Komputika: Jurnal Sistem Komputer
Volume 11, Nomor 2, Oktober 2022, hlm. 193 - 202
Terakreditasi Peringkat 3, SK No. 28/E/KPT/2019
DOI: 10.34010/komputika.v11i2.6498

ISSN: 2252-9039 (print)
ISSN: 2655-3198 (online)

Klasifikasi Customer Relationship Management Menggunakan Dataset KDD Cup 2009 dengan Teknik Reduksi Dimensi

Fahmi Ardiansyah[1*], Hamdan[2], Sugiyanto[3], Ilham Wahyudi Siadi[4]

[1,2,3,4)] Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Nusa Mandiri
Jl. Kramat Raya No. 18, Jakarta, Indonesia 10450
**email: 14207047@nusamandiri.ac.id*

(Naskah masuk: *21 Januari 2022*; diterima untuk diterbitkan: *01 Juni 2022*)

ABSTRAK *– Customer Relationship Management (CRM) merupakan teknologi yang menghubungkan antara pelanggan dengan bisnis, CRM dapat membantu pertumbuhan bisnis dan meningkatkan loyalitas dalam pelanggan. Pada awalnya CRM hanya berbentuk tulisan tangan, namun dengan berkembangnya teknologi saat ini CRM berkaitan dengan strategi bisnis secara keseluruhan, sistem CRM layaknya berbentuk repository yang mengintegrasikan aktivitas dari penjualan, pemasaran, dan dukungan pelanggan dengan menyederhanakan proses strategi dan pengelolaan penjualan dalam suatu sistem. Contohnya adalah pada data Knowledge Data Discovery (KDD) Cup 2009 yang merupakan Piala KDD 2009 menawarkan kesempatan untuk mengerjakan database pemasaran besar dari Perusahaan Telekomunikasi Prancis Orange untuk memprediksi kecenderungan pelanggan untuk beralih penyedia (churn), beli produk atau layanan baru (appetency), atau beli upgrade atau add-on yang diusulkan ke mereka untuk membuat penjualan lebih menguntungkan (up-selling). Masalahnya karena menangani database yang sangat besar, termasuk data yang heterogen (variabel numerik dan kategorik), dan distribusi kelas yang tidak seimbang ini membutuhkan efisiensi waktu yang cukup lama dalam pengelolaan dataset oleh karena itu dibutuhkan teknik reduksi dimensi yang merupakan teknik pengurangan dari jumlah dimensi dari dataset, dengan dimensi reduksi optimal hasilkan klasifikasi paling baik dengan PCA, PCA dengan klasifikasi Random Forest 96.93%. Klasifikasi LDA dengan Naïve Bayes 61.00%. Klasifikasi SVD dengan Random Forest 95.97%.*

Kata Kunci *– Customer Relationship Management (CRM), KDD Cup 2009, Reduksi Dimensi, Klasifikasi, Machine Learning*

Classification of Customer Relationship Management using KDD Cup 2009 Dataset with Dimension Reduction Technique

ABSTRACT *– Customer Relationship Management (CRM) is a technology that connects customers with business, CRM can help business growth and increase customer loyalty. At first CRM was only in the form of handwriting, but with the development of current technology CRM is related to overall business strategy, the CRM system is like a repository that integrates activities from sales, marketing, and customer support by simplifying the process of strategy and sales management in a system. An example is the 2009 KDD Cup data which is the 2009 KDD Cup offering the opportunity to work on a large marketing database from the French Telecommunications Company Orange to predict a customer's propensity to switch providers (churn), purchase a new product or service (appetency), or purchase upgrades or add-ons. proposed to them to make the sale more profitable (up-selling). The problem is that dealing with very large databases, including heterogeneous data (numeric and categorical variables), and this unbalanced class distribution requires a longtime efficiency in managing datasets, therefore dimension reduction techniques are needed which are techniques for reducing the number of dimensions from dataset, with optimal reduction dimensions resulted in the best classification with PCA, PCA with Random Forest classification 96.93%. LDA classification with Naïve Bayes 61.00%. SVD Classification with Random Forest 95.97%.*

Keywords *- Customer Relationship Management (CRM), KDD Cup 2009, Dimensionality Reduction, Classification, Machine Learning*

1. PENDAHULUAN

Pelanggan merupakan salah satu faktor penentu sebuah perusahaan untuk terus bersaing dalam bisnis. Hal ini dilakukan agar pelanggan menjadi loyal terhadap perusahaan. Tingkat loyalitas ini dapat tercapai dengan adanya *Customer Relationship Management* (CRM) yang tepat sasaran sehingga tercipta hubungan yang baik dan saling menguntungkan dengan para pelanggannya [1]. Dimasa yang sekarang ini CRM sudah merambah ke aplikasi berbasis web, android, maupun desktop karena pasarnya untuk saat ini cukup tinggi[2], mengakibatkan banyak saat ini *database* yang berkaitan dengan CRM[3]. Contoh pada data *Knowledge Data Discovery* (KDD) Cup 2009 ini merupakan bidang pemasaran dari perusahaan *Telecom company Orange* untuk memprediksi kecenderungan pelanggan untuk beralih ke provider (*churn*), membeli yang baru produk atau layanan (*appetency*)[4]. Atau membeli *upgrade* atau tambahan yang diusulkan kepada mereka untuk penjualan lebih menguntungkan (*Upselling*). Tantangan dalam KDD Cup 2009 ini adalah untuk mengalahkan sistem *in-house* yang dikembangkan oleh *Orange Labs*. *Database* yang sangat besar, termasuk data berisik yang heterogen (variabel numerik dan kategorik), dan distribusi kelas yang tidak seimbang. Waktu efisiensi sering menjadi poin penting.

Banyaknya tingkat loyalitas dalam CRM diperlukan metode–metode untuk memenuhi loyalitas tersebut salah satunya dengan menggunakan metode klasifikasi, regresi, klasterisasi[5]. Dalam penelitian ini metode yang digunakan untuk data berdimensi tinggi menggunakan metode klasifikasi, Klasifikasi adalah pembagian yang bertujuan memilih suatu objek kedalam suatu kelas atau kategori yang sudah dipengaruhi sebelumnya.

Data *churn* pada data KDD Cup 2009 digunakan untuk memprediksi kecenderungan pelanggan untuk beralih ke provider, dalam data tersebut memiliki dimensi yang tinggi. Untuk memecahkan data yang dimensi tinggi tersebut digunakan metode pengurangan fitur yaitu fitur ekstraksi dan fitur seleksi, fitur seleksi digunakan dalam data berdimensi tinggi dengan mengurangi jumlah fitur yang ada dalam data dengan cara memilih fitur relevan pada data, sedangkan fitur ekstraksi adalah suatu pengurangan dalam fitur dari suatu data yang nilainya didapatkan dari karakter intrinsik pada data[6]. Penelitian ini memiliki data CRM sebanyak 230 atribut, maka dari itu data ini tergolong data berdimensi tinggi, data berdimensi tinggi menjadi masalah karena memiliki banyaknya atribut dari data menyebabkan waktu komputasi cenderung lebih lambat, banyaknya data yang redudansi. Oleh karena itu dibutuhkan reduksi dimensi untuk mengatasi dari masalah-masalah tersebut.

Beberapa penelitian terkait yang pernah dilakukan oleh Inayati yang menjelaskan tentang reduksi dimensi pada kasus kualitas jasa *e-commerce* di Indonesia dengan menggunakan reduksi dimensi data yang didapatkan melalui kuesioner dengan jumlah item pertanyaan lebih dari 100 melibatkan lebih 1000 responden sehingga data yang terkumpul sangat besar , hasil pengelolaan data tersebut bahwa terdiri atas 5 dimensi penilaian kualitas jasa (*Tangible*, *Reliability*, *Responsiveness*, *Assurance*, dan *Empathy*), dapat direduksi menjadi dua dimensi, yakni *Responsiveness* dan *Assurance* [7].

Rahmi *et al* [8] pada penelitian tentang Optimasi Klasifikasi Bayesian Network Melalui Reduksi Attribute Menggunakan Metode Principal Component Analysis Principal Component Analysis dengan menggunakan dataset faktor-faktor yang mempengaruhi ketidakhadiran karyawan diambil dari Repository 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 kelima terjadi penurunan akurasi sebesar 89,7%.

Selanjutnya penelitian yang dilakukan Elvina *et al* [9] tentang reduksi dimensi pada optimasi portofolio mean-variance menggunakan *non-negative principal component analysis* menerangkan bahwa Optimasi portofolio mean variance merupakan suatu metode dasar untuk menentukan alokasi investasi yang optimal ke berbagai saham. Implementasi metode reduksi dimensi *Non-negative Principal Component Analysis* sebagai alat pra-pemrosesan pada optimasi portofolio mean-variance, menghasilkan *expected return* sebesar 0,107% dan variance 0,020%, lebih baik dibanding menggunakan metode reduksi dimensi Principal Component Analysis yang menghasilkan *expected return* 0,103% dan *variance* 0,019% dan optimasi portofolio tanpa reduksi dimensi yang menghasilkan *expected return* 0,095% dan *variance* 0,010%.

Penulisan dalam penelitian ini akan menggunakan pendekatan dimensi reduksi dengan metode ekstraksi fitur dan seleksi fitur secara bersamaan sehingga dapat menghasilkan tingkat akurasi dan presisi lebih baik dalam memprediksi kecenderungan pelanggan untuk beralih ke provider (*churn*), kemudian penulis juga menganalisa komparasi klasifikasi menggunakan *Random Forest, Naïve Bayes, SVM (Singular Value Decomposition).*

Tujuan dari penelitian ini adalah mereduksi dimensi tinggi, hasil data yang telah direduksi kemudian diklasifikasi untuk menemukan akurasi

dan presisi yang lebih baik dalam memprediksi pelanggan untuk beralih ke provider ya atau tidak dengan menggunakan teknik klasifikasi *Random Forest, Naïve Bayes, SVM (Singular Value Decomposition).*

2. METODE BAHAN

Data

Data yang digunakan dalam penelitian ini merupakan data KDD Cup 2009 yang diambil dari website KDD *repository* yang digunakan pada penelitian sebelumnya[4]. Data tersebut berisikan data pelanggan yang dikumpulkan sebelumnya oleh isabelle guyon, Vincent Lemaire, Gideon Dror, David Vogel. Data tersebut memiliki jumlah banyaknya data sebanyak 50000 data dan 230 kolom, data tersebut berisi data mentah yang tidak teratur dimana didalamnya masih ada angka dan nilai kosong yang tidak beraturan. KDD Cup 2009 pekerjaan dibidang pemesanan dari perusahaan *Telecom company Orange* untuk memprediksi kecenderungan pelanggan untuk beralih ke provider (*churn*), membeli yang baru produk atau layanan (*appetency*). Atau membeli upgrade atau tambahan yang diusulkan kepada mereka untuk penjualan lebih menguntungkan (Upselling).

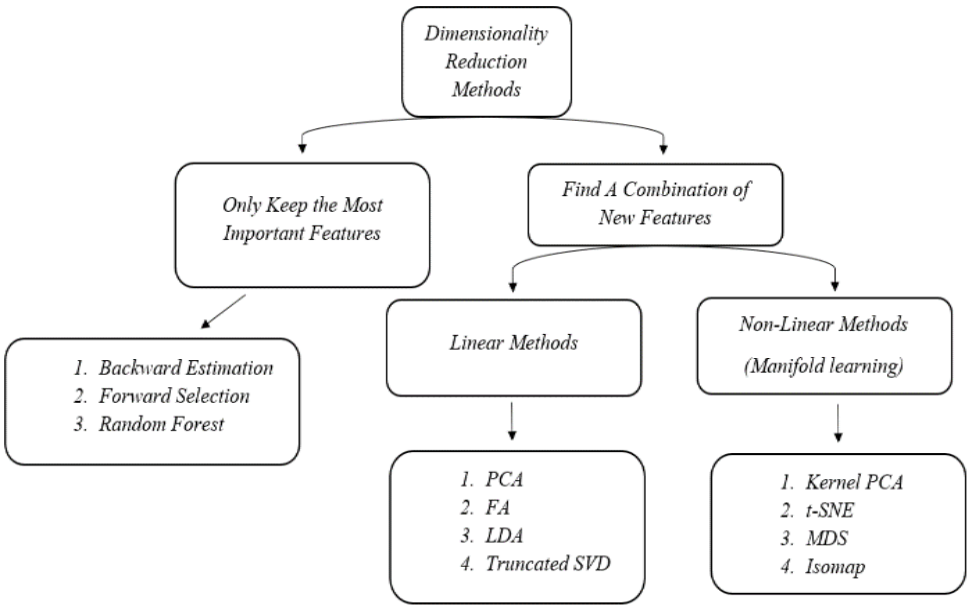
Dimensi Reduksi

Dimensi reduksi merupakan pengurangan dimensi suatu dataset dengan pertimbangan bahwa informasi-informasi penting tetap dipertahankan. Reduksi dimensi dapat digunakan untuk penyederhanaan komputasi pada data besar, dengan dimensi yang lebih rendah hasil analisis pada data hasil reduksi masih menghasilkan kesimpulan yang relevan[10]. Jenis–jenis dimensi reduksi diperlihatkan pada Gambar 1.

Principal Component Analysis (PCA)

Principal Component Analysis atau disebut juga teknik yang digunakan untuk menyederhanakan suatu data, dari jumlah data yang diperkecil tetapi tidak menghilangkan sifat dari data tersebut[11]. Kompleksitas membuat sulit dalam pengolahan datanya, dalam meningkatkan interpretabilitas tanpa menghilangkan informasi yang terkandung di dalam data membutuhkan reduksi dimensi dalam menentukan variabel-variabel utama. Berikut langkah–langkah pengerjaan dalam mereduksi atribut PCA dalam reduksi dimensi sebagai berikut:

a. Pembentukan Matriks
   Tahapan dalam PCA pertama kali adalah pembentukan matriks. Data direpresentasikan dalam matriks *Mxn*, dimana *M* merupakan jumlah data yang dilatih dan *n* merupakan banyaknya kelas atau feature. Dari matriks tersebut dilakukan proses ekstraksi menjadi data yang berdimensi lebih kecil dijadikan menjadi matriks.
b. Menghitung Matriks Kovarian
   Menghitung matriks kovarian biasanya setiap vector ditransformasikan secara linear dalam satu vektor yang baru disimbolkan dengan huruf s dinyatakan dalam rumus $S_t = U^T . xi$ dimana U merupakan matrik orthogonal mxm dengan kolom ke I, u merupakan nilai *eigenvector* dari sampel matrik kovarian.
c. Menghitung *Eigenvalue* dan *Eigenvector* Dari Matrik Kovarian.
d. Menghitung nilai transformasi orthogonal.



Gambar 1. Jenis Reduksi Dimensi

***Linear Discriminant Analysis* (LDA)**

LDA bekerja berdasarkan analisa matrik penyebaran (scatter matrix analysis) yang bertujuan menemukan suatu proyeksi optimal sehingga dapat memproyeksikan data input pada ruang dengan dimensi yang lebih kecil dengan pola yang dipisahkan[12]. Berikut langkah-langkah dalam mereduksi data dengan LDA sebagai berikut:

a. Tahapan pertama pada proses LDA adalah mengubah data latih dan uji menjadi vektor.
b. Membuat kelas berdasarkan data latih dan data uji.
c. Hitung rata-rata dari kelas dan rata-rata kelas keseluruhan dari seluruh data.
d. Menghitung matriks sebaran antarkelas.
e. Menghitung matriks sebaran dalam kelas.
f. Mencari matriks W.
g. Mencari vektor eigen dari W dan mengurutkan nilai eigen (λ) sesuai dengan urutan nilai yang ada pada nilai eigen dari besar ke kecil.
h. Melakukan reduksi dimensi dengan cara melakukan proyeksi dengan vector.

***t-Distributed Stochastic Neighbor Embedding* (TSNE)**

t-Distributed Stochastic Neighbor Embedding atau sering disebut t-SNE adalah teknik nonlinear unsupervised yang digunakan untuk reduksi dimensi, eksplorasi data, dan visualisasi data berdimensi tinggi. t-SNE menghitung ukuran kesamaan antara pasangan titik data di ruang berdimensi tinggi dan dimensi rendah, kemudian mengoptimalkan dua kesamaan ini[13]. Berikut langkah-langkah dalam mereduksi data dengan TSNE sebagai berikut:

a. Tahapan pertama adalah temukan persamaan yang berpasangan antar titik terdekat dalam ruang dimensi tinggi, karena TSNE mengubah jarak Euclidean yang tinggi antara titik data.
b. Petakan setiap titik dalam ruang dimensi tinggi ke berdimensi rendah berdasarkan keterkaitan berpasangan antara titik di ruang berdimensi tinggi.
c. Temukan representasi data yang berdimensi rendah menggunakan penurunan gradien.
d. Gunakan distribusi Student-t untuk menghitung kesamaan antara dua titik di ruang berdimensi rendah.

***Singular Value Decomposition* (SVD)**

Singular Value Decomposition (SVD) merupakan sebuah teknik komputasi numerik yang melakukan faktorisasi terhadap sebuah matriks tak nol sehingga diperoleh tiga matriks tak nol yang baru. Salah satu matriks yang diperoleh dari proses SVD akan memuat nilai-nilai singular dari matriks asal[14]. Berikut langkah-langkah dalam mereduksi data dengan SVD sebagai berikut:

a. Langkah pertama mendefinisikan vektor eigen ortonormal yang bersesuaian dengan nilai vektor eigen tidak-nol.
b. Mereduksi dimensi dari matriks adalah dengan mengurangi dimensi dari matriks S (nilai singular) yang berupa matriks diagonal.
c. mengalikan matriks baru U (vektor singular kiri) dan S (nilai singular) sehingga menghasilkan matriks H (matriks hasil pembobotan) baru.

***Multidimensional Scaling* (MDS)**

Analisis *multidimensional scaling* merupakan salah satu teknik peubah ganda yang dapat digunakan untuk menentukan posisi suatu objek lainnya berdasarkan penilaian kemiripannya[15]. Berikut langkah-langkah dalam mereduksi data dengan MDS sebagai berikut:

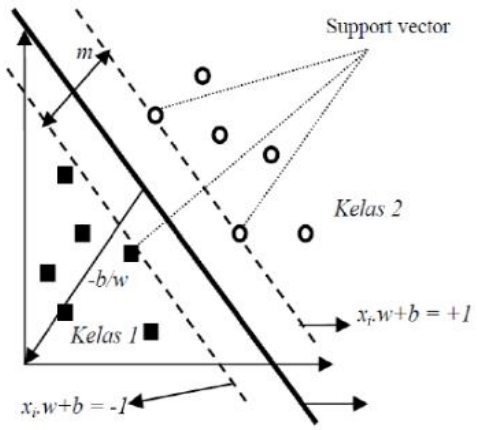
a. Langkah pertama temukan kemiripan dan ketidakmiripan antar objek.
b. Berikan hasil berupa titik-titik sehingga jarak antar titik menggambarkan tingkat kemiripan atau ketakmiripan.
c. Berikan petunjuk dalam mengidentifikasi faktor yang mempengaruhi munculnya kemiripan atau ketakmiripan.

Teknik Klasifikasi

Klasifikasi merupakan pembagian terstruktur mengenai artinya proses yang bertujuan memilih suatu objek kedalam suatu kelas atau kategori yang sudah dipengaruhi sebelumnya. Menurut pembagian terstruktur mengenai ialah proses dari pembangunan terhadap suatu contoh yang mengklasifikan suatu objek sinkron menggunakan atribut –atributnya. pembagian terstruktur mengenai data ataupun dokumen pula dapat dimulai asal menciptakan aturan pembagian terstruktur mengenai eksklusif menggunakan data training yang diklaim sebagai tahapan pembelajaran serta pengujian dipergunakan menjadi data testing[16].

***Support Vector Machine* (SVM)**

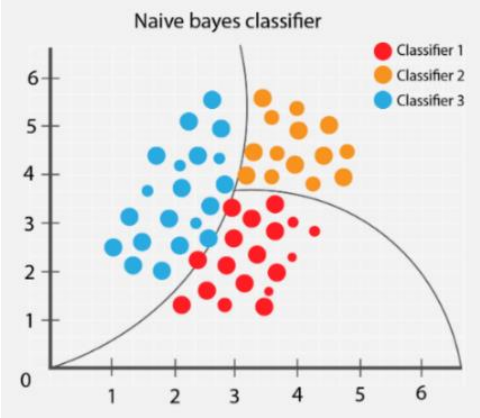
SVM adalah metode *supervised machine learning* yang digunakan untuk klasifikasi data. Dalam teknik ini, setiap objek statistik diplotkan sebagai titik dalam ruang berdimensi-n dengan nilai setiap fungsi menjadi biaya koordinat yang dipilih. Selanjutnya digambarkan pada Gambar 2 dilakukan klasifikasi dengan mendesain *hyperplane* yang membedakan kedua kelas tersebut. Dalam area dua dimensi, *hyperline* adalah garis yang membagi bidang menjadi dua bagian atau kelas[17].



Gambar 2. Hyperline SVM

Naïve Bayes

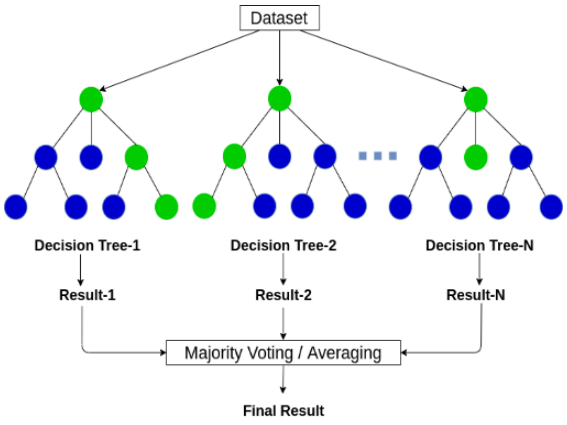
Naïve Bayes (NB) adalah salah satu algoritma data mining yang paling terkenal untuk klasifikasi. *Naïve Bayes Classifier* (NB) adalah pengklasifikasi yang sangat sederhana berdasarkan teorema *Bayes*. NB digambarkan pada Gambar 3 mengasumsikan semua node fitur di dalam kelas independen dan fitur variabel diasumsikan distribusi Gaussian jika bersifat kontinu[18].



Gambar 3. Naive bayes classifier

Random Forest

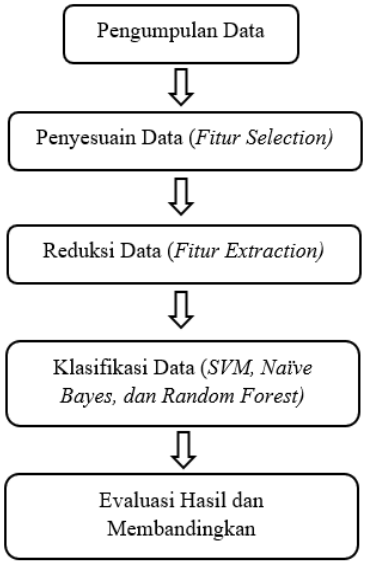
Random Forest adalah kombinasi pohon predictor dimana setiap pohon bergantung pada nilai vector acak yang dijadikan sampel independen dengan distribusi yang sama untuk semua pohon di hutan. Metode Random Forest yang digambarkan pada Gambar 4 merupakan metode klasifikasi yang dilakukan dengan menggunakan struktur pohon dalam jumlah besar daripada satu struktur pohon. Dalam metode ini, sampel dari kumpulan data dipilih dengan metode Bootstrap dan pohon klasifikasi dibuat [19].



Gambar 4. *Random Forest*

Langkah Penelitian

Langkah Penelitian yang dilakukan dalam penelitian ini digambarkan pada Gambar 5.



Gambar 5. Langkah Penelitian

a) Melakukan pemrosesan data dan pengumpulan data meliputi data uji dan data pelatihan KDD Cup 2009.
b) Melakukan penyesuaian data dengan menggunakan metode seleksi fitur
c) Reduksi data dengan menggunakan ekstraksi fitur menggunakan teknik dimensi reduksi PCA, LDA, T-SNE, MDS, SVD.
d) Mengklasifikasikan data menggunakan metode SVM, Naïve Bayes, dan Random Forest
e) Menginterpretasikan hasil analisis dan melakukan perbandingan akurasi dan presisi antara hasil yang telah direduksi serta menarik kesimpulan dari hasil perbandingan akurasi dan presisi yang memiliki akurasi dan presisi yang lebih baik.

3. HASIL DAN PEMBAHASAN

Hasil yang dilakukan dengan teknik dimensi reduksi dengan teknik dimensi reduksi fitur seleksi dan fitur ekstraksi diantaranya PCA (*Principal Component Analysis*), TSNE (*t-Distributed Stochastic Neighbor Embedding*), LDA (*Linear Discriminant Analysis*), SVD (*Singular Value Decomposition*), MDS (*Multidimensional Scaling*). Pada gambar 6 dan 7 merupakan dataset uji dan data pelatihan yang didapatkan dari website KDD yang berjumlah 50000 baris dengan jumlah feature sebanyak 230 nantinya akan digabungkan kedua data tersebut dan digunakan untuk reduksi dimensi setelah data tersebut telah direduksi langkah berikutnya dengan menggunakan metode klasifikasi dari SVM, Naïve Bayes, dan *Random Forest* masing-masing dari metode tersebut menggunakan kelasnya sebanyak 2,

```
In [3]:
# Baca data test dari file local
test = pd.read_table('C:/Users/Fahmi-PC/machine_learning/dimred-project/data/orange_small_test.data')
# tampil dataframe
test.head()
```

Out[3]:

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	...	Var221	Var222	Var223	Var224	Var225	Var226	Var227	Var228
0	NaN	NaN	NaN	NaN	NaN	1225.0	7.0	NaN	NaN	NaN	...	Al6ZaUT	P6pu4VI	LM8l689qOp	NaN	ELof	7P5s	Zl9m	R4y5gQQWY8OodqDV
1	NaN	NaN	NaN	NaN	NaN	259.0	0.0	NaN	NaN	NaN	...	oslk	S46Rt72	LM8l689qOp	NaN	NaN	Qu4f	RAYp	F2FyR07IdsN7I
2	NaN	NaN	NaN	NaN	NaN	861.0	14.0	NaN	NaN	NaN	...	oslk	CcdTy9x	LM8l689qOp	NaN	NaN	7aLG	RAYp	F2FyR07IdsN7I
3	NaN	NaN	NaN	NaN	NaN	1568.0	7.0	NaN	NaN	NaN	...	oslk	Q53Rkup	LM8l689qOp	NaN	kG3k	7P5s	RAYp	TCU50_Yjmm6GIBZ0lL_
4	NaN	NaN	NaN	NaN	NaN	1197.0	7.0	NaN	NaN	NaN	...	Al6ZaUT	WfsWw2A	LM8l689qOp	NaN	ELof	5Acm	Zl9m	iyHGyLCEkQ

5 rows × 230 columns

```
In [7]:
print(test.shape)
```

```
(50000, 230)
```

Gambar 6. Dataset Uji KDD 2009

```
In [6]:
# baca data training dari file local
train = pd.read_table('C:/Users/Fahmi-PC/machine_learning/dimred-project/data/orange_small_train.data')
# tampil dataframe sebanyak 20 baris
train.head(20)
```

Out[6]:

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	...	Var221	Var222	Var223	Var224	Var225	Var226	Var227	
0	NaN	NaN	NaN	NaN	NaN	1526.0	7.0	NaN	NaN	NaN	...	oslk	fXVEsaq	jySVZNlOJy	NaN	NaN	xb3V	RAYp	F2FyR0
1	NaN	NaN	NaN	NaN	NaN	525.0	0.0	NaN	NaN	NaN	...	oslk	2Kb5FSF	LM8l689qOp	NaN	NaN	fKCe	RAYp	F2FyR0
2	NaN	NaN	NaN	NaN	NaN	5236.0	7.0	NaN	NaN	NaN	...	Al6ZaUT	NKv4yOc	jySVZNlOJy	NaN	kG3k	Qu4f	02N6s8f	ib5G6X1e
3	NaN	NaN	NaN	NaN	NaN	NaN	0.0	NaN	NaN	NaN	...	oslk	CE7uk3u	LM8l689qOp	NaN	NaN	FSa2	RAYp	F2FyR0
4	NaN	NaN	NaN	NaN	NaN	1029.0	7.0	NaN	NaN	NaN	...	oslk	1J2cvxe	LM8l689qOp	NaN	kG3k	FSa2	RAYp	F2FyR0
5	NaN	NaN	NaN	NaN	NaN	658.0	7.0	NaN	NaN	NaN	...	zCkv	QqVuch3	LM8l689qOp	NaN	NaN	Qcbd	02N6s8f	Zy
6	NaN	NaN	NaN	NaN	NaN	1680.0	7.0	NaN	NaN	NaN	...	oslk	XlgxB9z	LM8l689qOp	NaN	kG3k	FSa2	RAYp	55
7	NaN	NaN	NaN	NaN	NaN	77.0	0.0	NaN	NaN	NaN	...	oslk	R2LdzOv	NaN	NaN	NaN	FSa2	RAYp	F2FyR0
8	NaN	NaN	NaN	NaN	NaN	1176.0	7.0	NaN	NaN	NaN	...	zCkv	K2SqEo9	jySVZNlOJy	NaN	kG3k	PM2D	6fzt	am14IcfM7tWLrUmR
9	NaN	NaN	NaN	NaN	NaN	1141.0	7.0	NaN	NaN	NaN	...	oslk	EPqQcw6	LM8l689qOp	NaN	kG3k	FSa2	RAYp	55
10	NaN	NaN	NaN	NaN	NaN	490.0	7.0	NaN	NaN	NaN	...	zCkv	catzS2D	LM8l689qOp	NaN	kG3k	WqMG	Zl9m	ib5G6X1e
11	NaN	NaN	NaN	NaN	NaN	798.0	14.0	NaN	NaN	NaN	...	oslk	VWtlBQt	LM8l689qOp	NaN	kG3k	FSa2	RAYp	F2FyR0
12	NaN	NaN	NaN	NaN	NaN	595.0	0.0	NaN	NaN	NaN	...	zCkv	WfsWw2A	LM8l689qOp	NaN	NaN	me1d	Zl9m	iyHGy
13	NaN	NaN	NaN	NaN	NaN	2268.0	0.0	NaN	NaN	NaN	...	oslk	QKXEsaq	LM8l689qOp	NaN	xG3x	Qu4f	RAYp	F2FyR0
14	NaN	NaN	NaN	NaN	NaN	3633.0	7.0	NaN	NaN	NaN	...	oslk	kYwEsaq	LM8l689qOp	NaN	NaN	PM2D	RAYp	F2FyR0
15	NaN	NaN	NaN	NaN	NaN	259.0	0.0	NaN	NaN	NaN	...	oslk	IlvQguc	LM8l689qOp	NaN	ELof	Qu4f	RAYp	55
16	NaN	NaN	NaN	NaN	NaN	5152.0	7.0	NaN	NaN	NaN	...	oslk	WfsWw2A	jySVZNlOJy	NaN	kG3k	FSa2	Zl9m	iyHGy
17	NaN	NaN	NaN	NaN	NaN	1449.0	7.0	NaN	NaN	NaN	...	QKW8DRm	catzS2D	LM8l689qOp	NaN	NaN	WqMG	Zl9m	ib5G6X1e
18	NaN	NaN	NaN	NaN	NaN	574.0	7.0	NaN	NaN	NaN	...	oslk	DQ3iZDY	jySVZNlOJy	NaN	kG3k	FSa2	RAYp	F2FyR0
19	NaN	NaN	NaN	NaN	NaN	658.0	7.0	NaN	NaN	NaN	...	oslk	VQZzrB7	LM8l689qOp	NaN	NaN	5Acm	RAYp	F2FyR0

20 rows × 230 columns

```
In [5]:
print(train.shape)
```

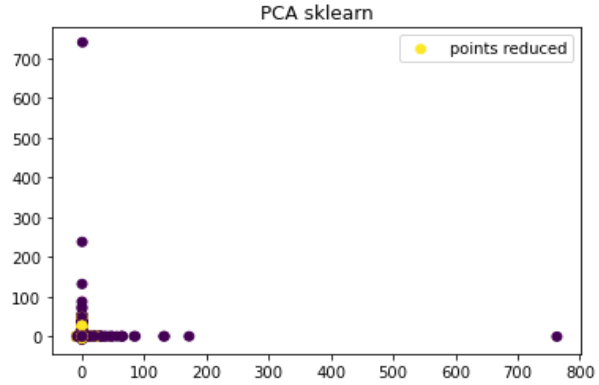
```
(50000, 230)
```

Gambar 7. Dataset Pelatihan KDD 2009

kelas tersebut dihasilkan dari teknik dimensi reduksi sebanyak 2 kelas.

PCA

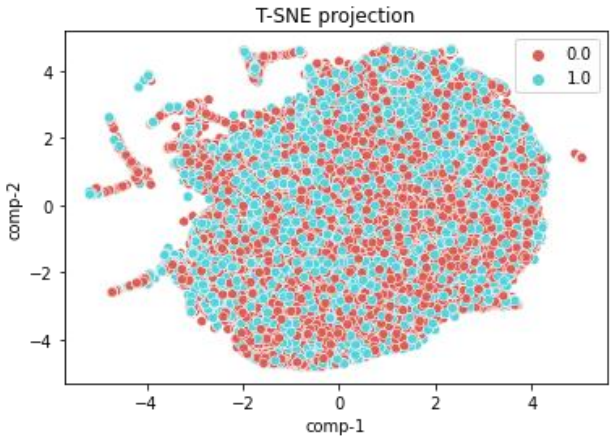
Dari hasil reduksi dengan PCA menggunakan software *anaconda,* jumlah data feature sebelumnya ada 230 *features*. Setelah direduksi dengan PCA mendapatkan 189 features. Gambar 8 merupakan hasil dimensi reduksi dengan ekstraksi fitur dengan PCA.



Gambar 8. Teknik PCA

TSNE

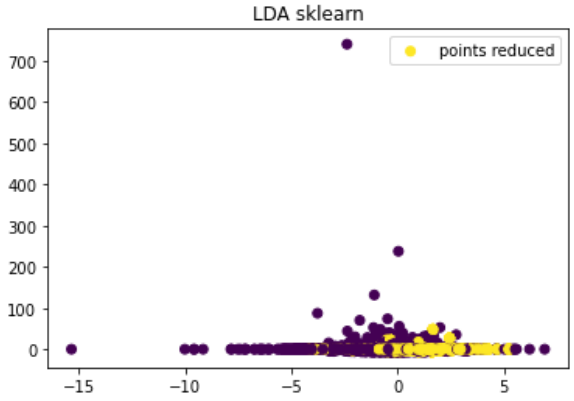
Dari hasil reduksi dengan TSNE menggunakan software *anaconda,* jumlah data feature sebelumnya ada 230 *features*. Setelah direduksi dengan TSNE mendapatkan 121 features. Gambar 9 merupakan bentuk gambar hasil dimensi reduksi dengan ekstraksi fitur dengan TSNE.



Gambar 9. Teknik TSNE

LDA

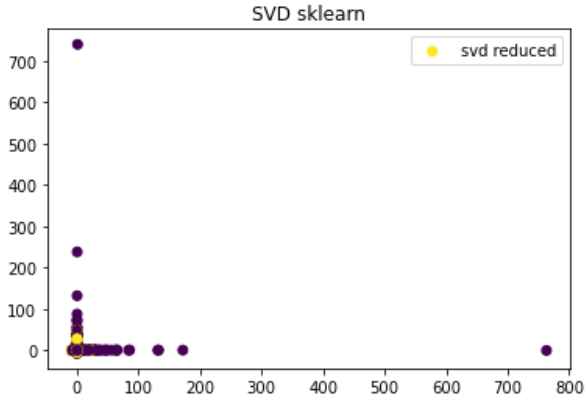
Dari hasil reduksi dengan LDA menggunakan software *anaconda,* jumlah data feature sebelumnya ada 230 *features*. Setelah direduksi dengan LDA mendapatkan 2 features. Gambar 10 merupakan bentuk gambar hasil dimensi reduksi dengan ekstraksi fitur dengan LDA.



Gambar 10. Teknik LDA

SVD

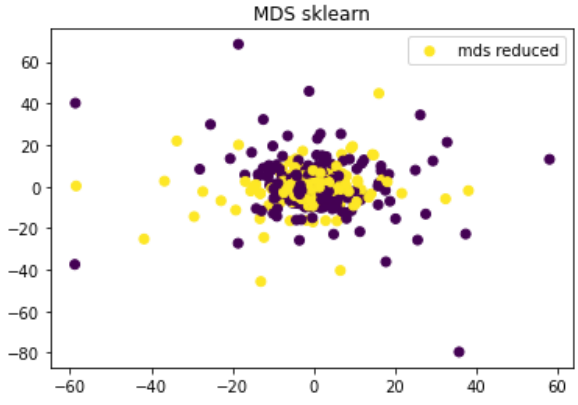
Dari hasil reduksi dengan SVD menggunakan software *anaconda,* jumlah data feature sebelumnya ada 230 *features*. Setelah direduksi dengan SVD mendapatkan 189 features. Gambar 11 merupakan bentuk gambar hasil dimensi reduksi dengan ekstraksi fitur dengan SVD.



Gambar 11. Teknik SVD

MDS

Dari hasil reduksi dengan MDS menggunakan software *anaconda,* jumlah data feature sebelumnya ada 230 *features*. Setelah direduksi dengan MDS mendapatkan 2 features. Gambar 12 merupakan bentuk gambar hasil dimensi reduksi dengan ekstraksi fitur dengan MDS.



Gambar 12. Teknik MDS

Dari hasil reduksi dimensi dengan fitur ekstraksi yang sudah didapatkan jumlah fitur setelah direduksi. Kemudian diuji coba dengan metode pengklasifikasi seperti Random Forest, Naive Bayes, SVM, *experiment setup* yang dilakukan dalam teknik klasifikasi.

Random Forest

Klasifikasi menggunakan Random Forest dengan menggunakan kedalaman *tree* 100, jumlah variabel yang terdapat pada setiap percabangan sebanyak 10, hasilnya dengan menggunakan data yang telah di direduksi menggunakan fitur seleksi dari ekstraksi fitur sebanyak 8000 data data menggunakan *random forest* klasifikasi yang mengklasifikasikan data pada data uji dengan benar sebesar 96.93%. Sehingga klasifikasi lumayan tepat mengklasifikasikan data pada data uji.

SVM

Klasifikasi menggunakan SVM, dilakukan dengan pengujian dengan jenis kernel hasil terbaik menggunakan *RBF* kernel, tipe SVM klasifikasi, dan cost 1 dan dilakukan pengulangan sebanyak 10 kali. Hasilnya dengan menggunakan data yang telah direduksi menggunakan fitur seleksi dari ekstraksi fitur sebanyak 8000 data data menggunakan SVM klasifikasi mendapatkan akurasi rata-rata 61.71%. Artinya, kemampuan klasifikasi yang dibentuk dengan metode SVM mengklasifikasikan data pada data uji cukup optimal pada data uji.

Naïve Bayes

Klasifikasi menggunakan Naïve Bayes, dilakukan menggunakan pada data uji dengan *laplace* 0. Hasilnya dengan menggunakan data yang telah di reduksi menggunakan fitur seleksi dari ekstraksi fitur sebanyak 8000 data menggunakan Naïve Bayes klasifikasi mendapatkan akurasi rata-rata 54.00%. Artinya, kemampuan klasifikasi yang dibentuk dengan metode Naïve Bayes mengklasifikasikan data pada data uji cukup buruk pada data uji.

Hasil *experiment setup* di atas ditampilkan pada Tabel 1.

Tabel 1. Hasil Penelitian dengan Dimensi Reduksi

	Reduksi	Jumlah Kelas	SVM	Naïve Bayes	Random Forest
PCA	18%	2	51.68%	50%	96.93%
LDA	18%	2	61%	61%	49.86%
SVD	18%	2	51.69%	50%	95.97%
TSNE	18%	2	57.68%	54%	96.85%
MDS	18%	2	60%	52%	49.5%

Tabel 1 merupakan hasil penelitian teknik klasifikasi menggunakan Random Forest, SVM, dan Naïve Bayes. Dari ketiga metode klasifikasi tersebut menggunakan Random Forest hasilnya lebih tinggi dibandingkan SVM dan Naïve Bayes. Untuk teknik reduksi dimensi MDS dan LDA lebih kecil hasil Random Forest dikarenakan setelah dilakukan reduksi dimensi dari 230 fitur menggunakan teknik reduksi dimensi LDA dan MDS hanya menghasilkan kelas 2 dari 230 kemungkinan dari 2 kelas ini yang menyebabkan hasil MDS dan LDA lebih kecil metode klasifikasi Random Forest, karena kelebihan Random Forest adalah semakin tinggi data yang dimiliki semakin tepat klasifikasi yang dilakukan Random Forest sebaliknya jika data fitur rendah berarti semakin rendah tingkatan dalam pohon keputusan semakin cepat juga proses klasifikasi namun hasil yang dihasilkan dari klasifikasi semakin rendah. Untuk SVM dan Naïve Bayes hasil dari klasifikasi dari metode tersebut relatif hasilnya lebih kecil dari Random Forest, dari SVM, SVM tidak cocok dengan data yang jumlahnya besar, SVM tidak bekerja baik dengan kumpulan data yang memiliki banyak data yang tumpang tindih *(redundant)*, untuk Naïve bayes membuat asumsi yang sangat kuat tentang bentuk distribusi data yaitu dua fitur tidak tergantung pada kelas keluaran. Karena ini, hasilnya bisa berpotensi sangat buruk - karenanya, pengklasifikasi naif, Masalah lain dari Naïve Bayes terjadi karena kelangkaan data. Untuk kemungkinan nilai fitur, hal ini dapat mengakibatkan probabilitas menuju 0 atau 1, yang pada gilirannya menyebabkan ketidakstabilan numerik dan hasil yang lebih buruk.

Tabel 2. Hasil Penelitian tanpa Dimensi Reduksi

Reduksi	Jumlah Kelas	SVM	Naïve Bayes	Random Forest
0%	192	52.52%	68%	99.39%

Berdasarkan pada Tabel 2 tanpa menggunakan dimensi reduksi hasil yang dihasilkan dari teknik klasifikasi antara SVM, Naïve Bayes, dan Random Forest. Ketiga metode klasifikasi tersebut menggunakan Random Forest hasilnya lebih tinggi dibanding SVM dan Naïve Bayes dapat disimpulkan bahwa reduksi dimensi mengurangi dimensi suatu dataset dengan pertimbangan bahwa informasi-informasi penting tetap dipertahankan dapat dilihat hasilnya tidak jauh berbeda antara menggunakan reduksi dimensi dengan yang tanpa reduksi dimensi.

4. KESIMPULAN

Berdasarkan hasil dan pembahasan pada bagian sebelumnya, dapat disimpulkan dari teknik dimensi reduksi dengan teknik klasifikasi, hasil penelitian bahwa hasil yang paling baik dalam data KDD cup 2009 ini ada di teknik klasifikasi Random Forest dapat menghasilkan nilai akurasi, dan presisi yang cukup tinggi dan stabil dengan masing masing dari semua metode dimensi reduksi pada data uji

direduksi sebanyak 18% dari jumlah kelas, yakni 96,93% dengan reduksi dimensi, pada akurasi dan 99,39% tanpa reduksi dimensi dapat disimpulkan bahwa reduksi dimensi mengurangi dimensi suatu dataset dengan pertimbangan bahwa informasi-informasi penting tetap dipertahankan, dapat dilihat hasilnya tidak jauh berbeda antara menggunakan reduksi dimensi dengan yang tanpa reduksi dimensi. Dan mengapa hasil Random Forest lebih tinggi secara keseluruhan baik menggunakan reduksi dimensi maupun tanpa reduksi dimensi, karena dataset yang digunakan cukup banyak karena kelebihan Random Forest adalah semakin tinggi data yang dimiliki semakin tepat klasifikasi yang dilakukan Random Forest.

Saran yang dapat dilakukan bagi peneliti lain adalah menggunakan metode Neural Network untuk data yang mempelajari tentang KDD cup 2009 karena menghasilkan nilai akurasi, presisi, dan penarikan yang lebih tinggi. Semakin tinggi ketiga nilai tersebut maka semakin bagus metode dan model yang digunakan.

DAFTAR PUSTAKA

[1] W. Maulana And D. L. Pramitaputri, "Pengaruh Customer Relationship Management (CRM) Terhadap Loyalitas Pelanggan Xl Axiata Sampang," *Makro J. Manaj. Dan Kewirausahaan*, Vol. 3, No. 2, Pp. 225–238, 2018, Doi: 10.36467/Makro.2018.03.02.07.

[2] Y. Irawan, "Sistem Informasi Pemasaran Busana Syar'i Dengan Penerapan Customer Relationship Management (CRM) Berbasis Web," *J. Inf. Technol. Comput. Sci.*, Vol. 8, No. 5, P. 55, 2019.

[3] S. Bahri And S. Dalis, "Rancang Bangun E-Enrollment Berbasis Web Menggunakan Customer Relationship Management (CRM) Pada Sekolah Dasar Islam Terpadu," *J. Tek. Komput.*, Vol. 4, No. 1, Pp. 205–211, 2018.

[4] I. Guyon, V. Lemaire, M. Boullé, G. Dror, And D. Vogel, "Design And Analysis Of The Kdd Cup 2009," *Acm Sigkdd Explor. Newsl.*, Vol. 11, No. 2, Pp. 68–76, 2010, Doi: 10.1145/1809400.1809414.

[5] R. Perangin-Angin *Et Al.*, "Comparison Detection Edge Lines Algoritma Canny Dan Sobel," *J. Times*, Vol. Viii, No. 2, Pp. 35–42, 2020.

[6] L. Zhao, Q. Gao, X. J. Dong, A. Dong, And X. Dong, "K- Local Maximum Margin Feature Extraction Algorithm For Churn Prediction In Telecom," *Cluster Comput.*, Vol. 20, No. 2, Pp. 1401–1409, 2017, Doi: 10.1007/S10586-017-0843-2.

[7] M. Inayati And T. A. Yogyakarta, "Usulan Reduksi Dimensi Penilaian Kualitas Jasa Pada Kasus Data Kualitas Jasa E-Commerce Di Indonesia Pada Kasus Data Kualitas Jasa E-Commerce Di Indonesia," *J. Teknol.*, No. May, Pp. 1–5, 2018.

[8] S. Rahmi, P. Sirait, And E. S. Panjaitan, "Optimasi Klasifikasi Bayesian Network Melalui Reduksi Attribute Menggunakan Metode Principal Component Analysis," *J. Media Inform. Budidarma*, Vol. 4, No. 4, Pp. 955–962, 2020, Doi: 10.30865/Mib.V4i4.2370.

[9] E. Oktavia, D. Saepudin, And A. A. Rohmawati, "Reduksi Dimensi Pada Optimasi Portofolio Mean-Variance Menggunakan Non-Negative Principal Component Analysis," *E-Proceeding Eng.*, Vol. 6, No. 2, Pp. 9978–9985, 2019.

[10] R. Susetyoko, "Reduksi Dimensi Menggunakan Komponen Utama Data Partisi Pada Pengklasifikasian Data Berdimensi Tinggi Dengan Ukuran Sampel Kecil," *Industrial Electronic Seminar*, 2010.

[11] G. Rahayu And Mustakim, "Principal Component Analysis Untuk Dimensi Reduksi Data Clustering Sebagai Pemetaan Persentase Sertifikasi Guru Di Indonesia," *Semin. Nas. Teknol. Inf. Komun. Dan Ind.*, Vol. 0, No. 0, Pp. 201–208, 2017, [Online]. Available: Http://Ejournal.Uin-Suska.Ac.Id/Index.Php/Sntiki/Article/View/3265.

[12] R. N. Azizah, "Pengenalan Wajah Dengan Metode Subspace Lda (Linear Discriminant Analysis)," *Proceeding Semin. Tugas Akhir Jur. Tek. Elektro FTI-ITS*, No. 6, Pp. 1–6, 2018.

[13] D. M. Chan, R. Rao, F. Huang, And J. F. Canny, "T-Sne-Cuda: Gpu-Accelerated T-Sne And Its Applications To Modern Data," *Proc. - 2018 30th Int. Symp. Comput. Archit. High Perform. Comput. Sbac-Pad 2018*, Pp. 330–338, 2019, Doi: 10.1109/Cahpc.2018.8645912.

[14] B. Utomo, "Dekomposisi Nilai Singular Pada Sistem Pengenalan Wajah," *J. Mat.*, Vol. 2, No. 1, Pp. 31–43, 2012.

[15] G. Walundungo, M. Paendong, And T. Manurung, "Penggunaan Analisis Multidimensional Scaling Untuk Mengetahui Kemiripan Rumah Makan Di Manado Town Square Berdasarkan Kerakteristik Pelanggan," *D'cartesian*, Vol. 3, No. 1, P. 30, 2014, Doi: 10.35799/Dc.3.1.2014.3806.

[16] I. M. Parapat, M. T. Furqon, And Sutrisno, "Penerapan Metode Support Vector Machine (Svm) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, Vol. 2, No. 10, Pp. 3163–

3169, 2018, [Online]. Available: Https://J-Ptiik.Ub.Ac.Id/Index.Php/J-Ptiik/Article/View/2577.

[17] D. I. Pushpita Anna Octaviani, Yuciana Wilandari, "Penerapan Metode Svm Pada Data Akreditasi Sekolah Dasar Di Kabupaten Magelang," *J. Gaussian*, Vol. 3, No. 8, Pp. 811–820, 2014.

[18] Bustami, "Penerapan Algoritma Naive Bayes," *J. Inform.*, Vol. 8, No. 1, Pp. 884–898, 2014.

[19] A. Primajaya And B. N. Sari, "Random Forest Algorithm For Prediction Of Precipitation," *Indones. J. Artif. Intell. Data Min.*, Vol. 1, No. 1, P. 27, 2018, Doi: 10.24014/Ijaidm.V1i1.4903.