/1,
Summarecon - Serpong, Tangerang, Indonesia
*e-mail *Corresponding Author: adi.wirayasa@gmail.com*

Abstrak

Kepuasan kerja pekerja sangat berhubungan dengan pekerjaan maupun kondisi dirinya ditempat kerja. Tingkat kepuasan kerja pekerja dapat di analisis dan menjadi bahan evaluasi perusahaan dalam menjalankan bisnis untuk mencapai target yang diinginkan. Kombinasi teknik *clustering* dan *classification* merupakan algoritma *machine learning* yang dapat membantu bagian Sumber Daya Manusia dalam menganalisis dan prediksi tingkat kepuasan kerja pekerja di perusahaan. Teknik *clustering* yang digunakan dalam penelitian ini adalah K-Means dan teknik *classification* menggunakan algoritma *classificafier* dari *library Pycaret*. Hasil analisis dari penggunaan teknik *clustering* dan *classification* dari ke-5 model *classifier* yang dipilih, 3 model yaitu LightGBM, Catboost dan XGBoost menunjukkan performa yang konsiten dan menghasilkan tingkat *accuracy* prediksi diatas 98% dengan jumlah cluster ideal 2, n-*component* 27, waktu proses rata-rata setiap model kurang dari 2 menit setiap tahapan proses dan menggunakan *K-means clustering.*
Kata kunci: *Kepuasan pekerja; Klaster; Klasifikasi; Pembelajaran mesin*

Abstract

Job satisfaction of workers is closely related to their work and conditions at work. The level of job satisfaction of workers can be analyzed and become an evaluation material for companies in running a business to achieve the desired target. The combination of clustering and classification techniques is a machine learning algorithm that can assist the Human Resources department with analyzing and predicting the level of job satisfaction of workers in the company. The clustering technique used in this research is K-Means in the classification technique using a binary classification algorithm from the Pycaret library. The results analysis of the clustering and classification techniques from the five selected classifier models, three models namely LightGBM, Catboost, and XGBoost shown consistent performance and the prediction accuracy levels above 98% with the ideal number of clusters 2, n-components 27, the average of processing time each model is less than 2 minutes each stage process and using K-means clustering.
Keywords: *Employee Satisfaction; Clustering; Classification; Machine Learning*

1. Pendahuluan

Kepuasan pekerja (*employee satisfaction*) merupakan ukuran dari tingkat kepuasan pekerja sesuai jenis pekerjaan mereka yang berkaitan dengan sifat dari tugas pekerjaannya, hasil kerja yang dicapai, bentuk pengawasan yang diperoleh maupun perasaan suka terhadap pekerjaan yang ditekuninya[1, 2]. Menurut Wexley and Yukl, yang dikutip oleh [3] menyatakan bahwa kepuasan kerja sebagai “the way an employee feels about his or her job”, adalah cerminan perasaan pekerja terhadap dirinya atau pekerjaanya. Terdapat beberapa hal yang dapat menyokong atau tidak menyokong kondisi dirinya yang berkaitan dengan pekerjaan dan perasaan yang berhubungan dengan dirinya seperti umur, kondisi kesehatan, kemampuan dan pendidikan serta kesempatan untuk pengembangan karier, dan hubungan dengan pekerja lain[3, 4].

Kepuasan kerja (*job satisfaction*) tidak hanya mengarah pada peningkatan kinerja individu tetapi juga mengarah pada peningkatan kinerja departemen dan organisasi. Penurunan tingkat kepuasan kerja pekerja dapat disebabkan oleh beberapa faktor seperti mendapatkan upah yang rendah, pertumbuhan karier terbatas, kurangnya minat, manajemen yang buruk, atasan yang tidak mendukung, kurangnya pekerjaan yang berarti, rendahnya insentif untuk pekerjaan yang bermakna dan keseimbangan antara pekerjaan dan kehidupan pekerja[5, 6].

Faktor kepuasan kerja pekerja dapat berasal dari dalam maupun luar seperti: (1). Kebijakan kompensasi dan tunjangan yang diterima, (2). Keamanan dan kondisi kerja yang meliputi rasa aman, nyaman, dan motivasi kerja, (3). Hubungan dengan otoritas atasan, (4). Promosi dan pengembangan karir, (5). Gaya Kepemimpinan, (6). Kerja kelompok atau tim, (7). Hal yang bersifat pribadi seperti kepribadian, harapan, usia, pendidikan, dan perbedaan jenis kelamin, (8). Hal lainnya seperti apakah anggota kelompok merasa seperti bagian dari keluarga, dorongan dan umpan balik, serta akses dan perlengkapan kerja[1, 7, 8]. Ketidakpuasan mengarahkan perilaku untuk meninggalkan organisasi, termasuk mencari posisi baru serta pengunduran diri, dan berhentinya pekerja secara kolektif yang dapat merugikan organisasi.

Algoritma *unsupervised* adalah algoritma *Machine Learning* (ML) yang digunakan untuk kumpulan data yang tidak berlabel, yaitu yang tidak memiliki variabel keluaran (*output*). Algoritma *unsupervised* memfasilitasi analisis kumpulan data, sehingga membantu menghasilkan informasi dari data yang tidak berlabel. Kemajuan terbaru dalam pembelajaran hierarkis, algoritma pengelompokan, analisis faktor, model laten, dan deteksi *outlier*, telah membantu secara signifikan dalam teknik *unsupervised machine learning*[9, 10].

Untuk menganalisis faktor kepuasan kerja, teknik *clustering* dapat digunakan. Teknik *clustering* digunakan untuk data dengan jumlah besar dan kompleks[11]. Penelitian ini fokus pada analisis antara kepuasan kerja dan faktor-faktor penentunya, dengan mengacu pada pertanyaan: (1). Apakah algoritma *clustering* dapat menangani data numerik dan non numerik untuk data dalam jumlah besar (*big data*)? (2). Bagaimana hasil algoritma untuk pengolahan *big data* dengan teknik *clustering* dan *classification*?

Penelitian ini juga bertujuan menganalisis dua algoritma *clustering*: algoritma *K-Means* dan *Hierarchical Clustering-Agglomerative* serta membandingkan kinerja kedua algoritma ini dalam membangun klaster kelas yang sesuai serta mereduksi variabel yang terdapat dalam *dataset* menggunakan teknik *Principal Component Analysis* (PCA), sebelum dibuat model *classification* menggunakan algoritma *classifier* yang terdapat pada *library Pycaret* untuk melihat *accuracy* dari model yang dibuat dengan menggunakan komparasi beberapa algoritma yang ada.

2. Tinjauan Pustaka

2.1 Studi Literatur

Beberapa penelitian dengan memanfaatkan metode ML dan *deep learning* untuk pengolahan data dalam jumlah besar, digunakan untuk memudahkan di dalam mengambil keputusan dan juga memberikan analisis lebih mendalam dalam proses klasifikasi, asosiasi, klustering dan prediksi[1, 10, 12]. Metodologi *data mining* dapat dimanfaatkan untuk membantu para peneliti dan praktisi untuk memahami hubungan antara kinerja perusahaan dan faktor-faktor dapat mendorong penilaian perusahaan seperti komentar individu pekerja tentang kepuasan kerja[13, 14].

Penelitian dilakukan untuk melihat tingkat kepuasan kerja pekerja dalam hal sikap mereka terhadap berbagai aspek pada situasi pekerjaan dan meneliti fitur utama apa yang berkontribusi dalam memprediksi kepuasan kerja secara akurat, dengan menggunakan perbandingan beberapa algoritma ML untuk melihat algoritma yang berkinerja terbaik dalam memprediksi kepuasan kerja[1].

Peneliti lain menganalisis metode ML menggunakan *linear regression*, *random forest* dan *extreme gradient boosting*, dengan eksperimental untuk mendapatkan solusi yang terbaik dalam memprediksi tingkat kepuasan pekerja[15]. Melalui penelitian yang dilakukan oleh[4], peneliti menghasilkan faktor kepuasan kerja yang dapat dipertimbangkan seperti: 1). Menentukan jumlah topik yang tepat dan memvalidasi kualitasnya melalui uji reliabilitas dan uji validitas eksternal. 2). Faktor kepuasan kerja yang kami berikan dan hasil dari beragam analisis yang kami jalankan berdasarkan faktor-faktor akan memungkinkan para manajer departemen SDM untuk merencanakan, merancang, dan mengimplementasikan aktivitas terkait SDM. Masing-masing kegiatan tersebut diharapkan dapat mengarahkan pekerja untuk memiliki

kepuasan kerja yang lebih baik dari sebelumnya, dan mereka pada akhirnya akan berubah perusahaan mereka lebih kompetitif dan menemukan lima faktor kepuasan kerja baru yang tidak dipertimbangkan dalam literatur[4].

Penelitian menggunakan analisis klaster untuk mengelompokkan pekerja ke dalam kelompok-kelompok sesuai dengan kinerjanya. Algoritma *decision tree* digunakan untuk membuat keputusan yang berarti bagi pekerja. Berdasarkan hasil kinerja pekerja dimungkinkan untuk mengambil keputusan apakah pelatihan lanjutan, pengayaan bakat atau kualifikasi lebih lanjut diperlukan atau tidak. Aplikasi ini juga dapat membantu bagian administrasi untuk meningkatkan kualitas organisasi[16].

2.2 Clustering

Clustering adalah teknik klasifikasi dengan tanpa pengawasan (*unsupervised*), di mana satu set pola atau sampel dikelompokkan ke dalam klaster sedemikian rupa menjadi klaster yang sama dan serupa serta terbagi dalam klaster yang berbeda-berbeda, namun memiliki anggota yang mirip atau sama dengan yang sampel dari kelompok lain[17, 18]. *Clustering* adalah prosedur adaptif di mana objek dikelompokkan bersama, berdasarkan prinsip memaksimalkan kesamaan intra-kelas dan meminimalkan kesamaan antar-kelas[11, 19]. Berbagai algoritma *clustering* telah dikembangkan yang menghasilkan kinerja yang baik pada kumpulan data untuk pembentukan klaster.

Pada algoritma *K-Means*, *clustering* bertujuan untuk mencari kelompok data (data tidak berlabel) dan di-*cluster* berdasarkan kemiripan fitur dengan menggunakan rumus jarak *euclidian*. Analisis *clustering* bermanfaat untuk mengeksplorasi data peubah ganda, mereduksi data, dan stratifikasi data sampling serta dapat memprediksi keadaan obyek[20, 21].

2.3 Pycaret

Library Pycaret dapat digunakan dengan instruksi *low code* ini memudahkan dalam penggunaan *coding* dan bersifat otomatis serta membantu mempercepat proses dari eksperimen yang dilakukan dan lebih efisien[22]. Prediksi menggunakan model *Pycaret* terhadap data yang telah di reduksi melalui proses *clustering* dan dibagi menjadi data *training* dan data *testing* akan menghasilkan performa dari model *classifier* mulai dari tahap pembentukan model, *tuning-parameter* dan *ensembling model*. Hasil *ensembling* akan memperlihatkan apakah model yang dipilih mengalami kenaikan performa dari sisi tingkat akurasi[23].

2.4 Principal Component Analysis (PCA)

Ketika *dataset* memiliki banyak fitur, model algoritma akan menjadi lebih kompleks. PCA adalah metode yang ideal untuk menangani kumpulan data dan mengurangi dimensi data dari kumpulan data yang kompleks[24]. PCA adalah cara untuk mengurangi jumlah variabel dan disaat bersamaan mempertahankan sebagian besar informasi yang penting. Proses Ini mengubah sejumlah variabel yang mungkin berkorelasi menjadi sejumlah kecil variabel yang tidak berkorelasi, yang dikenal sebagai komponen utama[25, 26]. PCA merupakan kombinasi linier yang diberi bobot pada variabel asli oleh variansnya (nilai *Eigen*) dalam dimensi ortogonal tertentu. Tujuan utama PCA adalah untuk menyederhanakan fitur model yang ada menjadi komponen yang lebih sedikit untuk membantu memvisualisasikan pola dalam data dan untuk membantu model berjalan lebih cepat. Menggunakan PCA juga mengurangi kemungkinan *overfitting* model dengan menghilangkan fitur dengan korelasi tinggi[25, 27].

3. Metodologi

Penelitian menggunakan *dataset Kaggle* – IBM *Analytics*, bahasa pemrograman *Python* dan memanfaatkan *library SKLearn* dan *Pycaret*. *Dataset* terdiri dari 1470 sampel dan 35 variabel, dan terbagi menjadi data numerik dan non-numerik dengan struktur data seperti ditunjukkan pada Gambar 1. *Dataset* akan dibagi menjadi data pelatihan dan testing dengan rasio 70 persen dan 30 persen.

	Features	Unique Number	Values
0	Age	43	[41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35, 2...
1	Attrition	2	[Yes, No]
2	BusinessTravel	3	[Travel_Rarely, Travel_Frequently, Non-Travel]
3	DailyRate	886	[1102, 279, 1373, 1392, 591, 1005, 1324, 1358,...
4	Department	3	[Sales, Research & Development, Human Resources]
5	DistanceFromHome	29	[1, 8, 2, 3, 24, 23, 27, 16, 15, 26, 19, 21, 5...
6	Education	5	[2, 1, 4, 3, 5]
7	EducationField	6	[Life Sciences, Other, Medical, Marketing, Tec...
8	EmployeeCount	1	[1]
9	EmployeeNumber	1470	[1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 14, 15, 16,...
10	EnvironmentSatisfaction	4	[2, 3, 4, 1]
11	Gender	2	[Female, Male]
12	HourlyRate	71	[94, 61, 92, 56, 40, 79, 81, 67, 44, 84, 49, 3...
13	JobInvolvement	4	[3, 2, 4, 1]
14	JobLevel	5	[2, 1, 3, 4, 5]
15	JobRole	9	[Sales Executive, Research Scientist, Laborato...

Gambar 1. Struktur Dataset

Prosedur *K-means clustering* dalam penelitian ini, dijelaskan sebagai berikut:

1. Tentukan jumlah *cluster* untuk di kelompokkan data nya kedalam *cluster* sebagai nilai input.
2. Tentukan nilai *centroids*.
3. Inisialisasi k- *cluster* pertama, k-*instance* pertama dari k-elemen.
4. Hitung mean dari masing-masing *cluster* yang terbentuk pada dataset.
5. Tetapkan *K-means* dari *dataset* untuk satu *cluster* awal.
6. Hitung ulang *mean* dari semua *cluster* dalam *dataset* sampai tidak ada yang berubah pada setiap kelompoknya.

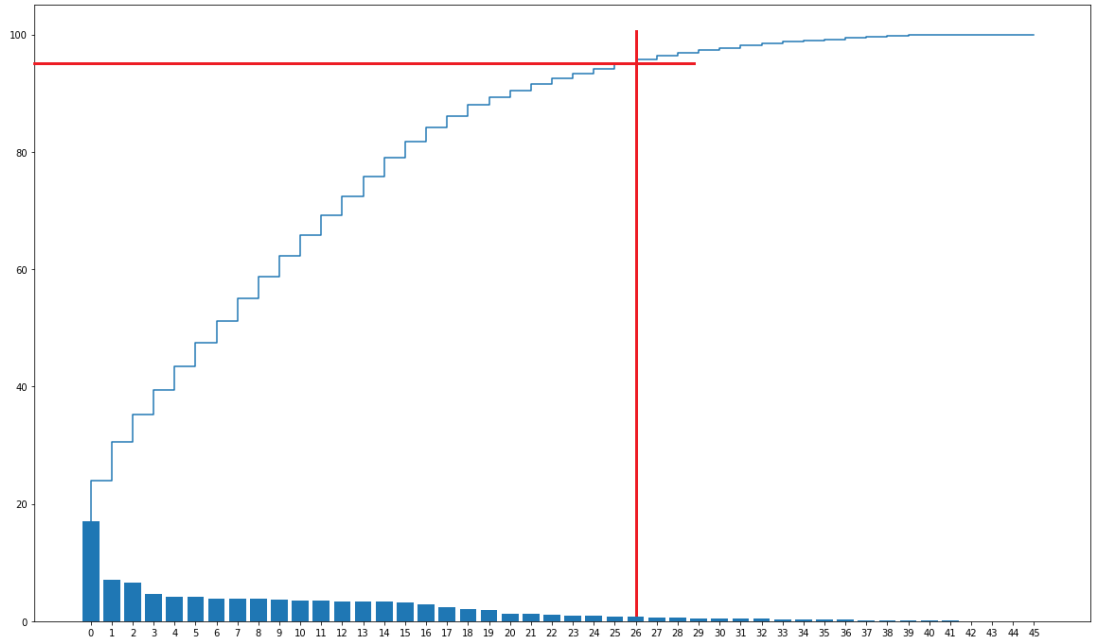
Pycaret merupakan aplikasi *open-source,* dengan teknik *low-code* dalam menjalankan instruksi ML[22]. Pada penelitian ini, peneliti memilih 5 model *classification*, dari 18 model algoritma yang disedikan *library Pycaret* seperti Gambar 2. Ekperimen ke 5 model algoritma melalui tahapan *training*, *testing* dengan *tuning-parameter* dan evaluasi model untuk prediksi dari model yang dipilih setelah process *clustering* dengan menggunakan teknik PCA dan *K-means*. Lima model *classification* yang digunakan dalam eksperimen penelitian ini adalah *Light Gradient Boosting Machine* (LightGBM), *Extreme Gradient Boosting* (XGBoost), *CatBoost Classifier* (CatBoost), *Random Forest* (RF) *Classifier*, dan *Decision Tree* (DT) *Classifier*.

ID	Name	Reference	Turbo
lr	Logistic Regression	sklearn.linear_model._logistic.LogisticRegression	True
knn	K Neighbors Classifier	sklearn.neighbors._classification.KNeighborsCl...	True
nb	Naive Bayes	sklearn.naive_bayes.GaussianNB	True
dt	Decision Tree Classifier	sklearn.tree._classes.DecisionTreeClassifier	True
svm	SVM - Linear Kernel	sklearn.linear_model._stochastic_gradient.SGDC...	True
rbfsvm	SVM - Radial Kernel	sklearn.svm._classes.SVC	False
gpc	Gaussian Process Classifier	sklearn.gaussian_process._gpc.GaussianProcessC...	False
mlp	MLP Classifier	sklearn.neural_network._multilayer_perceptron....	False
ridge	Ridge Classifier	sklearn.linear_model._ridge.RidgeClassifier	True
rf	Random Forest Classifier	sklearn.ensemble._forest.RandomForestClassifier	True
qda	Quadratic Discriminant Analysis	sklearn.discriminant_analysis.QuadraticDiscrim...	True
ada	Ada Boost Classifier	sklearn.ensemble._weight_boosting.AdaBoostClas...	True
gbc	Gradient Boosting Classifier	sklearn.ensemble._gb.GradientBoostingClassifier	True
lda	Linear Discriminant Analysis	sklearn.discriminant_analysis.LinearDiscrimina...	True
et	Extra Trees Classifier	sklearn.ensemble._forest.ExtraTreesClassifier	True
xgboost	Extreme Gradient Boosting	xgboost.sklearn.XGBClassifier	True
lightgbm	Light Gradient Boosting Machine	lightgbm.sklearn.LGBMClassifier	True
catboost	CatBoost Classifier	catboost.core.CatBoostClassifier	True

Gambar 2. Model algoritma – library pycaret

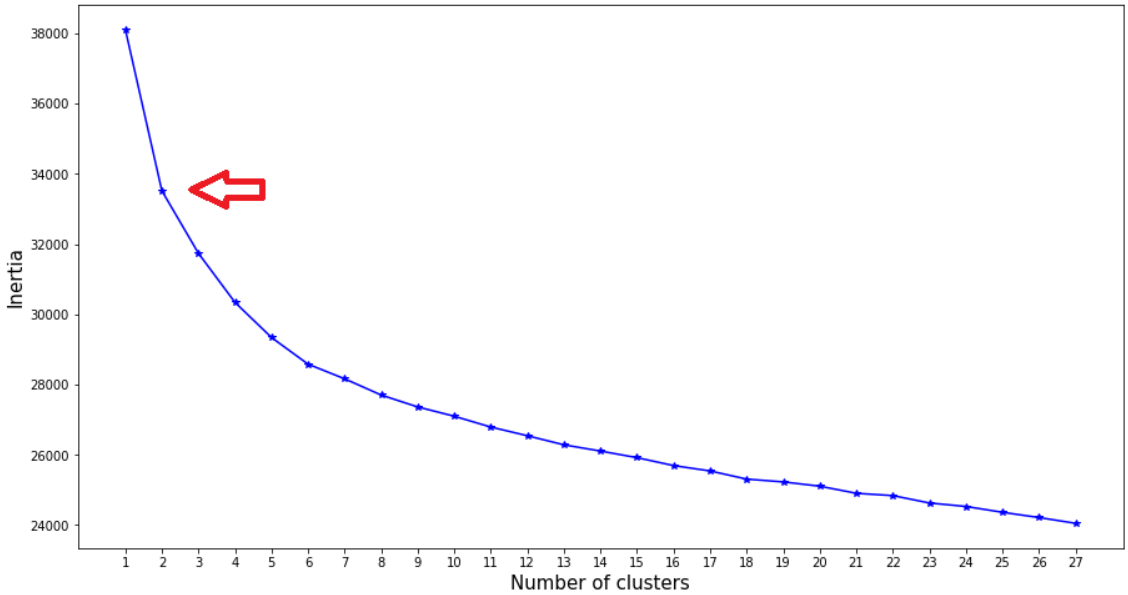
4. Hasil dan Pembahasan

Pengujian pada *dataset* IBM HR *Analytics*, menggunakan pemrograman *Python* dengan menghitung *instance cluster* yang sesuai secara proporsional. Persentase kinerja telah dihitung untuk menganalisis algoritma pada kumpulan data dengan menggunakan PCA sebagai parameter perbandingan. Berdasarkan grafik kumulatif pada Gambar 3, grafik rasio varian kumulatif menunjukkan score 94.98% dan berada di komponen ke-27. Setelah komponen ke-27 tidak ada banyak perbedaan nilai varian yang diperlihatkan.



Gambar 3. Plot kumulatif varian

Setelah jumlah komponen ditentukan, proses berikutnya adalah menentukan jumlah kluster atau nilai –k dengan menggunakan metode *elbow*. Pada tahapan ini dilakukan dengan menentukan koherensi internal sebuah *cluster* menggunakan *Inertia*. Data dikelompokkan menggunakan algoritma *K-Means* dengan membagi sampel menjadi 2 kelompok sesuai Gambar 4. Hasil dari *elbow plot* dan *inertia score*, nilai *cluster* terbaik ada di 2, yaitu 33518.95. Hal ini karena setelah *cluster* ke-2 perubahan nilai *inertia* sangat kecil dibandingkan dengan yang lain. Nampak terlihat patahan pada *cluster* ke-2, sehingga dari hasil ini akan di bangun model *K-means* menggunakan 2 *cluster.*



Gambar 4. Jumlah *Cluster vs. Inertia*

Setelah menentukan jumlah *cluster* berdasarkan skalabilitas banyaknya sampel. Setiap *cluster* di deskripsikan dengan nilai *mean* dari sampel. *Mean* biasanya disebut sebagai "centroids" dari *cluster*. Untuk mengukur kualitas hasil *clustering*, dengan melihat nilai *Adjusted Rand Index* (ARI) dan *Adjusted Mutual Information* (AMI). Semakin besar nilai ARI, semakin baik kualitas *cluster* yang terbentuk. Sedangkan nilai AMI menyatakan keseimbangan antara kelompok dan melihat kelompok kecil yang terbentuk dari *cluster*. Sesuai Tabel 1, hasil K-means *clustering* dibandingkan dengan *Agglomerative clustering*, nilai ARI K-means diperoleh nilai 0.995 dan lebih besar dari ARI – *Agglomerative* yang hanya 0.508. Hasil ini membuktikan pemilihan metode *K-means clustering* untuk dataset yang digunakan lebih baik dibandingkan *Agglomerative*.

Tabel 1. Perbandingan performa *K-means clustering*

Clustering	ARI	AMI	Homogeneity	Completeness	V-measure
K-means	0.995	0.986	0.986	0.985	0.986
Agglomerative	0.508	0.383	0.375	0.392	0.383

Hasil analisis menunjukkan bahwa algoritma *K-means* bekerja dengan baik tanpa memasukkan filter PCA dibandingkan dengan algoritma *hierarchical clustering-agglomerative*, karena memiliki lebih sedikit contoh objek yang salah *cluster* berdasarkan pengelompokan kelas. Pengelompokan hierarkis dibandingkan dengan pengelompokan cepat terjauh memberikan kinerja yang lebih baik dengan hasil *Silhouette* 0.43, *Calinski-Harabasz* 1708.82 dan *Davies-Bouldin* 0.74.

Selanjutnya tahap pembentukan model *classification* menggunakan *library Pycaret*. Pelatihan dan evaluasi dari *dataset* yang dilakukan menghasilkan skor akhir dari model yang lakukan secara independen oleh *Pycaret*. Penilaian dilakukan secara internal pada modul *PyCaret* menggunakan *cross-validation* bertingkat 10 kali lipat yang diurutkan berdasarkan akurasi. Hasil *Accuracy* klasifikasi ditampilkan dalam metrik performa untuk perbandingan dari

masing-masing model, termasuk nilai *Precision*, *Recall*, *score F1*, *Area Under Curve* (AUC), *Kappa*, dan MCC. Terdapat 15 model algoritma yang ada di *library pycaret* seperti ditunjukkan pada Gambar 4. Model XGBoost mencapai *accuracy* tertinggi dari hasil *training model Pycaret* dan memiliki waktu proses 0.128 detik. Sedangan *CatBoost* memiliki waktu proses yang paling lama dibandingkan dengan model lainnya saat pelatihan ini, yaitu 6.428 detik.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
xgboost	Extreme Gradient Boosting	0.9920	0.9987	0.9958	0.9918	0.9937	0.9826	0.9830	0.1280
rf	Random Forest Classifier	0.9893	0.9991	0.9958	0.9880	0.9917	0.9766	0.9775	0.2050
gbc	Gradient Boosting Classifier	0.9893	0.9861	0.9917	0.9920	0.9917	0.9768	0.9774	0.1280
catboost	CatBoost Classifier	0.9893	0.9997	0.9958	0.9880	0.9917	0.9766	0.9775	6.4280
lightgbm	Light Gradient Boosting Machine	0.9867	0.9984	0.9875	0.9920	0.9895	0.9712	0.9720	0.2310
dt	Decision Tree Classifier	0.9866	0.9846	0.9917	0.9878	0.9895	0.9710	0.9718	0.0110
ada	Ada Boost Classifier	0.9866	0.9959	0.9875	0.9918	0.9894	0.9712	0.9720	0.1030
lr	Logistic Regression	0.9785	0.9978	0.9875	0.9798	0.9835	0.9527	0.9533	0.9100
svm	SVM - Linear Kernel	0.9758	0.0000	0.9833	0.9800	0.9814	0.9469	0.9480	0.0110
nb	Naive Bayes	0.9653	0.9966	0.9792	0.9675	0.9730	0.9245	0.9259	0.0130
et	Extra Trees Classifier	0.9545	0.9967	0.9958	0.9378	0.9657	0.8981	0.9028	0.1750
ridge	Ridge Classifier	0.9383	0.0000	0.9792	0.9305	0.9534	0.8621	0.8678	0.0110
lda	Linear Discriminant Analysis	0.9383	0.9947	0.9792	0.9305	0.9534	0.8621	0.8678	0.0120
qda	Quadratic Discriminant Analysis	0.9116	0.9752	0.9415	0.9266	0.9313	0.8063	0.8139	0.0130
knn	K Neighbors Classifier	0.8819	0.9565	0.9750	0.8615	0.9138	0.7284	0.7474	0.0230

Gambar 5. Training model algoritma

Dari 5 model algoritma yang dipilih, setelah proses pelatihan dilakukan *tuning-parameter*. Semua algoritma klasifikasi akan menggunakan prosedur *cross-validation* sebanyak 10 kali lipat dengan maksimum iterasi sebanyak 7, *task* 42 dan *task* 100 dilakukan dengan lama waktu proses yang berbeda-beda. Waktu tercepat pada *task* 42 dihasilkan oleh model DT yaitu 1.25 detik, dan waktu terlama oleh model RF 11.08 detik. Sedangkan pada task ke 100, waktu tercepat adalah model DT 1.4 detik, dan waktu terlama model RF selama 24.9 detik.

Proses cross-validation 10 kali lipat membagi pengumpulan data menjadi sepuluh bagian yang sama. Sembilan bagian yang tersisa digunakan untuk melatih model, dan kesalahan pengujian dihitung dengan mengklasifikasikan bagian yang diberikan. Terakhir, hasil dari sepuluh tes dirata-ratakan dengan hasil performa seperti ditunjukkan pada Tabel 2.

Tabel 2. Tune-parameter classification model

Model	Accuracy	AUC	Recall	Prec.	F1
Lightgbm	0.9784	0.9950	0.984	0.9841	0.9839
XGboost	0.9838	0.9987	0.988	0.9885	0.9879
Catboost	0.9838	0.9983	0.988	0.9885	0.9879
RF classifier	0.9838	0.9973	0.984	0.9923	0.9878
DT classifier	0.9784	0.9797	0.976	0.9923	0.9835

Penurunan nilai *accuracy* model DT saat dilakukan *training* 0.9812 dan setelah *tuning-hyperparameter* 0.9784 seperti yang ditunjukkan pada Gambar 4, merupakan indikasi *overfitting* ekstrim yang kemungkinan tidak dapat diatasi dengan cara struktural atau *hyper-parameter*.

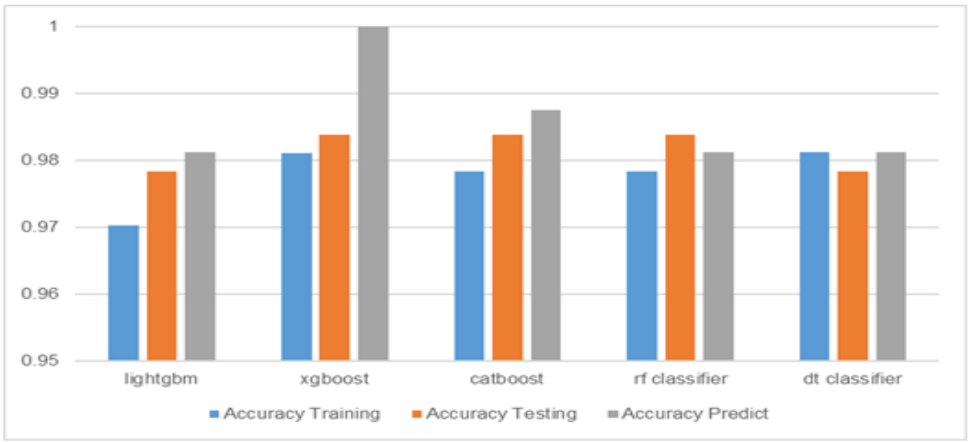
Setelah proses *tuning-parameter*, dilanjutkan proses prediksi dari ke 5 model. Model di-*deploy* secara lokal menggunakan *save_model,* model ini dapat digunakan untuk memprediksi data yang tidak terlihat menggunakan fungsi *predict_model*. Fungsi ini akan secara otomatis di terapkan pada eksperimen berdasarkan probabilitas 50% (*default*), dengan menggunakan *optimize threshold*, dan parameter *probability threshold* dalam *predict* model untuk

menghasilkan prediksi pada *hold-out / test set*. Hasil prediksi ke-5 model dapat dilihat pada Tabel 3. Dimana performa klasifikasi model XGboost mencapai hasil *accuracy* tertinggi.

Tabel 3. Hasil prediksi model

Model	Accuracy	AUC	Recall	Prec.	F1
Lightgbm	0.9812	0.9997	1.0000	0.9684	0.9840
XGboost	1.0000	1.0000	1.0000	1.0000	1.0000
Catboost	0.9875	0.9994	0.9891	0.9891	0.9891
RF classifier	0.9812	0.9996	0.9891	0.9785	0.9838
DT classifier	0.9812	0.9818	0.9783	0.9890	0.9836

Hasil prediksi model DT mengalami peningkatan setelah model di evaluasi dan nilai prediksi menghasilkan akurasi 0.9812 atau sama dengan saat pelatihan seperti ditunjukkan pada Gambar 6. Komparasi *accuracy* dari ke-5 model memperlihatkan bahwa model RF mengalami pernurunan performa pada setelah tahap evaluasi yaitu prediksi. Sedangkan nilai *accuracy* pada tahap proses pelatihan meningkat setelah dilakukan *tuning-parameter*. Model DT mengalami penurunan performa setelah proses *tuning parameter*. Hal ini mengindikasikan bahwa terjadi *overfitting* pada proses *tuning-parameter*. Sedangkan ke-3 model lainnya yaitu: *LightGBM, XGBoost* dan *Catboost* mengalami peningkatan performa setelah proses *tuning-parameter*, begitu juga pada tahap evaluasi untuk prediksi model yang dihasilkan mengalami peningkatan performa dari ke-3 model tersebut. Sedangkan model RF dan DT menunjukkan performa yang tidak konsisten seperti ditunjukkan pada Gambar 6, grafik bar untuk *accuracy training, accuracy testing (tuning-parameter)* dan *accuracy predict*.



Gambar 6. Performa Model Algoritma

Pada Tabel 4 menunjukkan bahwa pengklasifikasi berpasangan menggunakan *McNemar test* dimana nilai-p lebih besar dari 0.5 secara konsisten adalah *underpredicted* dan *overpredicted* pada *dataset* tes yang sama. Ini berarti pengklasifikasi memiliki proporsi kesalahan yang sama pada saat pengujian. Selain itu, model yang di pilih sebagian besar tidak signifikan. Dengan kata lain, jika model yang berbeda membuat kesalahan yang berbeda maka dapat dikatakan satu model lebih akurat daripada model yang lain. Oleh karena itu, hampir semua pengklasifikasi dianggap cocok untuk memprediksi kepuasan kerja pekerja dengan menggunakan variabel yang ada pada *dataset* setelah melalui proses *clustering* menggunakan *K-means* dan PCA.

Tabel 4. Komparasi *McNemar test*

Model	LightGBM	Xgboost	Catboost	RF classifier	DT classifier
Lightgbm	-	p = 1.000	p = 1.000	p = 1.000	p = 1.000
XGboost	-	-	p = 0.375	p = 0.688	p = 0.688
Catboost	-	-	-	p = 1.000	p = 1.000

Model	LightGBM	Xgboost	Catboost	RF classifier	DT classifier
RF classifier	-	-	-	-	p = 1.000
DT classifier	-	-	-	-	-

Analisis prediksi adalah teknik di mana pengguna memprediksi masa depan berdasarkan situasi saat ini. Analisis prediksi terdiri dari dua langkah. Tahap pertama adalah *clustering*, yaitu mengelompokkan jenis data yang sejenis dan tidak sejenis. Langkah kedua terdiri dari klasifikasi yang akan mengklasifikasikan data terklaster untuk analisis prediksi. Pada penelitian ini, *K-mean clustering* digunakan untuk *clustering* tersebut. Ke 5 model *classifier* digunakan untuk mengklasifikasikan kumpulan data untuk memprediksi data yang kompleks. Keakuratan pengelompokan dan klasifikasi berkurang ketika beberapa titik tetap tidak mengelompok atau salah mengelompok pada saat model DT di *testing*. Teknik normalisasi diterapkan untuk mengurangi kompleksitas kumpulan data besar, akan dapat menghitung jarak *euclidean* secara dinamis dan mempertahankan akurasi maksimum. Karena normalisasi bekerja lebih baik ketika kumpulan data bersifat kompleks. Hasil yang dinormalisasi membuat informasi dapat di analisis secara spesifik dan mengarah pada peningkatan akurasi klasifikasi.

Keterbatasan (*limitation*) dari hasil penelitian ini menemukan bahwa implementasi dari algoritma *K-means* bekerja dengan baik, analisis kompleksitas dari prosedur yang ada disarankan untuk diuji dalam penelitian selanjutnya. Namun, peneliti berpendapat bahwa kompleksitas proses akan lebih baik ketika membandingkannya dengan kompleksitas algoritma *K-means* reguler pada dataset secara penuh. Hal ini akan mempengaruhi jumlah iterasi K-means per kelompok. Hal lain yang perlu di perhatikan juga bahwa pada *dataset* yang besar, semua perhitungan *K-means* dapat dilakukan secara paralel, yaitu pada *datanode* yang berbeda. Oleh karena itu, peneliti berpendapat bahwa kompleksitas akan sangat dipengaruhi oleh besarnya ukuran grup yang akan dihasilkan.

5. Kesimpulan

Penerapan algoritma clustering K-means pada penelitian ini adalah untuk mempartisi atribut pekerja ke dalam cluster yang berbeda berdasarkan variabel pendukung lainnya. Ketika model yang sesuai dihasilkan, algoritma ini dapat dikembangkan untuk memprediksi tingkat kepuasan kerja pekerja di semua jenis organisasi. Kepuasan kerja pekerja adalah salah satu hal yang sangat penting dan dapat merugikan perusahaan dan pekerja jika tidak dikelola dengan baik.

Dari hasil eksperimen dari *dataset* yang digunakan, 3 model *classification* yaitu *XGBosst*, *CatBoost* dan *LightGBM* menunjukkan performa yang signifikan mulai tahap pembentukan model *training*, *testing* menggunakan *tuning parameter* dan evaluasi untuk menghasilkan prediksi dari model yang ada. Dengan Jumlah komponen variabel yang menjadi faktor analisis sebanyak 27 dan dibagi menjadi 2 *cluster*. Performa model algoritma menghasilkan nilai *accuracy* diatas 98% dan AUC rata-rata 0.99.

Prediksi terhadap faktor-faktor yang mempengaruhi tingkat kepuasan kerja pekerja berdasarkan *dataset* yang ada akan menghasilkan hasil yang lebih baik dan dapat digunakan dalam pengambilan keputusan manakala dilengkapi dengan *Program Employee Engagement* (EE) dan *Employee Value Proposition* (EVP) melalui hasil *survey* dan evaluasi perusahaan secara berkala. Metode analisis dengan menggabungkan teknik *clustering* dan *classficiation* dalam memprediksi tingkap kepuasan pekerja ini akan lebih optimal dengan memasukkan teknik *data mining*. Implementasi lebih lanjut dapat dikembangkan untuk menghasilkan data analisis dan sistem terpadu di organisasi..

Daftar Referensi

[1] S.J. Conlon, L.L. Simmons, & F. Liu, "Predicting Tech Employee Job Satisfaction Using Machine Learning Techniques". Int J Manag Inf Technol, vol. 16, pp. 72–88, 2021

[2] H. Gao, M. He, & G. Hou, "A Comparison of Machine Learning Approaches For Employee Satisfaction Prediction", Int J Mod Eng Res, vol. 10, pp. 46–53, 2020

[3] C.I. Ristianty, Nurwati, Nasrul, et al. "The influence of leadership on human resource empowerment and job satisfaction in improving employee performance". Eur J Mol Clin Med, vol. 7, no. 3844–3852, 2020

[4] Y. Jung, Y. Suh, "Mining The Voice of Employees: A Text mining Approach to Identifying

And Analyzing Job Satisfaction Factors From Online Employee Reviews", Decis Support Syst, vol. 123, no. 113074, 2019
[5] S. Andari, "Hubungan Antara Gaya Kepemimpinan Dengan Kinerja Karyawan", Pros Psikol, vol. 5, no.2, pp. 6, 2017.
[6] R.N. Londok, W.A. Areros, & S. Asaloei, "Pengaruh Kepuasan Kerja Terhadap Kinerja Karyawan CV. Diagram Global Mandiri Manado", J Adm Bisnis, no.9, pp. 6, 2019
[7] S.T. Dziuba, M. Ingaldi, & M. Zhuravskaya "Employees' Job Satisfaction and their Work Performance as Elements Influencing Work Safety", Syst Saf Hum - Tech Facil - Environ, vol. 2, pp.18–25, 2020
[8] C. KUZEY, "Impact of Health Care Employees ' Job Satisfaction on Organizational Performance Support Vector Machine Approach", J Econ Financ Anal, vol. 2, pp. 45–68, 2018
[9] H. Bodepudi, "Credit Card Fraud Detection Using Unsupervised Machine Learning Algorithms", Int J Comput Trends Technol, vol. 69, pp. 1–13, 2021
[10] M. Usama, J. Qadir, A. Raza, et al. "Unsupervised Machine Learning For Networking: Techniques, Applications And Research Challenges", *IEEE,* vol. 2, pp. 37, 2019
[11] X. Liu, N. Zhang, "Research on Customer Satisfaction of Budget Hotels Based on Revised IPA and Online Reviews", Science Journal of Business and Management, http://article.bizmgmt.org/pdf/10.11648.j.sjbm.20200802.11.pdf (2020).
[12] R. Hidayati, A. Zubair, P.A. Hidayat, et al. "Silhouette Coefficient Analysis in 6 Measuring Distances of K-Means Clustering", TechnoCom, vol. 20, pp. 186–197, 2021
[13] J.J. Shon, "Employee Satisfaction and Corporate Performance : Mining Employee Reviews on Glassdoor.com". Thirty Seventh Int Conf Inf Syst, pp. 1–16, 2016.
[14] O. Somantri, D. Apriliani, "Support Vector Machine Based on Feature Selection For Sentiment Analysis Customer Satisfaction on Culinary Restaurant at Tegal City". J Teknol Inf dan Ilmu Komput, vol. 5, pp. 537, 2018
[15] Y. Chen, X. Qin, J. Wang, et al. "Fedhealth: A federated transfer learning framework for wearable healthcare". *IEEE Intell Syst*, https://ieeexplore.ieee.org/abstract/ document/9076082, 2020.
[16] A. Sarker, S.M. Shamim, M. Shahiduz, et al. "Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm", Int Res J Softw Data Eng Glob J Comput Sci Technol, vol. 18, pp 7, 2018
[17] A.Al. Malki, M.M. Rizk, "El-Shorbagy MA, et al. Hybrid Genetic Algorithm with K-Means for Clustering Problems", Open J Optim, vol. 05, pp. 71–83, 2016
[18] D.Q. Zeebaree, H. Haron, A.M. Abdulazeez, et al. "Combination of K-Means Clustering With Genetic Algorithm: A Review", Int J Appl Eng Res, vol. 12, pp. 14238–14245, 2017
[19] S.G. Mathias, D. GroBmann, G.J. Sequeira, "A Comparison of Clustering Measures on Raw Signals of Welding Production Data", *International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*. Epub ahead of print 2019. DOI: 10.1109/deep-ml.2019.00019, *2019.*
[20] A. Adolfsson, M. Ackerman, N.C. Brownstein, "To Cluster , or Not to Cluster : An Analysis of Clusterability Methods", arXiv*,* vol. 1, pp. 30, 2018
[21] M. Masoud, Y. Jaradat, E. Rababa, et al. "Turnover Prediction using Machine Learning: Empirical Study", Int J Adv Soft Comput its Appl, vol. 13, pp. 193–207, 2021
[22] V. Mulpuru, N. Mishra, "In Silico Prediction of Fraction Unbound in Human Plasma from Chemical Fingerprint Using Automated Machine Learning", ACS Omega, vol. 6, pp. 6791–6797, 2021
[23] E. Larsen, D. Noever, K. Macvittie, et al. "Overhead-MNIST : Machine Learning Baselines For Image Classification", arXiv, vol. 2, pp. 6, 2021
[24] J. Judrups, R. Cinks, I. Birzniece, et al. "Machine learning based solution for predicting voluntary employee turnover in organization", In: *Engineering for Rural Development*, pp. 1359–1366, 2021
[25] F.L. Gewers, G.R. Ferreira, H.F. Arruda De, et al. "Principal Component Analysis: A Natural Approach to Data Exploration", arXiv, vol. 2, pp. 1–33, 2018
[26] G. Wijngaard. *Thesis : Clustering soccer players: investigating unsupervised learning on player positions*, 2019.
[27] E. Bisong, "Google Cloud Machine Learning Engine (Cloud MLE)". Building Machine Learning and Deep Learning Models on Google Cloud Platform, pp. 545–579, 2019.