



**SEKOLAH TINGGI MANAJEMEN INFORMATIKA & KOMPUTER
(STMIK) BANI SALEH BEKASI
UJIAN TENGAH SEMESTER GENAP
TAHUN AKADEMIK 2020/2021**

Mata Kuliah	: Data Mining	No. Absen	:
Prog Studi/Kelas	: S1/TI/06/AP, AM, BM	Nama Mahasiswa	:
Hari & Tanggal	: 24 April 2021	NPM	:
Nama Dosen	: Suhadi, ST., M.Kom	Tanda Tangan Mahasiswa	:
Waktu	: Online	Sifat Ujian	: All Open
Tanggal Verifikasi	:	Diverifikasi Oleh	:

Bismillahirrohmanirrohim

Petunjuk Soal:

1. Bila terdapat perintah “**Jelaskan**”, maka gunakan kalimat **Saudara** sendiri untuk menjelaskan
2. Kerjakan semua soal dibawah, mulai dari yang paling mudah menurut **Saudara**
3. Diperbolehkan menggunakan asumsi **Saudara** sendiri bila diperlukan

Soal:

1. Jelaskan pengertian metode data mining (a) *Estimation*, (b) *Forecasting*, (c) *Classification*, (d) *Clustering*, (e) *Association* dan contohkan masing-masing metode tersebut.
2. Buatkan analisis untuk model data mining menggunakan algoritma (a) *Multiple Linear Regression (MLR)*, (b) *Artificial Neural Network (ANN)*, (c) *Support Vector Machine (SVM)* dengan contoh kasus bebas **Saudara** pilih dalam bentuk grafik beserta deskripsinya.

Ketentuan:

UTS boleh Individu & Berkelompok (maksimal 2 orang)

Pengiriman UTS Terakhir Hari Senin, Tanggal 26 April 2021 Jam 23.59 WIB

Format Dalam Bentuk PDF

Dikirim Melalui E-Mail: hdivibst1@gmail.com

Format Subjet & Nama File: UTS_DM_NPM_2021

U*** Selamat Mengerjakan *****U**

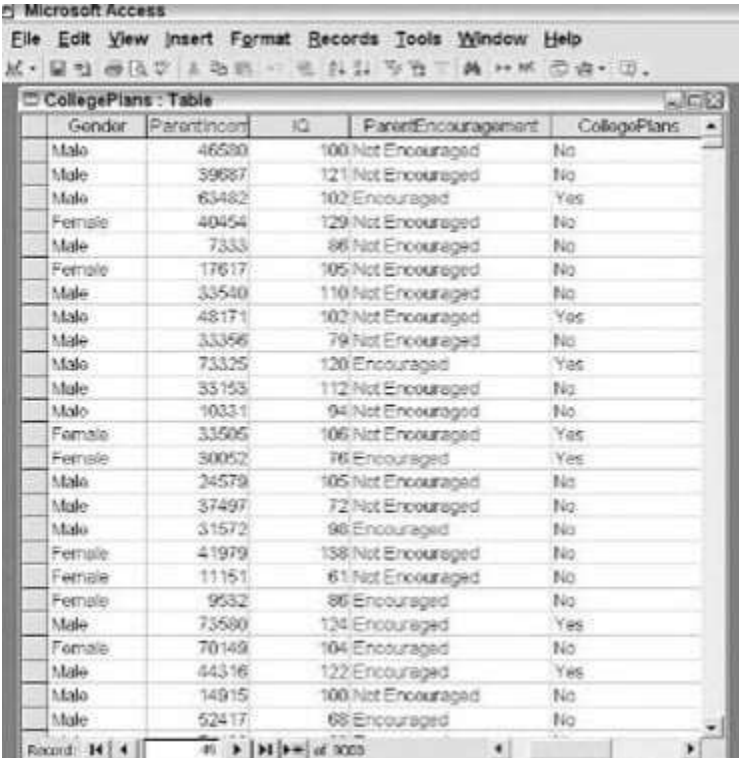
Jawaban UTS Data Mining

1. Pengertian metode data mining

1. Classification

Classification adalah metode yang paling umum pada data mining. Persoalan bisnis seperti Churn Analysis, dan Risk Management biasanya melibatkan metode Classification.

Classification adalah tindakan untuk memberikan kelompok pada setiap keadaan. Setiap keadaan berisi sekelompok atribut, salah satunya adalah class attribute. Metode ini butuh untuk menemukan sebuah model yang dapat menjelaskan class attribute itu sebagai fungsi dari input attribute.



Gender	ParentIncome	IQ	ParentEncouragement	CollegePlans
Male	46590	100	Not Encouraged	No
Male	39887	121	Not Encouraged	No
Male	63482	102	Encouraged	Yes
Female	40454	129	Not Encouraged	No
Male	7333	88	Not Encouraged	No
Female	17617	105	Not Encouraged	No
Male	33540	110	Not Encouraged	No
Male	48171	102	Not Encouraged	Yes
Male	33356	79	Not Encouraged	No
Male	73325	120	Encouraged	Yes
Male	33153	112	Not Encouraged	No
Male	10331	94	Not Encouraged	No
Female	33505	106	Not Encouraged	Yes
Female	30052	76	Encouraged	Yes
Male	24579	105	Not Encouraged	No
Male	37497	72	Not Encouraged	No
Male	31572	98	Encouraged	No
Female	41979	138	Not Encouraged	No
Female	11151	61	Not Encouraged	No
Female	9532	80	Encouraged	No
Male	73590	124	Encouraged	Yes
Female	70149	104	Encouraged	No
Male	44316	122	Encouraged	Yes
Male	14915	100	Not Encouraged	No
Male	52417	68	Encouraged	No

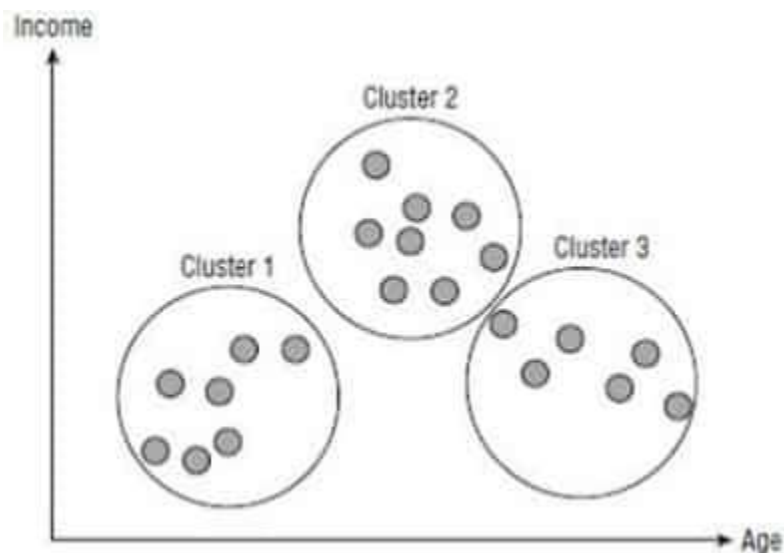
Class adalah attribute CollegePlans yang berisi dua pernyataan, Yes dan No, perhatikan ini. Sebuah Classification Model akan menggunakan atribut lain dari kasus tersebut (input attribut; yaitu kolom IQ, Gender, ParentIncome, dan ParentEncouragement) untuk dapat menentukan pola (pattern) class (Output Attribute; yaitu Kolom CollegePlans yang berisi Yes atau No).Algoritma Data Mining yang membutuhkan variabel target untuk belajar (sampai mendapatkan rule / pola yang berlaku pada data tersebut) kita standarkan dengan sebutan dengan Supervised Algorithm.

Nah, yang termasuk kepada Classification Algorithm adalah Decision Trees, Neural Network dan Naives Bayes.

2. Clustering

Clustering juga disebut sebagai segmentation. Metoda ini digunakan untuk mengidentifikasi kelompok alami dari sebuah kasus yang di dasarnya pada sebuah kelompok atribut, mengelompokkan data yang memiliki kemiripan atribut.

Gambar dibawah ini menunjukkan kelompok data pelanggan sederhana yang berisi dua atribut, yaitu **Age** (Umur) dan **Income** (Pendapatan).



Algoritma Clustering mengelompokkan kelompok data kedalam tiga segment berdasarkan kedua atribut ini.

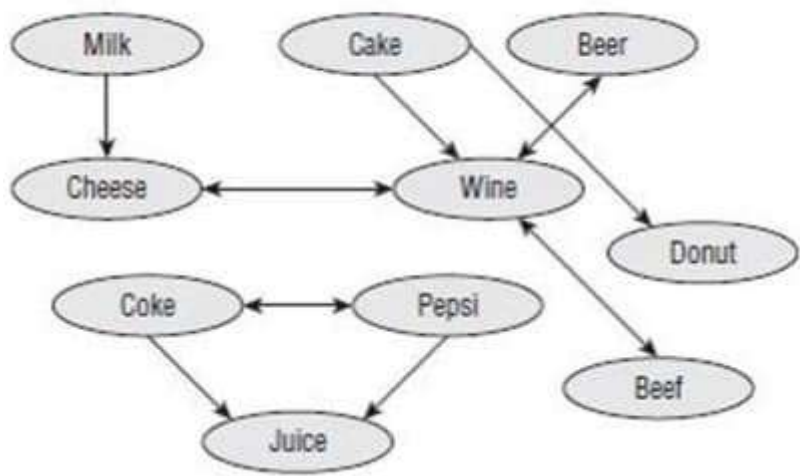
- Cluster 1 berisi populasi berusia muda dengan pendapatan rendah
- Cluster 2 berisi populasi berusia menengah dengan pendapatan yang lebih tinggi
- Cluster 3 berisi populasi berusia tua dengan pendapatan yang relatif rendah.

Clustering adalah metode data mining yang Unsupervised, karena tidak ada satu atributpun yang digunakan untuk memandu proses pembelajaran, jadi seluruh atribut input diperlakukan sama.

Kebanyakan Algoritma Clustering membangun sebuah model melalui serangkaian pengulangan dan berhenti ketika model tersebut telah memusat atau berkumpul (batasan dari segmentasi ini telah stabil).

3. Association

Association juga disebut sebagai **Market Basket Analysis**. Sebuah problem bisnis yang khas adalah menganalisa tabel transaksi penjualan dan mengidentifikasi produk-produk yang seringkali dibeli bersamaan oleh customer, misalnya apabila orang membeli sambal, biasanya juga dia membeli kecap. Kesamaan yang ada dari data pembelian digunakan untuk mengidentifikasi kelompok kesamaan dari produk dan kebiasaan apa yang terjadi guna kepentingan cross-selling seperti gambar dibawah ini.



Anda bisa lihat disini, beberapa hal dapat kita baca, misalnya :

- Ketika orang membeli susu, dia biasanya membeli keju
- Ketika orang membeli pepsi atau coke, biasanya dia membeli juice

Didalam istilah association, setiap item dipertimbangkan sebagai informasi. **Metode association memiliki dua tujuan:**

1. Untuk mencari produk apa yang biasanya terjual bersamaan
2. Untuk mencari tahu apa aturan yang menyebabkan kesamaan tersebut.

4. Forecasting

Forecasting juga adalah metode data mining yang sangat penting. Contohnya digunakan untuk menjawab pertanyaan seperti berikut:

- Seperti apa jadinya nilai saham dari Microsoft Corporation (pada NASDAQ, disimbolkan sebagai MSFT) pada keesokan hari?
- Sebanyak apa penjualan produk tertentu pada bulan depan?

Teknik Forecasting dapat membantu menjawab pertanyaan-pertanyaan diatas. Sebagai inputnya teknik Forecasting akan mengambil sederetan angka yang menunjukkan nilai yang berjalan seiring waktu dan kemudian Teknik Forecasting ini akan menghubungkan nilai masa depan dengan menggunakan bermacam-macam teknik machine-learning dan teknik statistik yang berhubungan dengan musim, trend, dan noise pada data.

Gambaranya dapat anda lihat sebagai berikut:



Gambar diatas menunjukkan dua kurva, garis yang tegas adalah time-series data sebenarnya dari nilai saham Microsoft, dan garis putus-putus adalah time series model yang memprediksi nilai saham berdasarkan nilai saham pada masa lalu.

5. Estimation

variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi.

2.Analisis untuk model data mining menggunakan algoritma (a) Multiple Linear Regression (MLR), (b) Artificial Neural Network (ANN), (c) Support Vector Machine (SVM) dengan contoh kasus bebas Saudara pilih dalam bentuk grafik beserta deskripsinya.

Pengujian

(a) Multiple Linear Regression (MLR)

Tabel 5 Data Pengujian		
Tgl	X ₁	X ₂
Jan-2014	26	33.3
Feb-2014	21	44.9
Mar-2014	24	49.4
Apr-2014	21	56.4
Mei-2014	27	62.3
Jun-2014	20	50.9
Jul-2014	12	64.2
Agt-2014	12	18

Maka berdasarkan model *multiple linear regression* nya, yaitu $\Sigma Y = 76.81641 + 11.74574 \Sigma X_1 - 1.70263 \Sigma X_2$ di dapatlah hasil curah hujan sebagai berikut :

Tabel 6 Hasil Pengujian						
Tgl	X ₁	X ₂	Y prediksi	Y kenyataan	Selisih (e _i)	Kuadrat selisih (e ²)
Jan- 2014	26	33.3	325.508	443	117.49	13804.38
Feb- 2014	21	44.9	247.028 7	220	-27.03	730.55
Mar- 2014	24	49.4	274.604 1	332	57.40	3294.29
Apr- 2014	21	56.4	227.448 4	223	-4.45	19.79
Mei- 2014	27	62.3	287.877 3	156	131.88	17391.63
Jun- 2014	20	50.9	225.067 2	221	-4.07	16.54
Jul- 2014	12	64.2	108.456 2	113	4.54	20.65
Agt- 2014	12	18	187.117 9	53	134.12	17987.61
TOTAL						53265.43

(a) Multiple Linear Regression (MLR)

Hasil pengujian model multiple linear regression dengan data curah hujan real menunjukkan adanya selisih. Hal ini memperlihatkan adanya error. Dari kuadrat selisih pada table diatas, dapat diketahui kesalahan baku (standart error) regresi adalah 22,52. Ini artinya besarnya penyimpangan atau ketidakakuratan nilai dugaan terhadap nilai sebenarnya adalah 22,52.

(b) Artificial Neural Network (ANN)

Artificial Neural Network Artificial (ANN) atau Jaringan Syaraf Tiruan merupakan sebuah teknik atau pendekatan pengolahan informasi yang terinspirasi oleh cara kerja sistem saraf biologis, khususnya pada sel otak manusia dalam memproses informasi. Elemen kunci dari teknik ini adalah struktur sistem pengolahan informasi yang bersifat unik dan beragam untuk tiap aplikasi.

Tabel 3.Analisis Korelasi terhadap Harga Kontrak Pekerjaan Struktur

Variabel Input	Nilai Kendals - Tau				Nilai Spearman			
	Koefisien Korelasi	Keterangan	Nilai Signifikan	Keterangan	Koefisien Korelasi	Keterangan	Nilai Signifikan	Keterangan
Luas Bangunan	0.929	Hub. Kuat (mendekati 1)	0.001	Ada Hubungan (Kurang dari 0.01)	0.976	Hub. Kuat (mendekati 1)	0	Ada Hubungan (Kurang dari 0.01)
Bentang Kolom Rata-Rata	0.929	Hub. Kuat (mendekati 1)	0.001	Ada Hubungan (Kurang dari 0.01)	0.976	Hub. Kuat (mendekati 1)	0	Ada Hubungan (Kurang dari 0.01)
Prosentase Pekerjaan Struktur Tanah dan Pondasi	0.571	Hub. Kuat (mendekati 1)	0.48	Tidak Ada Hubungan (lebih dari 0.05)	0.738	Hub. Kuat (mendekati 1)	0.37	Tidak Ada Hubungan (lebih dari 0.05)
Prosentase Pekerjaan Beton dan Pelat Lantai	0.143	Hub.Lemah (mendekati 0)	0.621	Tidak Ada Hubungan (lebih dari 0.05)	0.167	Hub.Lemah (mendekati 0)	0.693	Tidak Ada Hubungan (lebih dari 0.05)
Prosentase Pekerjaan Struktur Atap	0.357	Hub.Lemah (mendekati 0)	0.316	Tidak Ada Hubungan (lebih dari 0.05)	0.595	Hub. Kuat (mendekati 1)	0.12	Tidak Ada Hubungan (lebih dari 0.05)
IKK	0.357	Hub.Lemah (mendekati 0)	0.316	Tidak Ada Hubungan (lebih dari 0.05)	0.524	Hub. Kuat (mendekati 1)	0.183	Tidak Ada Hubungan (lebih dari 0.05)
Jarak Lokasi Proyek dengan Pusat Kota	0.214	Hub.Lemah (mendekati 0)	0.458	Tidak Ada Hubungan (lebih dari 0.05)	0.31	Hub.Lemah (mendekati 0)	0.456	Tidak Ada Hubungan (lebih dari 0.05)

*, Correlation is significant at the 0.05 level (2-tailed).
**, Correlation is significant at the 0.01 level (2-tailed)

SIMPULAN

Berdasarkan hasil analisis korelasi KendallTau dan Spearman yang telah dilakukan, dapat diidentifikasi faktor-faktor yang berpengaruh signifikan terhadap biaya pekerjaan struktural pembangunan konstruksi gedung Rumah Sakit Pratama (2015) antara lain: (a) Luas bangunan, (b) Bentang kolom rata-rata, dan (c) Prosentase pekerjaan struktur tanah dan pondasi Nilai IKK (Indeks Kemahalan Konstruksi) dan Jarak Lokasi Proyek dari pusat ibukota provinsi tidak memiliki nilai pengaruh sensitivitas yang tinggi terhadap besarnya nilai kontrak konstruksi Rumah Sakit Pratama (2015). Dari rumus empiris yang diperoleh dari metode ANN didapatkan prosentase error/ MMRE maksimal yang dihasilkan adalah sebesar 0.139%.

(c) Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah suatu teknik untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi (Santosa, 2007). SVM memiliki prinsip dasar linier classifier yaitu kasus klasifikasi yang secara linier dapat dipisahkan, namun SVM telah dikembangkan agar dapat bekerja pada problem non-linier dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi.

Analisis Data

Klasifikasi SVM Multikelas SLA dengan fungsi kernel Gaussian RBF menggunakan 5 nilai parameter σ sebagai analisis $\sigma = 1, \sigma = 2, \sigma = 3, \sigma = 4, \sigma = 5$ dengan nilai parameter $C = 1, C = 5, C = 10, C = 50, C = 100$. Hasil akurasi klasifikasi kernel Gaussian RBF pada data training dapat dilihat pada Tabel 2. Terlihat bahwa model yang dibentuk data training dapat digunakan untuk memprediksi kelas data training itu sendiri dengan akurasi sebesar 100% yang berarti kelas dapat diklasifikasikan tepat sesuai dengan kelas aslinya dengan error sebesar 0%. Berikut adalah tabel hasil akurasi untuk prediksi kelas pada data training

Tabel 2. Hasil Akurasi Klasifikasi Kernel Gaussian RBF pada Data Training

C	RBF					
	$\sigma=1$	$\sigma=2$	$\sigma=3$	$\sigma=4$	$\sigma=5$	$\sigma=6$
1	100	100	100	100	100	100
5	100	100	100	100	100	100
10	100	100	100	100	100	100
50	100	100	100	100	100	100
100	100	100	100	100	100	100

Selanjutnya untuk hasil akurasi klasifikasi kelas menggunakan data testing adalah sebagai berikut

Tabel 3. Hasil Akurasi Klasifikasi SVM Kernel Gaussian RBF pada Data Testing

C	RBF				
	$\sigma=1$	$\sigma=2$	$\sigma=3$	$\sigma=4$	$\sigma=5$
1	91.463	92.683	93.902	92.683	93.902
5	91.463	92.683	93.902	93.902	93.902
10	91.463	92.683	93.902	93.902	93.902
50	91.463	92.683	93.902	93.902	93.902
100	91.463	92.683	93.902	93.902	93.902

Tampak bahwa akurasi mencapai nilai maksimal saat $\sigma = 3$, sehingga prediksi terbaik yaitu dipilih nilai parameter $\sigma = 3$ untuk nilai $C = 1$, yaitu nilai C paling kecil dengan akurasi sebesar 93.902 %. Berikut adalah matriks konfusi Kernel Gaussian RBF $\sigma = 3, C = 1$ pada data testing

Tabel 4. Matriks Konfusi Kernel Gaussian RBF $\sigma = 3, C = 1$ pada Data Testing

F_{gh}		Kelas Hasil Prediksi (h)		
		Kelas 1	Kelas 2	Kelas 3
Kelas Asli (g)	Kelas 1	12	4	0
	Kelas 2	1	65	0
	Kelas 3	0	0	0

Tabel 5. Hasil Akurasi Klasifikasi SVM Kernel Polynomial pada Data Training

C	Polynomial	
	d=1	d=2
1	79.228	98.220
5	89.021	98.810
10	93.472	98.810
50	96.736	98.810
100	96.736	98.810

Terlihat pada Tabel 5. bahwa akurasi terbaik untuk pengklasifikasian data training adalah sebesar 98.810 % untuk parameter $d = 2$ dan $C = 5$. Berikut hasil akurasi klasifikasi pada data testing menggunakan fungsi kernel polynomial.

Tabel 6. Hasil Akurasi Klasifikasi SVM Kernel Polynomial pada Data Testing

C	Polynomial	
	d=1	d=2
1	80.488	85.366
5	87.805	84.146
10	91.463	78.049
50	92.683	84.146
100	92.683	84.146

Berdasarkan Tabel 6, terlihat bahwa klasifikasi terbaik menggunakan fungsi kernel polynomial adalah memakai $d = 1$ dan $C = 50$ dengan tingkat akurasi klasifikasi terbesar. Untuk mengetahui jumlah kelas yang diprediksi secara benar atau secara salah, dapat dilihat dari matriks konfusi berikut ini

Tabel 7. Matriks Konfusi Kernel Polynomial $d = 1, C = 50$ pada Data Testing

F_{gh}		Kelas Hasil Prediksi (h)		
		Kelas 1	Kelas 2	Kelas 3
Kelas Asli (g)	Kelas 1	13	3	0
	Kelas 2	3	63	0
	Kelas 3	0	0	0

Berdasarkan Tabel 7, diperoleh akurasi klasifikasi sebesar 92.682 % dan error klasifikasi sebesar 7.317 %. Artinya dengan akurasi 92.682 % sebanyak 76 SD dari 82 SD dapat diklasifikasikan secara benar sesuai dengan kelas asli, sedangkan sisanya sebanyak 6 SD diklasifikasikan berbeda dengan kelas asli.

Kesimpulan

Berdasarkan analisis yang telah dilakukan, diperoleh kesimpulan yaitu

1. Klasifikasi Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang menggunakan metode Support Vector Machine (SVM) dengan data training yang diujikan sebanyak 337 data memiliki akurasi klasifikasi sebesar 100% menggunakan fungsi kernel Gaussian Radial Basic JURNAL GAUSSIAN Vol. 3, No. 4, Tahun 2014 Halaman 820 Function (RBF). Sedangkan menggunakan fungsi kernel Polynomial akurasi klasifikasi adalah sebesar 98.810 %.
2. Pada pengujian dengan data testing sebanyak 82 data, akurasi klasifikasi yang didapat yaitu sebesar 93.902% menggunakan kernel Gaussian Radial Basic Function (RBF). Sedangkan menggunakan fungsi kernel Polynomial akurasi klasifikasinya adalah sebesar 92.683 % .
3. Dengan demikian akurasi klasifikasi terbaik yaitu menggunakan fungsi kernel Gaussian Radial Basic Function (RBF) karena menghasilkan akurasi yang lebih besar, yaitu dapat mengklasifikasikan 337 SD dari 337 SD secara benar sesuai dengan kelas asli pada data training. Sedangkan pada data testing dapat mengklasifikasikan 77 SD dari 82 SD secara benar sesuai dengan kelas asli.