JURNAL ILMIAH INFORMATIKA GLOBAL
VOLUME 13 Nomor 02 Agustus 2022
e-ISSN 2477-3786* | *p-ISSN 2302-500X
DOI : 10.36982/jiig.v13i2.2292

Perbandingan Akurasi Metode *Principal Component Analysis* (PCA) dan *Correlation-Based Feature Selection* (CFS) Pada Klasifikasi Perpanjangan Kontrak Karyawan Menggunakan Metode *Naïve Bayes*

Dewi Sartika[1)], Imelda Saluza[2)], Muhammad Haviz Irfani[3)]

[1),3)]Program Studi Teknik Informatika, Universitas Indo Global Mandiri
[2)]Program Studi Manajemen Informatika, Universitas Indo Global Mandiri
Jalan Jendral Sudirman No. 629 KM 4 Palembang 30129
Email : dewi.sartika@uigm.ac.id[1)], imeldasaluza@uigm.ac.id[2)], m.haviz@uigm.ac.id[3)]

ABSTRACT

PT. Oasis Waters International Palembang conducts regular staff performance reviews, the findings of which are utilized to make recommendations for employee contract extension. The Human Resource Department has assigned a numerical value to 25 qualities (HRD). The process of giving a label or class to a number of examples when the value of each characteristic is known as classification. The Naïve Bayes technique is a basic classification approach that makes use of probability estimates. Based on the observations, it was discovered that one of the 25 criteria was deemed the most relevant in determining the recommendation for an employee contract renewal. As a result, in this study, a comparison of the pre-processing Principal Component Analysis (PCA) approach and the Correlation-based Feature Selection (CFS) method on the categorization of employee contract extensions at PT Oasis Waters International Palembang will be performed. According to the data, the CFS approach has a positive influence on classification performance, while PCA does not. This is demonstrated by a 30% increase in accuracy when utilizing the CFS approach. Meanwhile, both strategies have a positive influence on the model's dependability. This is demonstrated by a reduction in Root Mean Square Error (RMSE) when using the CFS approach from 0.6325 to 0.1845, whereas using the PCA method results in 0.5123.

Keywords : *Naïve Bayes, Principal Component Analysis, Correlation-based Feature Selection, Confusion Matrix, Root Mean Square Error*

ABSTRAK

PT Oasis Waters International Palembang melakukan penilaian kinerja karyawan secara rutin yang hasilnya digunakan dalam penentuan rekomendasi perpanjangan kontrak karyawan. Terdapat sebanyak 25 atribut yang diberikan nilai berupa numerik oleh pihak *Human Resource Department* (HRD). Klasifikasi merupakan proses pemberian label atau kelas pada sejumlah kasus yang nilai dari tiap - tiap atributnya sudah diketahui. Salah satu metode klasifikasi sederhana yang memanfaatkan perhitungan probabilitas yaitu metode Naïve Bayes. Berdasarkan hasil observasi diperoleh bahwa dari 25 atribut terdapat 1 atribut yang dianggap paling penting dalam menentukan rekomendasi perpanjangan kontrak karyawan. Oleh karena itu, pada penelitian ini akan dilakukan analisis perbandingan metode *pre-processing Principal Component Analysis* (PCA) dan metode *Correlation-based Feature Selection* (CFS) terhadap pengklasifikasian perpanjangan kontrak karyawan PT Oasis Waters International Palembang. Berdasarkan hasil yang diperoleh menyatakan bahwa metode CFS memberikan dampak baik terhadap kinerja klasifikasi, sedangkan PCA tidak. Hal ini dinyatakan dengan adanya peningkatan akurasi sebesar 30% jika menggunakan metode CFS. Sedangkan dalam peningkatan kehandalan model, kedua metode memberikan dampak baik. Hal tersebut dapat dilihat dengan adanya penurunan *Root Mean Square Error* (RMSE) pada penggunaan metode CFS yang semula 0,6325 menjadi 0,1845, sedangkan pada penggunaan metode PCA menjadi 0,5123.

Kata Kunci : Naïve Bayes, Principal Component Analysis, Correlation-based Feature Selection, Confusion Matrix, Root Mean Square Error

Article History
Received : 10/05/2022
Revised : 24/06/2022
Accepted : 08/07/2022
Online : 01/08/2022

1. Pendahuluan

Karyawan kontrak merupakan karyawan pada suatu perusahaan yang dipekerjakan dalam batas waktu tertentu. Setiap habis masa kontrak karyawan tersebut akan dievaluasi berdasarkan kinerjanya. Seperti pada PT. Oasis Waters International Palembang rutin melakukan evaluasi kinerja pada karyawan kontrak yang akan habis masa kerja. Evaluasi kinerja dilakukan dengan melakukan pemberian nilai oleh kepala bagian yang hasilnya akan diusulkan ke *Human Resource Department* (HRD). Terdapat 25 kriteria yang harus dinilai oleh kepala bagian. Penilaian dilakukan dengan cara mengisi formulir yang kemudian akan dilakukan perhitungan oleh pihak HRD untuk penentuan perpanjangan kontrak karyawan.

Saat ini PT Oasis Waters International Palembang memiliki 140 karyawan yang terdiri dari karyawan bulanan kontrak dan karyawan harian lepas. Karyawan bulanan kontrak merupakan karyawan yang memiliki kontrak selama 12 bulan, sedangkan karyawan harian lepas memiliki kontrak selama 2 bulan. Oleh sebab itu penilaian kinerja karyawan kontrak pada PT Oasis Waters Indonesia Palembang dilakukan hampir setiap bulan, sehingga diperlukan alternatif yang dapat meringankan kerja HRD dalam melakukan pemutusan perpanjangan kontrak. Selain itu PT Oasis Waters International Palembang ingin berupaya mengurangi penilaian perpanjangan kontrak secara subjektif.

Klasifikasi merupakan proses pemberian label atau kelas pada sejumlah kasus yang nilai dari tiap - tiap atributnya sudah diketahui. Pemberian label atau kelas didasarkan pada sejumlah data yang telah diketahui label atau kelasnya, atau lebih sering dikenal sebagai data latih. Salah satu metode yang dapat digunakan untuk proses klasifikasi yaitu metode Naïve Bayes. Prinsip kerja dari metode ini adalah dengan melakukan perhitungan probabilitas dari data latih dengan kasus baru. Metode Naïve Bayes dalam menghitung estimasi parameter yang diperlukan dalam pengklasifikasian hanya membutuhkan sejumlah data yang kecil (Kusrini, 2017)

Pre-processing merupakan tahapan yang dilakukan sebelum proses data mining, baik pengklasifikasian maupun dalam pengklasterisasi. Salah satu teknik yang digunakan dalam *pre-processing* adalah seleksi fitur. Seleksi fitur bekerja dengan memilih fitur atau atribut yang berkorelasi dengan target kelas. Metode *Correlation based Feature Selection* (CFS) dibandingkan dengan metode seleksi fitur lainnya memiliki tingkat akurasi yang paling tinggi (Djatna & Morimoto, 2008). Terdapat juga pre-processing yang berupa perhitungan statistika, salah satunya adalah *Principal Component Analysis* (PCA). Prinsip kerja PCA adalah mengubah sejumlah data asli yang saling berkorelasi menjadi set data baru yang saling bebas (Noya van Delsen et al., 2017). Oleh karena itu, pada penelitian ini akan menganalisis pengaruh metode CFS dan PCA pada klasifikasi perpanjangan kontrak karyawan PT. Oasis Waters International Palembang menggunakan metode Naïve Bayes.

2. Pembahasan

Data mining secara deskriptif berfungsi untuk memahami data hingga mampu memahami karakteristik serta pola tertentu yang tersembunyi didalamnya. Sedangkan secara prediktif fungsi data mining sebagai proses dalam menentukan pola dari data, dimana pola tersebut diketahui berdasarkan variable - variable data untuk menentukan variable lainnya yang belum diketahui.

Klasifikasi merupakan proses dalam menentukan objek data ke dalam kelas. Klasifikasi termasuk dalam *supervised learning* yang artinya label atau kelas dari suatu data telah diketahui. Penentuan kelas atau label didasarkan pada set atribut. Atribut yang digunakan dalam penentuan kelas atau label dapat berupa data numerik, nominal, binary, maupun ordinal, sedangkan kelas atau label berupa data nominal.

Prinsip klasifikasi adalah menggunakan sebagaian data yang telah diketahui kelas atau labelnya digunakan sebagai data latih untuk menentukan data yang belum diketahui kelas atau labelnya. Tujuan dari klasifikasi adalah menentukan data kedalam kelas atau label yang benar. Selain itu, klasifikasi juga dapat digunakan dalam membantu memahami pengelompokan data.

Pre-processing merupakan teknik yang digunakan untuk menghasilkan data yang berkualitas. Tahapan ini dapat dilakukan sebelum melakukan proses data mining. *Pre-processing* menghasilkan data yang lebih sedikit dibandingkan data aslinya, namun secara signifikan dapat meningkatkan efisiensi data mining. Pre-processing ada yang bertujuan untuk (1) membersihkan data yang tidak konsisten, memiliki *noise*, maupun tidak lengkap; (2) pemilihan atribut maupun penghapusan duplikasi data; dan (3) memulihkan data yang tidak lengkap, menghapus kesalahan atau *outlier* maupun mengisi nilai atribut yang terlewat (Zhang et al., 2003).

Atribut yang digunakan dalam penentuan kelas atau label dari suatu data dimungkinkan untuk dipilih dengan tujuan menghindari *overfitting* maupun peningkatan kinerja dari proses data mining. Terdapat banyak atribut yang digunakan dalam pengklasifikasian memungkinkan untuk dipilih berdasarkan yang paling relevan dan berulang (Khan et al., 2021). Pemilihan fitur merupakan hal yang penting dilakukan agar dapat mempercepat proses belajar, untuk meningkatkan kualitas konsep (Kira & Rendell, 1992) serta untuk meningkatkan akurasi dalam klasifikasi (Ranjan et al., 2021) . Metode seleksi fitur terbagi menjadi dua yaitu metode *wrapper* (pembungkusan) dan metode filter. Perbedaanya pada metode wrapper bergantung dengan model klasifikasi yang digunakan, sedangkan metode filter tidak (Beniwal & Arora, 2012).

Principal Component Analysis (PCA) merupakan salah satu metode yang digunakan untuk mereduksi dimensi data pada tahap *Pre-processing* (Myoelectric et al., 2009). PCA termasuk salah satu metode ekstraksi fitur yang biasanya digunakan dalam data kontinu. Prinsip kerja metode ini adalah dengan mengekstaksi

atribut sehingga menyisakan atribut yang bertujuan untuk memperoleh hasil lebih optimal.

PCA disusun menggunakan komponen varian maksimum secara terurut dan setelahnya nilai varian akan menurun atau dengan kata lain komponen pertama merupakan komponen dengan nilai varian tertinggi, dan komponen seterusnya memiliki nilai varian yang lebih rendah (Widagdo et al., 2020). Metode ini terdiri dari 4 tahapan yaitu :

1. Mencari sejumlah data yang berdimensi m x n, dimana m adalah jumlah sampel data sedangkan n adalah jumlah atribut.

$$\boldsymbol{X^*_{i,j} = X_{i,j} - \bar{X}} \quad (1)$$

Dimana :
$X_{i,j}$ = Elemen Matrik X
$X^*_{i,j}$ = Elemen Matrik X*
$\bar{X}$ = nilai rata – rata matrik X

2. Mencari nilai kovarian (C_x) dari sejumlah data dengan persamaan 2 :

$$\boldsymbol{C_x = \frac{1}{m-1} \cdot X^{*T}_{i,j} \cdot X^*_{i,j}} \quad (2)$$

3. Menghitung nilai eigen (?) dengan persamaan 3, dimana I merupakan matrik identitas dan v merupakan *vector eigen*:

$$\boldsymbol{|C_x - \lambda I| = 0 \; dan \; (C_x - \lambda I) \cdot v = 0} \quad (3)$$

4. Menghitung *persentase* kontribusi komulatif variansi (V_r) , dimana d adalah jumlah atribut awal dan r adalah jumlah komponen yang dipilih.

$$\boldsymbol{V_r = \frac{\sum_j^r \lambda_j}{\sum_j^d \lambda_j} \cdot 100\%} \quad (4)$$

Correlation based Feature Selection (CFS) merupakan salah satu metode seleksi fitur yang bekerja dengan cara mengevaluasi subset dari atribut berdasarkan pada nilai korelasi. Atribut dipilih berdasarkan manfaat atribut secara individual dalam memprediksi kelas atau label dari suatu data sekaligus level inter-korelasi antar atribut. Dalam proses seleksinya CFS menggunakan matriks korelasi untuk mendapatkan nilai korelasi atribut yang tinggi pada kelas namun atribut tersebut tidak berkorelasi dengan atribut lainnya (Ifriza & Sam, 2021) dan (Nurul Yusufiyah & Gya Nur Rochman, 2021). CFS bekerja dengan tujuan untuk menghilangkan atribut yang redundan dan tidak relevan sebanyak mungkin (Hall, 1999). Redundan ditandai dengan korelasi yang tinggi satu sama lain sedangkan tidak relevan ditandai dengan korelasi yang rendah pada kelas (Pramadhana, 2021).

Pada CFS koefisien korelasi digunakan untuk mengukur korelasi berpasangan dari semua atribut (Jakob, 2016). CFS dapat digambarkan sebagai pengukuran nilai korelasi antara dua atribut atau lebih yang dilihat berdasarkan nilai korelasi atau di dalam statistik dinotasikan dengan nilai r. jika nilai $r = 1$ maka atribut memiliki korelasi positif yang sempurna, jika $r = -1$ maka atribut memiliki korelasi negatif yang sempurna dan $r = 0$ berarti tidak terdapat korelasi antar atribut. CFS mewajibkan setiap atribut bernilai numerik untuk didiskritkan terlebih dahulu secara *symmetrical uncertainty* guna mengestimasi derajat keterhubungan antar dua fitur diskrit (Purbasari et al., 2013). Perhitungan metode CFS dapat dilihat pada persamaan 8 (Karegowda et al., 2010).

$$\boldsymbol{r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k+k(k-1)\overline{r_{ii}}}}} \quad (5)$$

Dimana r_{zc} merupakan korelasi antar fitur dengan variable kelas, k adalah jumlah himpunan fitur, r_{zi} adalah rata – rata korelasi antara subset dalam penentuan variable kelas, sedangkan r_{ii} adalah rata – rata korelasi antara fitur subset. Selanjutnya untuk memperkirakan manfaat dari subset fitur diperlukan perhitungan ketergantungan antar atribut. Persamaan yang digunakan pada data diskrit adalah symmetrical uncertainty yang dapat dilihat pada persamaan 9.

$$\boldsymbol{SU = 2.0x\left[\frac{H(X)+H(Y)-H(X,Y)}{H(Y)+H(X)}\right]} \quad (6)$$

Dimana X dan Y merupakan pasangan fitur secara simetris. Sedangkan SU akan menghasilkan kisaran nilai [0,1], 1 menunjukkan bahwa fitur memprediksi nilai fitur yang lain sedangkan 0 menunjukan bahwa pasangan fitur saling bebas (Doshi & Chaturvedi, 2014).

Naïve Bayes merupakan salah satu metode klasifikasi yang menerapkan konsep peluang, sehingga dianggap sederhana untuk diterapkan. Prinsip yang digunakan pada metode ini merupakan teori Bayesian. Nilai dari setiap atribut akan digunakan dalam penentuan kelas. Pemahaman mendalam terkait karakteristik data yang digunakan sangat mempengaruhi kinerja dari metode ini, dimana metode ini akan memiliki kinerja yang baik pada fitur yang sepenuhnya independen dan fitur yang bergantung secara fungsional (Rish, 2001).

Tahapan dalam metode Naïve Bayes dalam melakukan klasifikasi data (Hakimah & Muhimah, 2021), yaitu :

1. Menghitung *prior probability* dari kelas yang diamati yaitu $P(C_i)$ dimana i = 0,1..dst

2. Menghitung *conditional probability* masing - masing atribut terhadap kelas dengan persamaan 7 :

$$\boldsymbol{P(X_j|C_i) = \frac{P(C_i|X_j)P(X_j)}{P(C_i)}} \quad (7)$$

Dimana :
$P(X_j|C_i)$: Probabilitas atribut X_j terhadap kelas C_i
$P(C_i|X_j)$: Probabilitas kelas C_i berdasarkan atribut X_j
$P(X_j)$: Probabilitas atribut ke- j
$P(C_i)$: Probabilitas kelas ke- i

3. Menghitung *posterior probability* kelas terhadap atribut menggunakan persamaan 8 :

$$P(C_i|X_1, X_2, \dots X_j) = P(C_i)x\prod P(X_j|C_i \quad (8)$$

Dimana $P(C_i|X_1, X_2, \dots X_j)$ merupakan probabilitas kelas ke-i berdasarkan data yang belum diketahui kelas atau labelnya.

4. Menentukan kelas dari objek data dengan persamaan 9 :

$$C(x) = \arg\max P(C_i|X_1, X_2, \dots X_j) \quad (9)$$

Menguraikan hasil penelitian analisis kualitatif, Penelitian ini dilakukan dalam beberapa tahapan, yaitu :

1. Studi Literatur
   Pada tahapan ini dilakukan pengumpulan literatur yang berkaitan dengan penelitian, yaitu data mining, klasifikasi, Naïve Bayes, pra-processing, seleksi fitur, *Correlation-based Feature Selection, Principal Component Analysis, Root Mean Square Error*, dan *Confusion Matrix*.
2. Perumusan Masalah
   Pada tahapan ini dilakukan observasi secara langsung ke PT Oasis Waters International Palembang untuk mengetahui proses yang selama ini berjalan dalam penentuan perpanjangan kontrak karyawan.
3. Pengumpulan Data
   Pada tahapan ini dilakukan pengumpulan data penilaian kinerja karyawan kontrak PT Oasis Waters International Palembang yang digunakan dalam menentukan perpanjangan kontrak. Berdasarkan formulir penilaian kinerja karyawan terdapat 25 kriteria yang akan digunakan dalam menentukan rekomendasi perpanjang kontrak atau direkomendasikan.
   Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari PT Oasis Waters International Palembang. Kriteria yang digunakan terdiri dari 25 atribut yaitu : loyalitas(A); disiplin(B); kesediaan memperbaiki diri(C), kesadaran akan biaya, bahan dan waktu(D); kerja keras(E); pengetahuan/penguasaan kerja(F); ketelitian(G), ketekunan (H), keuletan/daya tahan(I); kecepatan kerja(J); kemandirian kerja(K); perencanaan kerja(L); kualitas hasil kerja(M); kuantitas hasil kerja(N); kerjasama dan koordinasi(O); penyampaian ide/pemikiran(P); pengambilan keputusan(Q); stabilitas emosi(R); kepemimpinan(S); motivasi(T); tanggung jawab(U); inisiatif(V); penyesuaian diri(W); komunikasi(X); dan kreativitas(Y). Namun berdasarkan informasi yang diperoleh dari perusahaan menyatakan bahwa dari 25 atribut yang paling penting adalah atribut disiplin.
4. Analisis Data
   Pada tahapan ini data yang telah dikumpulkan akan digunakan sebagai data latih dan data uji. Analisis yang dilakukan adalah melihat pengaruh metode pre-processing *Correlation based Feature Selection* (CFS) pada klasifikasi menggunakan metode Naïve Bayes. Tahapan ini dilakukan dengan menggunakan *case tool* WEKA versi 3.8.4.
5. Penarikan Kesimpulan
   Penarikan kesimpulan dilakukan dengan melakukan perhitungan akurasi menggunakan *Confusion Matrix* dan kehandalan model menggunakan *Root Mean Square Error* (RMSE).

Berdasarkan formulir penilaian kinerja karyawan kontrak pada PT Oasis Waters International Palembang diperoleh 25 kriteria yang digunakan untuk menentukan rekomendasi perpanjangan masa kontrak. Berdasarkan informasi yang diberikan oleh *Human Resources Department* (HRD) dari 25 kriteria yang ada, terdapat 1 kriteria yang dianggap paling penting yaitu kriteria disiplin. Namun 1 kriteria tersebut tidak bisa menentukan rekomendasi perpanjangan kontrak, sedangkan kriteria yang lainnya dianggap sama penting. Semua kriteria yang digunakan bernilai numerik (1-100), sedangkan kelas rekomendasi terdiri dari perpanjangan kontrak dan dipertimbangkan. Mode pengujian yang dilakukan dalam penelitian ini adalah metode *supplied test set*, dimana sejumlah data yang digunakan dalam pengujian sudah diketahui kelas rekomendasinya.

Analisis yang akan dilakukan adalah hasil akurasi klasifikasi Naïve Bayes dengan dan tanpa *pre-processing* berdasarkan *Confusion Matrix* dan *Root Mean Square Error* (RMSE). *Pre-processing* yang digunakan berupa seleksi fitur menggunakan metode *Correlation based Feature Selection* (CFS) dan *Principal Component Analysis* (PCA) . Hasil dari penerapan masing – masing metode ini akan diperoleh kriteria terpilih dari 25 kriteria yang lebih berpengaruh dalam penentuan kelas rekomendasi. Hasil dari penerapan metode CFS diperoleh 8 kriteria terpilih yaitu disiplin (B), kesediaan memperbaiki diri (C), ketelitian (G), kecepatan kerja(J), stabilitas emosi (R), kepemimpinan(S), motivasi (T), dan inisiatif (V). Sedangkan dari penerapan metode PCA diperoleh 22 kriteria yang terpilih yaitu loyalitas(A), disiplin(B), kesediaan memperbaiki diri(C), kesadaran akan biaya, bahan dan waktu(D), kerja keras(E), pengetahuan/penguasaan kerja(F), ketelitian(G), ketekunan (H), keuletan/daya tahan(I), kecepatan kerja(J), kemandirian kerja(K), perencanaan kerja(L), kualitas hasil kerja(M), kuantitas hasil kerja(N), kerjasama dan koordinasi(O), penyampaian ide/pemikiran(P), pengambilan keputusan(Q), stabilitas emosi(R), kepemimpinan(S), motivasi(T), tanggung jawab(U), dan inisiatif(V).

Confusion Matrix dapat digunakan untuk menghitung nilai akurasi. Nilai akurasi yang mendekati 100% menandakan model tersebut semakin baik. Model dari *Confutsion Matrix* dapat dilihat pada Tabel 1 sedangkan perhitungannya bisa dilihat pada persamaan 10 [2].

Tabel 1. *Confusion Matrix*

		True Class	
		Positive	*Negative*
Predicted Class	*Positive*	*True Positives Count (TP)*	*False Negatives Count (FP)*
	Negative	*False Positives Count (FN)*	*True Negatives Count (TN)*

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (10)$$

Berdasarkan perhitungan menggunakan persamaan 10 diperoleh hasil pengujian tingkat akurasi klasifikasi *Naïve Bayes, Naïve Bayes* dengan CFS dan *Naïve Bayes* dengan PCA pada kasus perpanjangan kontrak karyawan PT Oasis Waters International Palembang yang dapat dilihat pada Tabel 2.

Tabel 2. *Hasil Tingkat Akurasi dengan Confusion Matrix*

Metode	***Confusion Matrix***		**Akurasi (%)**
Naïve Bayes	TP = 6	FP = 0	60
	FN = 4	TN = 0	
Naïve Bayes with CFS	TP = 6	FP = 0	90
	FN = 1	TN = 3	
Naïve Bayes with PCA	TP = 6	FP = 0	60
	FN = 4	TN = 0	

Root Mean Square Error (RMSE) merupakan persamaan yang dapat digunakan untuk menguji kehandalan suatu model. Semakin kecil nilai RMSE menyatakan bahwa suatu model tersebut semakin baik. Perhitungan RMSE dapat dilihat pada persamaan 11 [22].

$$RMSE = \sqrt{?_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{n}} \quad (11)$$

Dimana :
n : jumlah data
Y_i : Nilai sejati
$\hat{Y}_i$: Nilai prediksi

Kehandalan model yang digunakan dilihat berdasarkan nilai dari RMSE yang dihitung menggunakan persamaan 11 yang hasilnya dapat dilihat pada Tabel 3.

Tabel 3. *Kinerja Model Klasifikasi*

Metode	**RMSE**
Naïve Bayes	0,6325
Naïve Bayes with CFS	0,1845
Naïve Bayes with PCA	0,5123

3. Kesimpulan

Berdasarkan hasil diperoleh menyatakan bahwa Metode *Correlation based Feature Selection* (CFS) memberikan pengaruh yang cukup signifikan terhadap kinerja metode klasifikasi *Naïve Bayes* pada studi kasus perpanjangan kontrak karyawan PT Oasis Waters International Palembang. Hal ini dapat dilihat berdasarkan terjadinya peningkatan akurasi sebesar 30% serta penurunan RMSE yang semula 0,6325 menjadi 0,1845, yang berarti kehandalan model klasifikasi semakin meningkat. Sedangkan *Metode Principal Component Analysis* (PCA) tidak memberikan pengaruh terhadap kinerja metode klasifikasi *Naïve Bayes*, hal tersebut terlihat tidak adanya peningkatan akurasi, namun kehandalan dari model klasifikasi meningkat berdasarkan penurunan RMSE yang semula 0,6325 menjadi 0,5123.

Daftar Pustaka

Beniwal, S., & Arora, J. (2012). Classification and Feature Selection Techniques in Data Mining. *International Journal of Engineering Research & Technology (IJERT)*, *1*(6), 1–6.

Defiyanti, S. (2017). Integrasi Metode Clustering dan Klasifikasi untuk Data Numerik. *Citee*, *July*, 256–261.

Djatna, T., & Morimoto, Y. (2008). Pembandingan Stabilitas Algoritma Seleksi Fitur Menggunakan Transformasi Ranking Normal. *Jurnal Ilmiah Ilmu Komputer*, *6*(2), 245006.

Doshi, M., & Chaturvedi, S. K. (2014). Correlation Based Feature Selection (CFS) Technique to Predict Student Perfromance. *International Journal of Computer Networks & Communications*, *6*(3), 197–206. https://doi.org/10.5121/ijcnc.2014.6315

Hakimah, M., & Muhimah, R. R. (2021). Klasifikasi Penderita Penyakit Jantung Menggunakan Metode Naive Bayes dengan Chi-Square untuk Pemilihan Atribut. *Seminar Nasional Teknik Elektro, Sistem Informasi Dan Teknik Informatika*, *1*, 257–262.

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. *April*.

Ifriza, Y. N., & Sam, M. (2021). Irrigation management of agricultural reservoir with correlation feature selection based binary particle swarm optimization. *Journal of Soft Computing Exploration*, *2*(1), 40–45. https://doi.org/10.52465/joscex.v2i1.23

Jakob, R. (2016). Disease Classification. *International Encyclopedia of Public Health*, 332–337. https://doi.org/10.1016/B978-0-12-803678-5.00116-8

K, Gupta, G. (2014). *Introduction to Data Mining with Case Studies* (Third Edit).

Karegowda, A. G., Manjunath, A. S., Ratio, G., & Evaluation, C. F. (2010). Comparative study of Attribute Selection Using Gain Ratio. *International Journal of Information Technology and Knowledge and Knowledge Management*, *2*(2), 271–277. https://pdfs.semanticscholar.org/3555/1bc9ec8b6ee3c97c524f9c9ceee798c2026e.pdf%0Ahttp://csjournals.com/IJITKM/PDF 3-1/19.pdf

Khan, M. A., Akram, T., Sharif, M., Alhaisoni, M., Saba, T., & Nawaz, N. (2021). A probabilistic segmentation and entropy-rank correlation-based feature selection approach for the recognition of fruit

diseases. *Eurasip Journal on Image and Video Processing*, *2021*(1). https://doi.org/10.1186/s13640-021-00558-2

Kira, K., & Rendell, L. A. (1992). A Practical Approach to Feature Selection. In *Machine Learning Proceedings 1992*. Morgan Kaufmann Publishers, Inc. https://doi.org/10.1016/b978-1-55860-247-2.50037-1

Kusrini, S. E. D. A. (2017). Algoritma K-Means untuk Diskretisasi Numerik Kontinyu Pada Klasifikasi Intrusion Detection System Menggunakan Naive Bayes. *Konferensi Nasional Sistem & Informatika*, 61–66.

Myoelectric, P., Hudgins, B., Control, P. M., Hargrove, L. J., Li, G., Member, S., Englehart, K. B., & Member, S. (2009). *Principal Components Analysis Preprocessing for Improved Classification Accuracies in Principal Components Analysis Preprocessing for Improved Classification Accuracies in*. *56*(October 2016), 1407–1414.

Noya van Delsen, M. S., Wattimena, A. Z., & Saputri, S. (2017). Penggunaan Metode Analisis Komponen Utama Untuk Mereduksi Faktor-Faktor Inflasi Di Kota Ambon. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, *11*(2), 109–118. https://doi.org/10.30598/barekengvol11iss2pp109-118

Nurul Yusufiyah, H. K., & Gya Nur Rochman, J. P. (2021). Efektivitas Penggunaan Seleksi Ciri CFS pada Klasifikasi Ciri Bentuk Nodul Kanker Payudara dengan Citra Ultrasonografi. *Physics Education Research Journal*, *3*(1), 11–18. https://doi.org/10.21580/perj.2021.3.1.6667

Pramadhana, D. (2021). Klasifikasi Penyakit Diabetes Menggunakan Metode CFS dan ROS dengan Algoritma J48 Berbasis Adaboost. *Edumatic: Jurnal Pendidikan Informatika*, *5*(1), 89–98. https://doi.org/10.29408/edumatic.v5i1.3336

Purbasari, I. Y., Nugroho, B., & Implementasi, D. A. N. (2013). Benchmarking Algoritma Pemilihan Atribut Pada Klasifikasi Data Mining. *Snastia*, 47–54.

Ranjan, B., Sun, W., Park, J., Mishra, K., Schmidt, F., Xie, R., Alipour, F., Singhal, V., Joanito, I., Honardoost, M. A., Yong, J. M. Y., Koh, E. T., Leong, K. P., Rayan, N. A., Lim, M. G. L., & Prabhakar, S. (2021). DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nature Communications*, *12*(1), 1–12. https://doi.org/10.1038/s41467-021-26085-2

Rish, I. (2001). *An Empirical Study of The Naive Bayes Classifier*. 41–46. https://doi.org/10.1039/b104835j

Widagdo, K. A., Adi, K., & Gernowo, R. (2020). Kombinasi Feature Selection Fisher Score dan Principal Component Analysis (PCA) untuk Klasifikasi Cervix Dysplasia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *7*(3), 565. https://doi.org/10.25126/jtiik.2020702987

Zhang, S., Zhang, C., & Yang, Q. (2003). Data Prepartion for Data Mining. *Appl. Artif. Intel.*, *17(5–6)*, 375–381. https://doi.org/10.1080/08839510390219264