JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 4, Oktober 2020, Page 955-962
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i4.2370

Optimasi Klasifikasi Bayesian Network Melalui Reduksi Attribute Menggunakan Metode Principal Component Analysis

Surizar Rahmi*, Pahala Sirait, Erwin Setiawan Panjaitan

Program Studi Magister Teknologi Informasi, STMIK Mikroskil, Medan Indonesia
Email Penulis Korespondensi: surizar.rdanur@gmail.com

Abstrak−Reduksi dimensionalitas merupakan topik yang sedang hangat diperbincangkan dalam perkembangannya telah dilakukan diberbagai bidang penelitian salah satunya *machine learning* dengan melakukan reduksi dapat menurunkan kapasitas dimensi tanpa mengurangi (menghilangkan) informasi yang terkandung pada data tersebut. *Principal Component Analysis* merupakan salah satu teknik mereduksi yang telah teruji mampu mengurangi kapasitas data tanpa menghilangkan informasi yang terkandung pada dataset secara siginifikan. Pada penelitian ini dilakukan reduksi atribut menggunakan *Principal Component Analysis* dengan menggunakan dataset faktor-faktor yang mempengaruhi ketidakhadiran karyawan diambil dari *Repositori University of California di Irvine (UCI).* Kombinasi dengan *Bayesian Network* untuk mengklasifikasi data sebagai perbandingan antara sebelum dan sesudah dilakukan reduksi atribut. Hal tersebut dapat terlihat pada hasil akurasi awal sebelum dilakukan reduksi dengan akurasi sebesar 100% dan setelah dilakukan reduksi atribut ke lima terjadi penurunan akurasi sebesar 89,7%.

Kata Kunci: Reduksi, Atribut, *Principal Component Analysis*, Klasifikasi *Bayesian Network*

Abstract−Dimensionality reduction is a hot topic being discussed in its development has been carried out in various fields of research one of which is machine learning by reducing can reduce the capacity of dimensions without reducing (eliminating) information contained in the data. Principal Component Analysis is one of the proven reduction techniques capable of reducing data capacity without significantly eliminating the information contained in the dataset. In this research attribute reduction using principal component analysis using a dataset of factors affecting employee absence was taken from the University of California repository at Irvine (UCI). Combination with Bayesian Network to classify data as a comparison between before and after attribute reduction. This can be seen in the initial results before the reduction with an accuracy of 100% and after the fifth attribute reduction there is a decrease in accuracy by 89,7%.

Keywords: Reduction, Attribute, Principal Component Analysis, Bayesian Network Classification

1. PENDAHULUAN

Teknik reduksi atau reduksi dimensionalitas merupakan topik yang sedang hangat diperbincangkan dalam perkembangannya telah dilakukan diberbagai bidang penelitian seperti statistika dan *machine learning* dengan melakukan reduksi dapat menurunkan kapasitas data tanpa mengurangi informasi yang terkandung pada data tersebut[1]. Seperti pada penelitian yang telah dilakukan oleh Rehman *et al* dengan melakukan reduksi data dalam penentuan pola terhadap data operasional pada proses *data mining* sehingga dapat meminimalisir biaya-biaya yang dianggap dapat di alokasikan ke dalam biaya yang lebih penting[2]. Setelah dilakukannya reduksi pada kompleksitas data, proses pengolahan data menjadi lebih cepat dan tepat seperti yang telah dilakukan sebelumnya tentang *preprocessing* dalam mengukur waktu proses dan memori yang telah dilakukan reduksi data sehingga memberikan hasil lebih cepat dan akurat[3]. Penelitian serupa juga dilakukan oleh [4] dengan menggunakan teknik reduksi data sehingga menghasilkan peningkatan kinerja serta penghematan waktu dan bandwidth pada penggunaan *data mining*[5]. Fleksibilitas sistem dalam melayani permintaan umumnya memiliki lebih dari satu cara misal pada pendistribusian sistem penyimpanan dimana setiap server diberikan tugas untuk menyimpan data sehingga terjadi redudansi data yang mengakibatkan waktu layanan lebih panjang, seperti yang telah dilakukan dengan mengurangi latency melalui perancangan kebijakan penjadwalan dengan memberikan pengaturan karakterisasi untuk mengurangi latency[6].

Feature dalam melakukan reduksi data dapat melalui beberapa pendekatan seperti: *T-Distrib Stochastic Neib Embedding, Principle Component Analysis, Canonical Correlation Analysis, Linier Discriminant Analysis*[3]. *Principal Component Analysis* (PCA) merupakan teknik dalam meningkatkan interpretabilitas tanpa menghilangkan (mengurangi) informasi yang dibutuhkan pada dimensi dataset tersebut[7]. Teknik PCA telah banyak digunakan pada pengurangan struktur dimensi dalam pengolahan data menjadi salah satu hal penting dalam pengolahan seperti: gambar, teks, video dan pencarian web sehingga dapat meningkatkan efisiensi waktu serta pemakaian memori[8][9].

Seperti yang telah dilakukan dengan melakukan pengurangan dimensi menggunakan Tensor Robust Principal Component Analysis dalam mengembalikan data yang telah rusak sehingga meningkatkan efektifitas[10]. Dalam melakukan reduksi terdapat beberapa kendala seperti: *outlier data, noisy data, anomaly data,* dan *missing value*[11]. Pada penelitian sebelumnya yang telah dilakukan tentang menguji sensitivitas *outlier* data terhadap dimensi rendah menghasilkan pendekatan lebih efektif[12]. Pada penelitian sebelumnya yang telah dilakukan oleh Zhao Z, et al., 2016 tentang keakuratan dan efektitfitas PCA terhadap gambar 2D dalam mengurangi *noisy* data[13]. Pada penelitian sebelumnya yang telah dilakukan tentang menentukan variabel-variabel utama dalam menentukan spesifikasi enzim dengan melakukan rekayasa terhadap metabolisme dengan hasil lebih 40% setiap produksi[14].

Submitted: **25/08/2020**; Accepted: **05/09/2020**; Published: **20/10/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 4, Oktober 2020, Page 955-962
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i4.2370

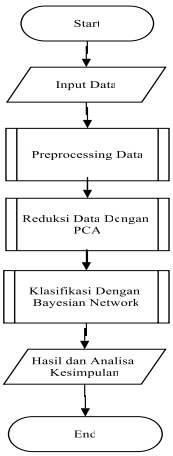
Klasifikasi merupakan proses pencarian sekumpulan model atau fungsi yang menggambarkan dan membedakan kelas data yang bertujuan untuk memprediksi dari suatu objek yang belum diketahui kelasnya[15]. Klasifikasi sendiri memiliki dua proses yaitu membangun model klasifikasi dari sekumpulan kelas data yang sudah didefinisikan sebelumnya (*training data set*) dan menggunakan model tersebut untuk klasifikasi data test serta mengukur akurasi model tersebut[16]. Pada penelitian penerapan *Bayesian Network* dalam memprediksi penyakit *Pneumonia* dan diare pada anak – anak memberikan hasil secara otomatis cukup akurat tanpa melibatkan seorang pakar sehingga dapat dilakukan pencegahan secara dini dengan mempertimbangkan penyebab penyakit tersebut [17]. *Bayesian Network* merupakan model grafis probabilitik yang mempresentasikan serangkaian variabel dan keterkaitan antar variabel dimana *Bayesian Network* dapat menampilkan probabilitas hubungan antara kejadian-kejadian yang saling berhubungan maupun tidak[18]. *Bayesian Network* mempresentasikan kebebasan kondisional dan kebergantungan antara satu variabel dengan variabel lainnya, sehingga proses klasifikasi lebih sederhana dan akurat[19]. Pada penelitian sebelumnya dilakukan tentang pemanfaatan data yang kecil serta ketidaklengkapan data dengan mengklasifikasi menggunakan *Bayesian Network* menghasilkan akurasi dan kinerja waktu yang lebih efisien[20]. Penelitian sebelumnya juga dilakukan dengan mengidentifikasi faktor-faktor penting pada turunnya kinerja karyawan menggunakan metode *Bayesian Network* dengan memilah potensi utama stress kerja sehingga mengakibatkan gangguan kerja, dengan kerterkaitan seperti; hubungan depresi dengan suasana hati, minimnya keterlibatan kerja, hubungan masalah keluarga dengan pekerjaan[21]. Penerapan Metode *Bayesian Network* juga digunakan oleh Izadi *et al*., tentang hubungan kesehatan dalam mengukur pemanfaatan pelayanan kesehatan pada pembelian obat dengan melakukan evaluasi pada pengawasan kesehatan masyarakat menghasilkan prediksi jangka pendek pada pengawasan *realtime* sehingga didapatkan hasil perkiraan yang tepat[21].

Penelitian ini bertujuan untuk menentukan variable-variable utama dengan tekhnik mereduksi data menggunakan *Principal Component Analysis* dan mengkombinasi *Bayesian Network* untuk klasfikasi sehingga dapat menghasilkan klasifikasi yang relatif optimal. Adapun langkah – langkah yang akan dilakukan adalah sebelum dilakukannya reduksi, terlebih dahulu dilakukan seleksi data, normalisasi data, selanjutnya dilakukan pemisahan data yaitu data training dan testing dimana pada data testing akan dilakukan pengujian dengan mereduksi data menggunakan *Principal Component Analysis* dan selanjutnya dilakukan klasifikasi dengan data training menggunakan *Bayesian Network* dengan hasil akhir dilakukan penilaian presisi, akurasi nilai error setelah dilakukan klasifikasi.

2. METODOLOGI PENELITIAN

2.1 Metodologi Penelitian

Metodologi penelitian merupakan tahapan-tahapan yang sistematis dilakukan pada penelitian ini sehingga penelitian ini terarah dengan baik. Adapun gambaran metode penelitian ini dapat dilihat pada gambar dibawah ini:



Gambar 1. Metodologi Penelitian

Pada penelitian yang dilakukan kali ini akan fokus untuk menyelesaikan 3 (tiga) masalah secara bersamaan, yaitu :

1. Masalah pertama yang ingin diselesaikan pada penelitian ini tentang preprocessing data.
2. Masalah kedua yang ingin diselesaikan adalah bagaimana mereduksi atribut yang ada tanpa mengurangi atau meminimalkan hasil informasi.
3. Masalah ketiga yang ingin diselesaikan adalah klasifikasi dataset dengan metode *Bayesian Network*.

2.2 Reduksi Data

Submitted: **25/08/2020**; Accepted: **05/09/2020**; Published: **20/10/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 4, Oktober 2020, Page 955-962
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i4.2370

Reduksi data merupakan langkah penting pada preprocessing data mining dengan tujuan memperoleh keakuratan hasil, kecepatan, dan kemudahan dalam beradaptasi pada kompleksitas data yang ditandai dengan kecepatan objek dalam merespon perubahan data[21] Reduksi data merupakan proses analisis untuk memilih, memusatkan perhatian, penyederhanaan abstraksi serta mentransformasi data yang diperoleh dari catatan – catatan lapangan dengan tujuan untuk lebih mudah dipahami dan tetap mempertahankan informasi yang terkandung dengan berfokus pada hal-hal yang dianggap penting, penentuan pola dan mencari tema serta membuang data yang dianggap tidak penting[21]. Pada perusahaan penerapan reduksi data dilakukan berbasis data mining dan machine learning untuk mencapai berbagai tujuan dalam menciptakan value yang berbeda

2.3 Metode Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah sebuah teknik menganalisa pada sebuah table data observasi ke dalam sebuah data baru yang memiliki kemiripan korelasi dengan tujuan untuk menyederhanakan data observasi yang sebelumnya kompleks agar menjadi lebih sederhana sehingga mudah untuk di proses atau dianalisis. P*principal component analysis* adalah teknik statistic yang secara linier mengubah sekumpulan bentuk variable asli menjadi variable yang lebih kecil atau sederhana yang tidak berkorelasi yang dapat mewakili informasi dari sekumpulan variable asli[13][18]

Kompleksitas data semakin umum ditemukan dan sulit dalam pengolahannya, dalam meningkatkan interpretabilitas tanpa menghilangkan informasi yang terkandung didalam data membutuhkan reduksi dimensi dalam menentukan variabel-variabel utama. Berikut langkah – langkah pengerjaan dalam mereduksi atribut sebagai berikut :

a. Pembentukan Matriks
Tahapan awal dari ekstraksi PCA adalah pembentukan matriks. Data dipresentasikan dalam ukuran matrik mxn, dimana m adalah jumlah citra yang dilatih dan n merupakan dimensi dari citra tersebut, kemudian dilakukan proses ekstraksi menjadi citra dengan dimensi yang lebih kecil yang hasilnya diproyeksikan menjadi sebuah matriks seperti pada persamaan berikut:

$$X = \begin{vmatrix} X_{11} & X_{12} & X_{1j} & X_{1N} \\ X_{21} & \dots & \dots & X_{2N} \\ X_{31} & \dots & \dots & X_{3N} \\ X_{i1} & \dots & \dots & X_{iN} \\ X_{M1} & X_{M2} & X_{Mj} & X_{MN} \end{vmatrix}$$

Dimana x = matrik citra. Setelah matrik data citra terbentuk, maka proses berikutnya adalah proses perhitungan untuk mencari rata-rata hasil seluruh citra. Pencarian nilai rata-rata ini bertujuan untuk mengetahui noise atau persamaan tiap vektor yang dapat mengganggu keakuratan perhitungan pada PCA, yang dapat dihitung dengan menggunakan rumus:

$$[A - \lambda I]\,[X] = [0] \qquad (1)$$

b. Menghitung Matrik Kovarian
Input vektor x_t (t=1, ⋯,1 dan $x_t = 0$) dengan dimensi m $x_t = [x_t(1), x_t(2), \cdots, x_t(m)]^T$ biasanya m < 1, setiap vektor x_t ditransformasikan secara linier kedalam satu vektor baru *s* yang dinyatakan sebagai berikut :

$$S_t = U^T . x_i \qquad (2)$$

Dimana *U* adalah matrik orthogonal m x m dengan kolom ke i, u_i adalah nilai eigenvector dari sampel matrik kovarian

$$C = \frac{1}{l} \sum_{t=1} l \; Xt . x_t^T \qquad (3)$$

c. Menghitung Eigenvalue dan Eigenvector Dari Matrik Kovarian

$$\lambda_t U_t = C.u_i \,, i=1, \cdots m \qquad (4)$$

Dimana *i* adalah salah satu eigenvalue dari C, u_i adalah nilai eigenvector.

d. Menghitung nilai transformasi orthogonal
Berdasarkan nilai estimasi u_i dan komponen s_t, yang kemudian dihitung sebagai transformasi orthogonal dari x_t

$$S_t\,(i) = u_i^T\, x_t \,, \text{i} = 1, \ldots., \text{m} \qquad (5)$$

Komponen yang baru tersebut disebut dengan principal component. Dengan menggunakan hanya beberapa nilai pertama eigenvector yang telah diurutkan berdasarkan nilai eigennya, jumlah principal component dari s_t (*i*) yang tidak saling terkorelasi, mempunyai varian maksimum yang berurutan dan estimasi error rata-rata dari representasi data input asli adalah minimal.

2.4 Bayesian Network

Bayesian Network didasarkan pada Teorema Bayes yaitu conditional probability (peluang bersyarat) yang dinotasikan dengan P(A|B) artinya peluang keadaan A jika keadaan B telah terjadi. Berbeda dari naive bayes yang mengabaikan hubungan antar atribut atau variabel, pada *Bayesian Network* antar variabel atau atribut bisa saling dependent atau berhubungan, berikut Rumus Teorema Bayes yaitu:

$$P\,(A|\text{B}) = \frac{P(B|A)P(A)}{P(B)} \qquad (6)$$

Submitted: **25/08/2020**; Accepted: **05/09/2020**; Published: **20/10/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 4, Oktober 2020, Page 955-962
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i4.2370

Atau

$$P(A|B) = \frac{P(A)P(B|A)}{P(B|A)P(A)+P(B|\hat{A})P(\hat{A})} \quad (7)$$

Dimana:

$P(A|B)$ = disebut posterior probability, yaitu peluangA terjadi setelah B terjadi
$P(A \cap B)$ = peluang B dan A terjadi bersamaan
$P(B|A)$ = disebut juga likehood, yaitu peluang B terjadi setelah A terjadi
$P(A)$ = disebut juga prior, yaitu peluang kejadian A
$P(B)$ = peluang kejadian B

Adapun langkah-langkah untuk menerapkan *Bayesian Network* yaitu:

1. Membangun struktur *Bayesian Network*
2. Menentukan parameter
3. Membuat *Conditional Probability Table* (CPT)
4.Membuat *Joint Probability Distribution* (JPD), untuk menghitung Joint Probability Distribution adalah mengalikan nilai *Conditional Probability* dengan *Prior Probability*.
5. Menghitung *Posterior Probabilistik*, didapatkan dari hasil JPD yang telah diperoleh.
6. Inferensi Probabilistik yaitu penelusuran yang dilakukan berdasarkan variabel input yang diberikan pengguna sehingga menghasilkan suatu nilai probabilitas.

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan Dataset yang digunakan adalah dataset faktor yang mempengaruhi ketidakhadiran karyawan diambil dari *repositori University of California di Irvine (UCI)*[1] digunakan pada penelitian akademik di Universidade Nove de Julho' Program Pascasarjana di Informatika dan Manajemen Pengetahuan. Data tersebut telah digunakan pada penelitian – penelitian sebelum dalam memprediksi ketidakhadiran karyawan. Pada dataset memiliki 740 data dan 21 attribut dengan target kelas dalam satuan jam. Adapun kriteria 21 attribut tersebut dapat dilihat pada table 1.

Tabel 1. Kriteria Attribut Dataset

No	Nama Attribut	Rentang (Range)	Keterangan
1	ID	1 sampai 36	Individual Identification
2	Reason For Absence	1 sampai 28	Terdapat 21 kategori penyakit dan 7 kategori tanpa CID (Code of Diseases)
3	Month Of Absence	Bulan 1 sampai bulan 12	Setahun terdapat 12 bulan
4	Day Of the week	1 hari sampai 5 hari	Senin, selasa, rabu, kamis, dan jumat
5	Seasons	1 musim sampai 4 musim	Terdapat 4 musim yaitu; musim panas, musim semi, musim gugur, dan musim dingin
6	Transportation expense	118 sampai 388	Biaya transportasi
7	Distance from residence to work	5 km sampai 52 km	Satuan kilometer
8	Service time	1 menit sampai 29 menit	Waktu pelayanan
9	Age	27 tahun sampai 58 tahun	Umur karyawan
10	Work load average/ day	205.917 menit sampai 378.884 menit	Rata – rata waktu kerja
11	Hit target	81 target sampai 100 target	Pencapaian target
12	Disciplinary failure	0 dan 1	Disiplin kerja
13	Education	1 sampai 4	High school, graduate, postgraduate, master and doctor
14	Son	0 sampai 4 anak	Jumlah anak
15	Social drinker	0 dan 1	Peminum
16	Social smoker	0 dan 1	Perokok
17	Pet	0 sampai 8 hewan	Hewan peliharaan
18	Weight	56kg sampai 108kg	Berat badan karyawan

Submitted: **25/08/2020**; Accepted: **05/09/2020**; Published: **20/10/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 4, Oktober 2020, Page 955-962
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i4.2370

No	Nama Attribut	Rentang (Range)	Keterangan
19	Height	163cm sampai 196cm	Tinggi badan karyawan
20	Body mass index	19kg/m^2 sampai 38 kg/m^2	Index masa tubuh karyawan
21	Absenteeism time in hours (target)	0 sampai 120 jam	Klasifikasi karyawan

Sebelum memasuki tahapan utama dilakukan normalisasi data agar data yang digunakan memiliki rentang jarak suatu nilai yang telah diatur sehingga memudahkan proses pengolahan data. Pada tahap normalisasi data penulis menggunakan standar *multiple regression* dengan perhitungan rumus pencarian nilai minimun dan maksimum.

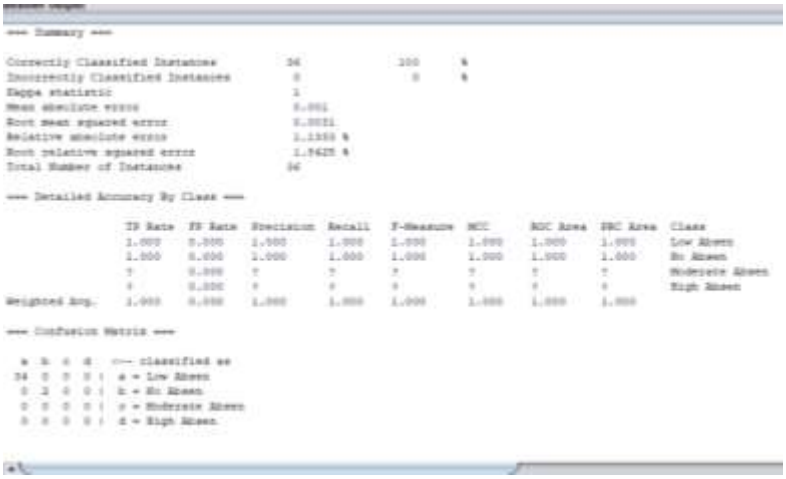
Tabel 2. Data Hasil Normalisasi

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16	N17	N18	N19
0,93	0,25	0,11	0,04	10,32	1,29	0,46	1,18	8555,50	3,46	0,00	0,04	0,07	0,04	0,00	0,04	3,21	6,14	1,07
0,00	0,25	0,11	0,04	4,21	0,46	0,64	1,79	8555,50	3,46	0,04	0,04	0,04	0,04	0,00	0,00	3,50	6,36	1,11
0,82	0,25	0,14	0,04	6,39	1,82	0,64	1,36	8555,50	3,46	0,00	0,04	0,00	0,04	0,00	0,00	3,18	6,07	1,11
0,25	0,25	0,18	0,04	9,96	0,18	0,50	1,39	8555,50	3,46	0,00	0,04	0,07	0,04	0,04	0,00	2,43	6,00	0,86
0,82	0,25	0,18	0,04	10,32	1,29	0,46	1,18	8555,50	3,46	0,00	0,04	0,07	0,04	0,00	0,04	3,21	6,14	1,07
0,82	0,25	0,21	0,04	6,39	1,82	0,64	1,36	8555,50	3,46	0,00	0,04	0,00	0,04	0,00	0,00	3,18	6,07	1,11
0,79	0,25	0,21	0,04	12,89	1,86	0,11	1,00	8555,50	3,46	0,00	0,04	0,04	0,04	0,00	0,14	2,86	6,14	0,96

3.1 Hasil Pengujian

Berikut ini merupakan hasil yang didapatkan pada penelitian yang dilakukan dengan mengkombinasi metode *Principal Component Analysis* dengan *Bayesian Network*. Hasil yang didapatkan pada pengujian berdasarkan menggunakan dataset faktor yang mempengaruhi ketidakhadiran karyawan dimana terdapat 21 attribut sudah termasuk target kelas dan 740 baris. Berikut tahapan dalam pengujian hasil analisis pada faktor – faktor yang mempengaruhi ketidakhadiran karyawan. Reduksi atribut berfungsi untuk mengurangi dimensi data (kapasitas) tanpa mempengaruhi nilai hasil. Pada penelitian reduksi dilakukan menggunakan aplikasi Weka versi 3.8. Pada penelitian ini penulis membatasi hanya mereduksi 5 atribut dari 19 atribut tanpa atribut target yaitu atribut *Absenteesim time in hours*. Pengujian dilakukan dengan membagi dataset menjadi 2 bagian yaitu: data training dan data testing. Pada data training digunakan untuk pengujian klasfikasi dan data testing digunakan untuk pengujian reduksi.

Pada pengujian reduksi 1 atribut menggunakan data testing sebesar 5% dengan menggunakan data sebesar 37 instances dan terdapat 19 atribut yang akan di uji. Sebelum dilakukan reduksi pada atribut dataset faktor yang mempengaruhi ketidakhadiran karyawan terlebih dahulu dilakukan klasifikasi terhadap data tersebut untuk mengetahui tingkat akurasi yang dimiliki pada dataset. Hasil klasifikasi pada dataset pertama dengan data testing sebesar 5% dapat dilihat pada gambar dibawah ini.



Gambar 2. Hasil Akurasi Klasifikasi

Tabel 3. Hasil Klasifikasi Dataset Sebalum Reduksi Pengujian 5%

Confusion matrix	Low absen	No absen	Moderate absen	High absen
Low absen	34	0	0	0

Submitted: **25/08/2020**; Accepted: **05/09/2020**; Published: **20/10/2020**

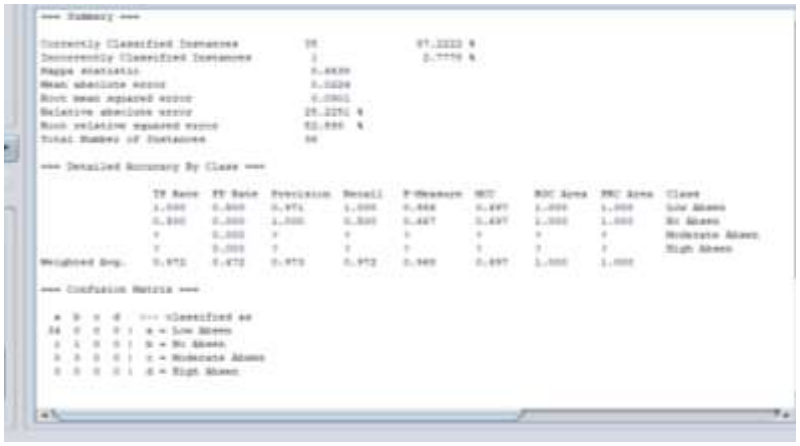
JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 4, Oktober 2020, Page 955-962
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i4.2370

Confusion matrix	Low absen	No absen	Moderate absen	High absen
No absen	0	2	0	0
Moderate absen	0	0	0	0
High absen	0	0	0	0
Hasil	34	2	0	0
Presisi	1.00			
Recall	1.00			
Akurasi	100%			

Selanjutnya dilakukan reduksi satu atribut. Berikut tampilan reduksi 1 atribut terlihat pada gambar dibawah.

Gambar 3. Reduksi 1 Atribut

Pada gambar diatas dapat dilihat atribut *work load average day* merupakan atribut dengan nilai eigen value terendah sehingga reduksi atribut dapat dilakukan dengan pencarian bobot terendah berdasarkan eigen value. Berikut hasil klasifikasi setelah dilakukan reduksi satu atribut.



Gambar 4. Klasifikasi Reduksi 1 Atribut

Proses reduksi dan juga klasifikasi dilakukan berulang hingga sampai proses reduksi atribut terakhir selesai dengan menggunakan data teseting yang berbeda – beda. Data testing yang digunakan sebesar 5%, 10%, 15% dan 20%. Setelah selesai proses reduksi atribut dan juga proses klasifikasi, maka proses selanjutnya adalah menampilkan seluruh hasil proses pengujian dataset. Sebelum dilakukan tahap reduksi didapatkan hasil akurasi sebesar 100% pada pengujian data testing sebesar 5% dan 10%, sedangkan pada pengujian data testing 15% nilai akurasi sebesar 98,1% dan data testing 20% nilai akurasi sebesar 98,6%. Sebelum dilakukan pengujian data training memiliki akurasi sebesar 100% dan setelah dikukan pengujian didapat nilai akurasi paling rendah sebesar 89,7%.

3.2 Pembahasan

Pada bagian ini, akan dilakukan evaluasi terhadap hasil pengujian yang telah dilakukan pada dataset faktor – faktor ketidakhadiran karayawan dengan dilakukan pengambilan data sampel sebanyak empat tahap, yaitu tahap pertama pengujian dilakukan dengan mengambil sampel data sebanyak 5% dari dataset dan dilakukan pengujian reduksi sebanyak 4 kali pengujian terhadap sampel, selanjutnya sampel yang digunakan sebanyak 10% dengan 4 kali pengujian dilakukan terhadap data, selanjutnya pengujian data sampel digunakan sebanyak 15% dan dilakukan 4 kali pengujian dan terakhir pengujian menggunakan data sampel sebanyak 20% dari dataset dan dilakukan sebanyak 4 kali pengujian terhadap data.

Submitted: **25/08/2020**; Accepted: **05/09/2020**; Published: **20/10/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 4, Oktober 2020, Page 955-962
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i4.2370

Pengujian dataset faktor ketidakhadiran karyawan pada tahap awal menggunakan data testing sebesar 5% dari jumlah keselurahan data training yaitu sebesar 36 data. Pada pengujian testing dilakukan reduksi sebanyak 4 kali, dengan atribut awal sebesar 19 atribut sehingga menyisakan atribut sebesar 15 atribut. Pada pengujian data sebesar 5% atribut yang direduksi adalah *work load average day*, *weight*, *distance from residence to work*, dan *transportation expense*. Setelah dilakukan reduksi tahap awala selanjutnya dilakkukan reduksi tahap dua dengan menggunakan dataset sebesar 10% atribut pada atribut yang direduksi pada tahap dua adalah *work load average day*, *distance from to work*, *transportation expense* dan *weigth*. Selanjutnya pada tahap reduksi tiga menggunakan data testing sebesar 15% pada tahap pengujian, didapatkan hasil atribut yang dierduksi adalah *work load average day*, *transportation expense*, *distance from residence to work* dan *weight.* Selanjutnya dilakukan reduksi tahap empat dengan menggunakan data testing sebesar 20%, didapatkan hasil atribut yang direduksi adalah *work load average day*, *transportation expense*, *distance from residence to work* dan *weight*. Penilaian reduksi atribut berdasarkan *eigen value*, *proportion* dan *cumulative.* Dari penjelasan diatas dapat diperoleh secara garis besar atribut yang merupakan variabel yang memiliki pengaruh adalah *reason for absence*, *month of absence*, *day of the week*, *seasons*, *service time*, *age*, *hit target*, *disciplinary failure*, *education*, *son*, *social drinker*, *social smoker*, *pet*, *height*, dan *body mass index.*

Pengujian klasifikasi pada dataset faktor ketidakhadiran karyawan dengan mengklasifikasi atribut dengan data testing sebesar 5%, selanjutnya dilakuakn klasifikasi kembali dengan menggunakan data testing sebesar 10%, pengujian kembali dilakukan dengan mengguanakan data testing sebesar 15% dan pegnujian terakhir menggunakan data testing sebesar 20%. Pada hasil pengujian klasifikasi diperoleh nilai akurasi sebagai berikut:

Tabel. 4. Hasil Klasifikasi Dataset 5%

Nomor	Data testing	Tingkat rata-rata akurasi
1	Reduksi 0 atribut	100%
2	Reduksi 1 atribut	97,2%
3	Reduksi 3 atribut	94,4%
4	Reduksi 3 atribut	94,4%
5	Reduksi 4 atribut	94,4%

Pada pengujian dengan menggunakan data testing sebesar 5% hasil akurasi terjadi penurunan sebesar rata-rata 2,6%. Selanjutnya untuk hasil pengujian klasifikasi faktor ketidakhadiran karyawan dengan data testing sebesar 10% adalah sebagai berikut:

Tabel 5. Hasil Klasifikasi Dataset 10%

Nomor	Data testing	Tingkat rata-rata akurasi
1	Reduksi 0 atribut	100%
2	Reduksi 1 atribut	98,6%
3	Reduksi 3 atribut	98,6%
4	Reduksi 3 atribut	95,3%
5	Reduksi 4 atribut	91,5%

Pada pengujian dengan menggunakan data testing sebesar 5% hasil akurasi terjadi penurunan sebesar rata-rata 3,2%. Selanjutnya untuk hasil pengujian klasifikasi faktor ketidakhadiran karyawan dengan data testing sebesar 15% adalah sebagai berikut:

Tabel 6. Hasil Klasifikasi Dataset 15%

Nomor	Data testing	Tingkat rata-rata akurasi
1	Reduksi 0 atribut	98,1%
2	Reduksi 1 atribut	95,4%
3	Reduksi 3 atribut	94,4%
4	Reduksi 3 atribut	90,6%
5	Reduksi 4 atribut	89,7%

Pada pengujian dengan menggunakan data testing sebesar 5% hasil akurasi terjadi penurunan sebesar rata-rata 2,06%. Selanjutnya untuk hasil pengujian klasifikasi faktor ketidakhadiran karyawan dengan data testing sebesar 20% adalah sebagai berikut:

Tabel 7. Hasil Klasifikasi Dataset 20%

Nomor	Data testing	Tingkat rata-rata akurasi
1	Reduksi 0 atribut	98,6%
2	Reduksi 1 atribut	96,5%
3	Reduksi 3 atribut	95,8%
4	Reduksi 3 atribut	93,1%
5	Reduksi 4 atribut	91,7%

Submitted: **25/08/2020**; Accepted: **05/09/2020**; Published: **20/10/2020**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 4, Nomor 4, Oktober 2020, Page 955-962
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v4i4.2370

Pada pengujian dengan menggunakan data testing sebesar 5% hasil akurasi terjadi penurunan sebesar rata-rata 1,66%.

4. KESIMPULAN

Berdasarkan dari penjelasan permasalah serta hasil dari penelitian maka dapat ditarik kesimpulan:

1. Metode Principal Component Analysis (PCA) dapat digunakan untuk menentukan peringkat didalam reduksi atribut, untuk mengetahui atribut mana saja yang memberikan pengaruh terhadap dataset sehingga dapat mereduksi data yang tidak memiliki pengaruh besar terhadap dataset.
2. Proses klasifikasi menggunakan *Bayesian Network* menghasilkan akurasi yang berbeda pada data yang telah direduksi dengan nilai akurasi awal sebesar 100% menurun menjadi 89,7% dengan atribut rata-rata yang direduksi yaitu: *work load average day*, *distance from residence to work*, *transportation expense*, dan *weight* sehingga dapat disimpulkan dengan mengkombinasi metode Principal Component Analysis dengan *Bayesian Network* menghasilkan klasifikasi relative optimal pada faktor yang mempengaruhi ketidakhadiran karyawan.

REFERENCES

[1] Y. Luo, K. Li, Y. Li, D. Cai, C. Zhao, and Q. Meng, "Three-Layer Bayesian Network for Classification of Complex Power Quality Disturbances," *IEEE Trans. Ind. Informatics*, vol. 14, no. 9, pp. 3997–4006, 2018, doi: 10.1109/TII.2017.2785321.

[2] M. Habib, V. Chang, A. Batool, and T. Ying, "International Journal of Information Management Big data reduction framework for value creation in sustainable enterprises," *Int. J. Inf. Manage.*, vol. 36, no. 6, pp. 917–928, 2016, doi: 10.1016/j.ijinfomgt.2016.05.013.

[3] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, 2017, doi: 10.1016/j.neucom.2017.01.078.

[4] A. A. Yildirim, C. Özdoğan, and D. Watson, "Parallel data reduction techniques for big datasets," *Big Data Manag. Technol. Appl.*, no. December 2015, pp. 72–93, 2013, doi: 10.4018/978-1-4666-4699-5.ch004.

[5] L. Shiyue, Y. Dong, D. Song, and Z. Liping, "Data filtering algorithm based on attribute reduction and gene expression programming," *2018 IEEE 3rd Int. Conf. Big Data Anal. ICBDA 2018*, pp. 248–253, 2018, doi: 10.1109/ICBDA.2018.8367686.

[6] N. B. Shah, K. Lee, and K. Ramchandran, "When Do Redundant Requests Reduce Latency?," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 715–722, 2016, doi: 10.1109/TCOMM.2015.2506161.

[7] I. T. Jollife and J. Cadima, "Principal component analysis: A review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, 2016, doi: 10.1098/rsta.2015.0202.

[8] K. J. Galinsky *et al.*, "Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia," *Am. J. Hum. Genet.*, vol. 98, no. 3, pp. 456–472, 2016, doi: 10.1016/j.ajhg.2015.12.022.

[9] Y. Aït-Sahalia and D. Xiu, "Principal Component Analysis of High-Frequency Data," *J. Am. Stat. Assoc.*, vol. 114, no. 525, pp. 287–303, 2019, doi: 10.1080/01621459.2017.1401542.

[10] C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan, "Lu_Tensor_Robust_Principal_CVPR_2016_paper.pdf," pp. 5249–5257.

[11] T. Metsalu and J. Vilo, "ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W566–W570, 2015, doi: 10.1093/nar/gkv468.

[12] S. Yi, Z. Lai, Z. He, Y. ming Cheung, and Y. Liu, "Joint sparse principal component analysis," *Pattern Recognit.*, vol. 61, pp. 524–536, 2017, doi: 10.1016/j.patcog.2016.08.025.

[13] Z. Zhao, Y. Shkolnisky, and A. Singer, "Fast Steerable Principal Component Analysis," *IEEE Trans. Comput. Imaging*, vol. 2, no. 1, pp. 1–12, 2016, doi: 10.1109/tci.2016.2514700.

[14] J. Alonso-Gutierrez *et al.*, "Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering," *Metab. Eng.*, vol. 28, pp. 123–133, 2015, doi: 10.1016/j.ymben.2014.11.011.

[15] S. H. Wang *et al.*, "Multiple Sclerosis Detection Based on Biorthogonal Wavelet Transform, RBF Kernel Principal Component Analysis, and Logistic Regression," *IEEE Access*, vol. 4, pp. 7567–7576, 2016, doi: 10.1109/ACCESS.2016.2620996.

[16] C. A. Magee, P. Caputi, and J. K. Lee, "Distinct longitudinal patterns of absenteeism and their antecedents in full-time australian employees," *J. Occup. Health Psychol.*, vol. 21, no. 1, pp. 24–36, 2016, doi: 10.1037/a0039138.

[17] A. C. Constantinou, N. Fenton, W. Marsh, and L. Radlinski, "From complex questionnaire and interviewing data to intelligent Bayesian network models for medical decision support," *Artif. Intell. Med.*, vol. 67, pp. 75–93, 2016, doi: 10.1016/j.artmed.2016.01.002.

[18] S. C. Ng, "Principal component analysis to reduce dimension on digital image," *Procedia Comput. Sci.*, vol. 111, pp. 113–119, 2017, doi: 10.1016/j.procs.2017.06.017.

[19] D. Ballabio, R. Todeschini, and V. Consonni, *Recent Advances in High-Level Fusion Methods to Classify Multiple Analytical Chemical Data*, vol. 31. Elsevier, 2019.

[20] Y. You, J. Li, and N. Xu, "A constrained parameter evolutionary learning algorithm for Bayesian network under incomplete and small data," *Chinese Control Conf. CCC*, pp. 3044–3051, 2017, doi: 10.23919/ChiCC.2017.8027825.

[21] J. Lee, R. Henning, and M. Cherniack, "Correction workers' burnout and outcomes: A bayesian network approach," *Int. J. Environ. Res. Public Health*, vol. 16, no. 2, 2019, doi: 10.3390/ijerph16020282.

Submitted: **25/08/2020**; Accepted: **05/09/2020**; Published: **20/10/2020**