

Penerapan Metode *Principal Component Analysis* (PCA) Pada Klasifikasi Status Kredit Nasabah Bank Sumsel Babel Cabang KM 12 Palembang Menggunakan Metode *Decision Tree*

Dewi Sartika¹, Imelda Saluza²
¹ Teknik Informatika, ² Manajemen Informatika
^{1,2} Universitas Indo Global Mandiri
Palembang, Indonesia
email : dewi.sartika@uigm.ac.id

Abstrak— Bank Sumsel Babel merupakan penyedia jasa layanan keuangan daerah yang berwenang dalam menghimpun dana dari masyarakat dan menyalurkannya kembali kepada masyarakat dalam bentuk pinjaman modal kerja dengan tujuan untuk meningkatkan taraf hidup masyarakat umum. Produk yang disediakan oleh Bank Sumsel Babel diantaranya tabungan, deposito, pinjaman dan lain sebagainya. Pemberian pinjaman pada suatu nasabah dilakukan berdasarkan beberapa variabel, salah satunya adalah melihat status pembayaran calon nasabah pada proses peminjaman sebelumnya. Oleh karena itu, pada penelitian ini akan dilakukan penerapan metode *Principal Component Analysis* (PCA) dalam menentukan status kredit nasabah Bank Sumsel Babel Cabang KM 12 Palembang menggunakan metode *Decision Tree*. Hasil dari penelitian ini akan diperoleh pengaruh metode *pre-processing* PCA pada metode klasifikasi *Decision Tree* dalam penentuan status kredit nasabah. Analisis dilakukan dengan memperhatikan tingkat akurasi serta nilai *precision*, *recall* dan *f1-score* yang dihasilkan. Berdasarkan analisis yang dilakukan diperoleh bahwa metode PCA memberikan dampak baik terhadap kinerja dari metode *Decision Tree* dalam klasifikasi, hal tersebut terlihat dengan adanya peningkatan tingkat akurasi sebesar 10% dengan model klasifikasi dengan kinerja lebih baik.

Kata Kunci— *Principal Component Analysis*, *Pre-Processing*, *Decision Tree*, *precision*, *recall*, *f1-score*

I. PENDAHULUAN

Berdasarkan Undang – Undang Republik Indonesia Nomor 10 Tahun 1998, Bank merupakan badan usaha yang menghimpun dana dari masyarakat dalam bentuk simpanan dan menyalurkannya Kembali kepada masyarakat dalam bentuk kredit dan atau bentuk – bentuk lainnya dengan tujuan untuk meningkatkan taraf hidup orang banyak. Bank Sumsel Babel merupakan Bank pembangunan daerah yang telah didirikan sejak 6 November 1957. Bank Sumsel Babel menyediakan produk layanan jasa keuangan diantaranya tabungan, deposito, giro, kredit/pinjaman dan lain sebagainya.

Proses pemberian pinjaman kepada nasabah dilakukan dengan tahapan pengisian formulir disertai dengan kelengkapan berkas. Selanjutnya berdasarkan data yang diberikan oleh nasabah pihak Bank akan menganalisis kemungkinan persetujuan pinjaman yang diajukan. Salah satu faktor yang menjadi bahan pertimbangan pihak Bank dalam persetujuan pengajuan pinjaman adalah reputasi nasabah dalam proses pembayaran pada pinjaman lain atau sebelumnya. Berdasarkan status kredit nasabah, pihak Bank

dapat melihat dan memastikan cara nasabah melakukan pembayaran cicilan pinjaman sampai lunas.

Pre-processing merupakan tahapan yang dilakukan pada sekumpulan data yang akan digunakan sebagai data latih sebelum proses mining dilakukan, baik itu untuk pengklasifikasian maupun klusterisasi. Bentuk dari *pre-processing* dapat berupa *data cleaning*, *feature scaling*, maupun *data reduction*. *Data cleaning* merupakan bentuk *pre-processing* yang bertujuan untuk membersihkan dan melengkapi data dari *missing value*, *error* maupun yang tidak konsisten; *feature scaling* dilakukan untuk memastikan bahwa data yang digunakan skala yang sama untuk setiap variabel; sedangkan *data reduction* dilakukan untuk menyesuaikan data ke dalam pola, focus, kategori, atau pokok permasalahan tertentu [1]. *Principal Component Analysis* (PCA) merupakan metode yang digunakan untuk proses *pre-processing* yang bertujuan untuk melakukan *feature scaling* maupun *data reduction*. Metode PCA memanfaatkan perhitungan statistik dengan menggunakan persamaan *eigen* dan *kovarian*. Prinsip kerja PCA adalah mengubah sejumlah data dengan atribut yang saling berkorelasi menjadi data baru dengan atribut yang saling bebas [2].

Decision Tree merupakan suatu metode yang memiliki struktur menyerupai pohon yang memiliki sebuah *root node* yang digunakan untuk mengumpulkan data, *inner node* yang berada pada *root node* yang berisikan rule/pertanyaan tentang data dan *leaf node* yang digunakan dalam membuat keputusan [3]. Algoritma ini dikenal memiliki tingkat akurasi yang baik dan melakukan pemilihan fitur secara fleksibel. Pada penelitian ini akan dilakukan penerapan metode *pre-processing* PCA pada klasifikasi status kredit nasabah pada Bank Sumsel Babel Cabang KM 12 Palembang menggunakan metode *Decision Tree*. Hasil klasifikasi akan dimanfaatkan oleh pihak yang berwenang dalam penentuan ajuan kredit baru.

II. TINJAUAN PUSTAKA

A. Data Mining

Data mining merupakan penggabungan dari beberapa disiplin ilmu yang dimanfaatkan untuk menggali informasi dari sekumpulan data. Informasi yang diperoleh selanjutnya dapat dimanfaatkan dalam proses penarikan kesimpulan, manajemen informasi, pengendalian proses dan sebagainya

[4]. Data mining digunakan sebagai *Knowledge Discovery from Database* terbagi menjadi [5] :

1. Karakterisasi
2. Diskriminasi
3. Asosiasi
4. Klasifikasi
5. Prediksi
6. Pengelompokkan
7. Analisis outlier
8. Analisis Evolusi dan Deviasi

B. Klasifikasi

Klasifikasi merupakan teknik data mining klasik yang didasarkan pada pembelajaran mesin. Tujuan dari klasifikasi adalah memberikan label/kelas pada suatu data yang belum diketahui berdasarkan sekumpulan data yang telah diketahui label/kelasnya. Teknik yang umum digunakan dalam klasifikasi diantaranya [6] :

1. Regresi
2. Perhitungan Jarak
3. Pohon Keputusan
4. Rules
5. Jaringan Syaraf Tiruan

Klasifikasi terdiri dari 2 tahapan [7], yaitu :

1. Membangun Model : tahapan ini dilakukan dengan memanfaatkan data latih yang telah diketahui label/kelas-nya untuk diproses menggunakan algoritma klasifikasi membentuk model klasifikasi
2. Penggunaan Model untuk Pengklasifikasian : tahapan ini dilakukan dengan memanfaatkan model klasifikasi yang telah terbentuk untuk menguji sekumpulan data yang belum diketahui label/kelasnya

C. Pre-processing

Pre-processing data merupakan teknik yang bertujuan untuk menganalisis sekumpulan data menjadi data yang berkualitas. *Pre-processing* terdiri dari pengumpulan data, integrasi data, transformasi data, *data cleaning*, reduksi data dan diskritisasi data. Tiga aspek pentingnya perlunya *pre-processing* karena [8] :

1. Data umumnya tidak lengkap, noisy dan tidak konsisten
2. Merubah data menjadi berukuran lebih kecil
3. Data berkualitas menghasilkan model yang berkualitas

D. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) merupakan salah satu metode yang digunakan untuk mereduksi dimensi data pada tahap *Pre-processing* [9]. PCA termasuk salah satu metode ekstraksi fitur yang biasanya digunakan dalam data kontinu. Prinsip kerja metode ini adalah dengan mengekstraksi atribut sehingga menyisakan atribut yang bertujuan untuk memperoleh hasil lebih optimal. Metode ini terdiri dari 4 tahapan yaitu :

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

$$X^*_{ij} = X_{ij} - \bar{X} \quad (1)$$

Dimana :

X_{ij} = Elemen Matrik X

X^*_{ij} = Elemen Matrik X^*

\bar{X} = nilai rata – rata matrik X

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

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

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

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

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

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

E. Decision Tree

Decision Tree merupakan algoritma *supervised learning* yang umum digunakan untuk klasifikasi dan regresi. *Decision Tree* merupakan model yang strukturnya menyerupai pohon, yaitu terdiri dari node teratas yang disebut root, node selanjutnya merepresentasikan atribut sedangkan cabang merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas/label/keputusan [3]. Cara kerja algoritma *Decision Tree* yaitu :

1. Menentukan *root* dari struktur pohon dengan menghitung *information gain* dari masing – masing atribut. Atribut yang memiliki nilai *information gain* terbesar yang menjadi root pada struktur pohon.

$$Gain(S, A) = \sum_{v \in v(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (5)$$

Dimana :

Gain(S,A) : information gain himpunan S dengan Atribut A

$V(A)$: range dari atribut A

S_v : subset himpunan S dengan nilai atribut v

2. Untuk menghitung nilai dari *information gain* tiap atribut, perlu dilakukan perhitungan *entropy* terlebih dahulu. Nilai *entropy* digunakan untuk menentukan seberapa informatif suatu atribut

$$Entropy(S) = \sum_{i=1}^c P_i \log 2^{P_i} \quad (6)$$

Dimana :

S : Himpunan Kasus

P_i : Rasio dari sejumlah himpunan kasus

i : Nilai dari atribut

c : partisi/state

3. Ulangi langkah 2 hingga semua record dari dataset terpartisi. Proses dihentikan saat :
 - a. Semua record simpel mendapat kelas yang sama
 - b. Tidak ada atribut dalam record yang bisa dipartisi lagi
 - c. Tidak ada record di dalam cabang yang kosong

F. Confusion Matrix

Confusion Matrix merupakan perhitungan yang digunakan untuk merepresentasikan tingkat akurasi dari suatu klasifikasi. Dasar perhitungan dari *confusion matrix* dengan membandingkan total data yang diklasifikasikan pada kelas yang benar dengan total seluruh data yang ada pada *matrix*.

TABEL 1 CONFUSION MATRIX

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

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (7)$$

Dimana :

- TN : Model memprediksi data ada di kelas negatif dan yang sebenarnya data memang ada di kelas negatif
- TP : Model memprediksi data ada di kelas positif dan yang sebenarnya data memang ada di kelas positif
- FN : Model memprediksi data ada di kelas negatif dan yang sebenarnya data memang ada di kelas positif
- FP : Model memprediksi data ada di kelas positif dan yang sebenarnya data memang ada di kelas negatif

G. Precision, Recall, F1-Score

Nilai *precision*, *recall* dan *f1-score* digunakan untuk melihat kinerja dari model klasifikasi yang terbentuk [10]. *Precision* adalah perbandingan antara jumlah prediksi data kelas positif yang merupakan kelas positif dengan banyak data yang diprediksi positif. Sedangkan *recall* adalah perbandingan antara jumlah prediksi data kelas positif yang merupakan kelas positif dengan banyaknya data yang sebenarnya positif.

$$precision = \frac{TP}{TP + FP} \quad (8)$$

$$recall = \frac{TP}{TP + FN} \quad (9)$$

F1-Score merupakan *harmonic mean* dari *precision* dan *recall*. Nilai terbaik *F1-Score* adalah 1.0 sedangkan nilai terburuknya adalah 0.

$$\frac{1}{F1-Score} = \frac{1}{2} \left(\frac{1}{precision} + \frac{1}{recall} \right) \quad (10)$$

III. METODOLOGI PENELITIAN

Penelitian ini dilakukan dalam beberapa tahapan, yaitu :

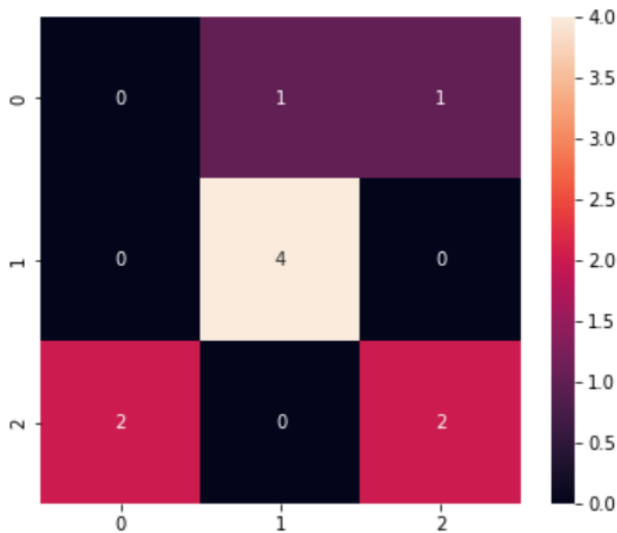
1. Studi Literatur
Pada tahapan ini dilakukan pengumpulan literatur yang berkaitan dengan penelitian, yaitu data mining, klasifikasi, *pre-processing*, *Principal Component Analysis*, *Decision Tree*, *precision*, *recall* dan *f1-score*
2. Perumusan Masalah
Pada tahapan ini dilakukan observasi secara langsung ke Bank Sumsel Babel Cabang KM 12 Palembang untuk mengetahui penentuan status kolektibilitas kredit nasabah.
3. Pengumpulan Data
Pada tahapan ini dilakukan pengumpulan data status kolektibilitas kredit dari nasabah Bank Sumsel Babel Cabang KM 12 Palembang yang digunakan dalam menentukan analisis status pembayaran kredit nasabah. Data yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari Bank Sumsel Babel Cabang KM 12 Palembang. Kriteria yang digunakan terdiri dari 11 atribut/variabel independen yaitu : jenis kelamin, umur, nama, suku bunga, jangka waktu, plafond kredit, sisa hutang, bunga, angsuran pokok dan jumlah. Sedangkan variabel dependen yaitu kolektibilitas terdiri dari : lancar, diragukan, dan macet.
4. Analisis Data
Pada tahapan ini data yang telah dikumpulkan akan digunakan sebagai data latih dan data uji. Pembagian dari dataset dengan komposisi 80% sebagai data latih dan 20% digunakan sebagai data uji. Analisis yang dilakukan adalah pengaruh *pre-processing Principal Component Analysis (PCA)* dalam menentukan status kredit nasabah menggunakan metode *Decision Tree*
5. Penarikan Kesimpulan
Penarikan kesimpulan dilakukan dengan melakukan membandingkan tingkat akurasi metode *Decision Tree* dengan dan tanpa *pre-processing*. Selain itu dilakukan juga analisis terkait kinerja model klasifikasi berdasarkan nilai *precision*, *recall* dan *f1-score*.

IV. HASIL DAN PEMBAHASAN

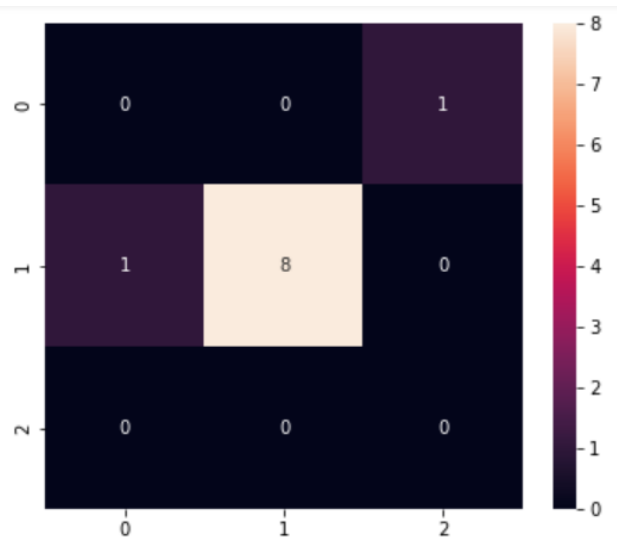
Berdasarkan hasil observasi dan wawancara yang dilakukan pada bagian kredit Bank Sumsel Babel Cabang KM 12 Palembang terdapat 10 atribut yang dijadikan bahan pertimbangan dalam menentukan kolektibilitas atau status kelancaran pembayaran kredit nasabah. 10 atribut tersebut adalah jenis kelamin, umur, nama, suku bunga, jangka waktu, plafond, sisa hutang, bunga, angsuran pokok, dan jumlah. Jumlah dataset yang diperoleh dari Bank Sumsel Babel Cabang KM 12 Palembang sebanyak 49 data kredit nasabah yang telah memiliki kelas kolektibilitas dengan rincian sebanyak 33 data dengan status kolektibilitas lancar,

5 data dengan status kolektibilitas diragukan, dan 11 data dengan status kolektibilitas macet.

Hasil *confusion matrix* dari klasifikasi status kelancaran pembayaran kredit nasabah menggunakan metode *Decision Tree* dapat dilihat pada Gambar 1 sedangkan *Decision Tree* dengan *pre-processing* PCA dapat dilihat pada Gambar 2. Berdasarkan perhitungan menggunakan persamaan 7 diperoleh tingkat akurasi yang dapat dilihat pada Tabel 2, sedangkan untuk melihat kinerja dari model klasifikasi yang terbentuk berdasarkan nilai *precision*, *recall* dan *f1-score* dapat dilihat pada Tabel 3.



Gambar 1. *Confusion Matrix Decision Tree*



Gambar 2. *Confusion Matrix Decision Tree with PCA*

TABEL 2 TINGKAT AKURASI

Metode	Akurasi
Decision Tree	0,70
Naïve Bayes with CFS	0,80

TABEL 3 *PRECISION, RECALL, F1-SCORE*

<i>Decision Tree</i>			
	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
Diragukan	0.00	0.00	0.00
Lancar	0.80	1.00	0.89
Macet	0.67	0.50	0.57
<i>Decision Tree with PCA</i>			
	<i>Precision</i>	<i>Recall</i>	<i>F1-Score</i>
Diragukan	0.00	0.00	0.00
Lancar	1.00	0.89	0.94
Macet	0.00	0.00	0.00

V. KESIMPULAN

Kesimpulan dari hasil penelitian yang telah dilakukan menyatakan bahwa Metode *Principal Component Analysis* memberikan pengaruh yang cukup signifikan terhadap kinerja metode klasifikasi *Decision Tree* pada studi kasus status kredit nasabah pada Bank Sumsel Babel Cabang KM 12 Palembang. Hal ini dapat dilihat berdasarkan terjadinya peningkatan akurasi sebesar 10% serta peningkatan nilai *precision*, penurunan nilai *recall* dan nilai *f1-score* yang mendekati 1.0 yang berarti bahwa model klasifikasi yang terbentuk memiliki kinerja yang lebih baik.

DAFTAR PUSTAKA

- [1] P. Misra and A. S. Yadav, "Impact of Preprocessing Methods on Healthcare Predictions," *SSRN Electron. J.*, no. July, 2019, doi: 10.2139/ssrn.3349586.
- [2] A. Macintosh, R. Euis, and T. A. Eds, *Applications and Innovations in Intelligent Systems XII*. 2005.
- [3] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi: 10.38094/jastt20165.
- [4] M. S. Chen, J. Han, and P. S. Yu, "Data mining: An overview from a database perspective," *IEEE Trans. Knowl. Data Eng.*, vol. 8, no. 6, pp. 866–883, 1996, doi: 10.1109/69.553155.
- [5] B. N. Lakshmi and G. H. Raghunandhan, "A conceptual overview of data mining," *Proc. Natl. Conf. Innov. Emerg. Technol. NCOIET'11*, pp. 27–32, 2011, doi: 10.1109/NCOIET.2011.5738828.
- [6] K. M. R. B. Com, "Data mining techniques," *SpringerBriefs Appl. Sci. Technol.*, vol. 179, no. 10, pp. 13–30, 2016, doi: 10.1007/978-3-319-22294-3_3.
- [7] I. Journal, "A Study Some Data Mining Classification Techniques," *Int. J. Mod. Trends Eng. Res.*, vol. 4, no. 1, pp. 210–215, 2017, doi: 10.21884/ijmter.2017.4031.zt9tv.
- [8] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 375–381, 2003, doi: 10.1080/713827180.
- [9] P. Myoelectric *et al.*, "Principal Components Analysis Preprocessing for Improved Classification

Accuracies in Principal Components Analysis Preprocessing for Improved Classification Accuracies in,” vol. 56, no. October 2016, pp. 1407–1414, 2009.

[10] R. Yacouby and D. Axman, “Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models,” pp. 79–91, 2020, doi: 10.18653/v1/2020.eval4nlp-1.9.