UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 2 Tahun 2024 Halaman 213-218
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

The Comparison of C4.5 and C5.0 Algorithms in Classifying the Nutritional Status of Stunted Toddlers

Dhea Afrilia Harelvi, Admi Salma*, Yenni Kurniawati, dan Fadhilah Fitri
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: admisalma1@fmipa.unp.ac.id

Submitted : 21 Mei 2024
Revised : 31 Mei 2024
Accepted : 31 Mei2024

ABSTRACT

Stunting is one of the health conditions that reflect aspects of nutrition and child growth, allowing us to observe the nutritional status of toddlers. The aim of this study is to determine the classification results of the C4.5 and C5.0 algorithms in cases of stunted toddler nutritional status and to compare the results between the C4.5 and C5.0 algorithms in classifying stunted toddler nutritional status using k-fold cross-validation. The data in this study are secondary data. Which is collected from Puskesmas IV Pesisir Selatan Regency. The research variables are divided into two, namely the response variable Y, which is Toddler Nutritional Status, and predictor variables X including Age, Toddler Gender, Toddler Weight, and Toddler Height. The result of the study obtain the algorithm C5.0 produse accuracy value of the C5.0 algorithm is higher than that of the C4.5 algorithm. The C5.0 algorithm provides an average accuracy result of 83% while the C4.5 algorithm provides an accuracy result of 79%. Thus, it can be concluded that the C5.0 algorithm is better at classifying stunted toddler nutritional status.

Keywords: *C4.5 Algorithm, C5.0 Algorithm, Classifying, Nutrional Status*

I. PENDAHULUAN

Data mining merupakan suatu proses analisis data yang menggunakan teknik matematika dan statistika, untuk mengidentifikasi pola atau pengetahuan yang bermanfaat dari kumpulan data besar. Salah satu teknik dari *data mining* adalah klasifikasi. Klasifikasi adalah proses mengelompokkan objek atau data ke dalam kategori atau kelas tertentu berdasarkan ciri-ciri dari atribut yang dimiliki oleh data tersebut. Salah satu metode klasifikasi yang paling populer digunakan yaitu *decision tree*. *Decision tree* adalah suatu model yang ditampilkan ke dalam bentuk pohon yang digunakan untuk membuat keputusan atau prediksi. Dalam pembentukkan *decision tree* ada beberapa algoritma yang dapat digunakan yaitu ID3, CART, C4.5, dan C5.0 (Kohavi, 1995 ; Prasetyo, 2012).

Algoritma yang diciptakan J. Ross Quinlan yaitu C4. 5, menjadi dasar pengembangan algoritma ID3. Kemampuan C4. 5 untuk menangani fitur tipe numerik dan kategorikal membedakannya dari algoritme ID3. Di sisi lain, ID3 terbatas pada fitur yang memiliki kategori kategoris. Algoritma C5.0 pada dasarnya sama dengan algoritma C4.5, dan merupakan pengembangan dari algoritma C4.5 sebelumnya. Berbeda dengan algoritma C4.5, algoritma C5.0 membutuhkan lebih sedikit langkah kerja. Algoritma C5.0 dapat menghasilkan prediksi dengan tingkat akurasi yang tinggi (Pandya, 2015 ; Prasetyo, 2014).

Klasifikasi banyak dimanfaatkan dalam bidang kesehatan untuk mengelompokkan informasi tentang penyakit. Salah satu permasalahan kesehatan yang dapat menggunakan klasifikasi yaitu status gizi balita *stunting*. Kecukupan gizi merupakan faktor kunci dalam meningkatkan kualitas sumber daya manusia dan indikator keberhasilan pembangunan negara. Kecukupan gizi berdampak signifikan pada kecerdasan dan produktivitas kerja. Memahami status gizi yang mencerminkan keseimbangan antara asupan dan kebutuhan nutrisi untuk fungsi metabolisme tubuh sangatlah penting untuk memenuhi kecukupan gizi. Balita adalah kelompok usia yang rentan mengalami masalah gizi, yang tercermin dalam kesehatan dan pertumbuhan mereka. Salah satu masalah gizi yang terjadi pada balita yang tinggi di Indonesia adalah *stunting*, yaitu kondisi di mana tinggi badan anak lebih pendek dari standar usia, mencerminkan kekurangan gizi kronis (Almatsier, 2001 ; Proverawati, 2009).

Survei Status Gizi Indonesia (SSGI) 2022 menunjukkan bahwa 25,20% balita di Sumatera Barat mengalami stunting, naik 1,9% dari 2021. Salah satu daerah yang memiliki angka stunting tertinggi di Sumatera Barat terdapat pada Kabupaten Pesisir Selatan, dimana mencapai 29,80% pada 2022. Untuk menurunkan angka ini, pemerintah daerah meluncurkan Program Bapak/Bunda Asuh Stunting (BAAS). Dalam upaya mendukung program tersebut, klasifikasi

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/172

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

status gizi balita *stunting* dengan algoritma *decision tree* perlu dilakukan untuk memperoleh informasi balita *stunting* di Kabupaten Pesisir Selatan. Algoritma *decision tree* dipilih karena kemampuannya untuk menghasilkan model yang mudah dipahami dan diinterpretasikan, yang dapat membantu dalam mengidentifikasi variabel-variabel penting yang mempengaruhi status gizi balita. Selain itu, *decision tree* mampu menangani data dengan baik bahkan jika terdapat hubungan non-linear antara variabel-variabel independen dan status gizi balita. (Kemenkes, 2022).

Penelitian oleh Putri dkk (2016) yang menyatakan bahwa algoritma C4.5 menghasilkan akurasi yang lebih tinggi, namun penelitian yang dilakukan oleh Dewi dkk (2019) menyatakan bahwa algortima C5.0 memberikan nilai akurasi yang lebih tinggi. Karena hasil dari penelitian sebelumnya yang tidak konsisten, maka pada penelitian ini akan dilakukan perbandingan algortima C4.5 dengan algoritma C5.0 menggunakan metode *k-fold cross validation* untuk melihat metode mana yang baik dalam menangani kasus status gizi balita *stunting*. *K-fold cross* validation merupakan salah satu jenis pengujian cross *validation* yang digunakan untuk mengevaluasi kinerja suatu metode algoritma. Metode ini bekerja dengan melakukan pembagian acak pada sampel data dan mengelompokkan data tersebut ke dalam *K* kelompok.

II. METODE PENELITIAN

A. Sumber Data dan Variabel Penelitian

Data yang digunakan merupakan data sekunder berupa rekapitulasi data status gizi balita di Puskesmas IV Koto Mudik Kabupaten Pesisir Selatan Tahun 2022. Variabel penelitian yang digunakan dalam kasus ini terdiri dari variabel respon Y sebanyak 1 variabel dan variabel prediktor X sebanyak 4 variabel dengan uraian pada Tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Keterangan
Usia (X1)	Bulan
Jenis Kelamin Balita (X2)	Kategori : 0 : Perempuan 1 : Laki - Laki
Berat Badan Balita (X3)	Kilogram
Tinggi Badan Balita (X4)	Centimeter
Status Gizi Balita TB/U (Y)	Kategori : Stunting Tidak Stunting

Berdasarkan Tabel 1 variabel pada penelitian ini mencakup usia balita(X1) yang diukur dalam bulan, jenis kelamin balita(X2) yang dicatat dengan kode kategori 0 untuk perempuan dan 1 untuk laki-laki, berat badan balita(X3) yang diukur dalam kilogram, tinggi badan balita(X4) yang diukur dalam sentimeter, serta status gizi balita(Y) berdasarkan indeks TB/U yang terkategori sebagai *stunting* atau tidak *stunting*.

B. Teknik Analisis Data

Penelitian ini menggunakan metode klasifikasi dengan menggunakan algoritma *decision tree* algoritma C4.5 dan algoritma C5.0 melalui tahapan sebagai berikut:

1. Melakukan input data.
2. Memisahkan data *training* dan data *testing* menggunakan metode uji validasi *k-fold cross validation* dengan jumlah *k* sebesar 10. *k-fold cross validation* mengatur data secara acak ke dalam *k* kelompok, di mana setiap kelompok kemudian dibagi menjadi data *training* dan data *testing*. Proses ini diulang sebanyak *k* kali, dengan satu kelompok ditinggalkan sebagai data uji pada setiap iterasi (*fold*). *k-fold cross validation* menggunakan kembali dataset yang sama, sehingga menghasilkan *k* partisi data yang tidak memiliki elemen yang sama pada setiap iterasi. (Raschka, 2018)
3. Algortima C4.5
   a. Menghitung nilai *entropy* setiap atribut pada data *traning* menggunakan persamaan (1).

$$Entropy(S) = \sum_{i=1}^{n} -\frac{|S_i|}{|S|} \times log_2 \frac{|S_i|}{|S|} \quad (1)$$

Dimana :
$|S_i|$: Total atribut dengan kelas ke i
$|S|$: Total himpunan kasus S

   b. Menghitung nilai *gain* setiap atribut pada data *traning* menggunakan persamaan (2).

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/172

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$Gain\ (S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} \times Entropy(S_i) \quad (2)$$

Dimana :
$|S_i|$: Proporsi S_i terhadap S
$|S|$: Jumlah kasus dalam
A : Atribut/variabel
Entropy (S_i) : *Entropy* untuk kelas ke-i

c. Menghitung nilai *splitinfo* setiap atribut pada data *traning* menggunakan persamaan (3).

$$SplitInfo\ (S,A) = -\sum_{i=1}^{n} \frac{S_i}{S} log_2 \frac{S_i}{S} \quad (3)$$

Dimana :
n : Total kelas atribut A dengan himpunan kasus S
S_i : Total atribut A dengan kelas ke-i
S : Total himpunan kasus S

d. Menghitung nilai *gain ratio* pada data *traning* setiap atribut menggunakan persamaan (4).

$$GainRatio\ (S,A) = \frac{Gain\ (S,A)}{SplitInfo\ (S,A)} \quad (4)$$

(Prasetyo, 2014)

e. Menentukan *gain ratio* akar dan cabang pohon berdasarkan nilai *gain ratio* tertinggi.
f. Ulangi langkah a sampai e sampai semua atribut telah digunakan, sampai terbentuk pohon klasifikasi.
g. Melihat akurasi dengan menggunakan *k-fold cross validation* dan menghitung kesalahan prediksi pada iterasi ke-j untuk $j = 1,2, \dots, k$ dan nilai I yaitu *miss clasification rate*, *miss clasification rate* atau tingkat kesalahan klasifikasi pada *k-fold cross-validation* adalah pengukuran evaluasi yang mengukur sejauh mana model klasifikasi melakukan kesalahan dalam memprediksi kelas target yang benar dengan nilai I yaitu $y_i \neq \hat{y}_i$ atau data asli ≠ data prediksi. Rata - rata atau prediksi galat pada kelompok dapat dilihat pada persamaan 6:

$$\bar{E}_j = \frac{\sum_{i=1}^{n_{uji(j)}} I(y_i \neq \hat{y}_i)}{n} \quad (6)$$

Keterangan :
I : *Miss clasification rate*
n : Jumlah pengamatan data uji pada iterasi ke-j
$\bar{E}_j$: Prediksi kesalahan pada iterasi ke-j
$n_{uji(j)}$: Jumlah observasi data uji pada iterasi ke-j

Sebagaimana persamaan (6) maka dihasilkan rumus untuk menghitung kesalahan prediksi menggunakan metode *k-fold cross validation* sebagai berikut:

$$\hat{E}^{CV} = \frac{\sum_{i=1}^{k} \bar{E}_j}{k} \quad (7)$$

(Wood, 2007)

Keterangan:
$\hat{E}^{CV}$: Perdiksi galat dengan metode *k-fold cross validation*
k : Jumlah kelompok data

4. Algoritma C5.0
a. Menghitung nilai *entropy* setiap atribut menggunakan persamaan (1).
b. Menghitung nilai *gain* setiap atribut menggunakan persamaan (2).
c. Menghitung nilai *gain ratio* setiap atribut pada data *traning* menggunakan persamaan (5).

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/172

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

$$GainRatio\ (S,A) = \frac{Gain\ (S,A)}{\sum_{i=1}^{n} Entropy\ (S_i)} \quad (8)$$

d. Menentukan *gain ratio* akar dan cabang pohon berdasarkan nilai *gain ratio* tertinggi.
e. Ulangi langkah a sampai d sampai semua atribut telah digunakan, sampai terbentuk pohon klasifikasi.
f. Melihat akurasi dengan menggunakan *k-fold cross validation.* Dengan menggunakan persamaan (6) dan (7).

5. Membandingkan Algoritma C4.5 dan Algoritma C5.0 menggunakan *k-fold cross validation.*
6. Mengambil keputusan.

III. HASIL DAN PEMBAHASAN

Hasil akurasi perbandingan klasifikasi pada algoritma C4.5 dan algoritma C5.0 dapat dihitung melalui *kfold cross validation* jumlah nilai k sebanyak 10. Tabel 2 merupakan hasil perbandingan ketepatan klasifikasi algoritma C4.5 dan algoritma C5.0 dalam mengklasifikasikan status gizi balita *stunting.*

Tabel 2. Hasil Klasifikasi Pada Algoritma C4.5 dan Algoritma C5.0

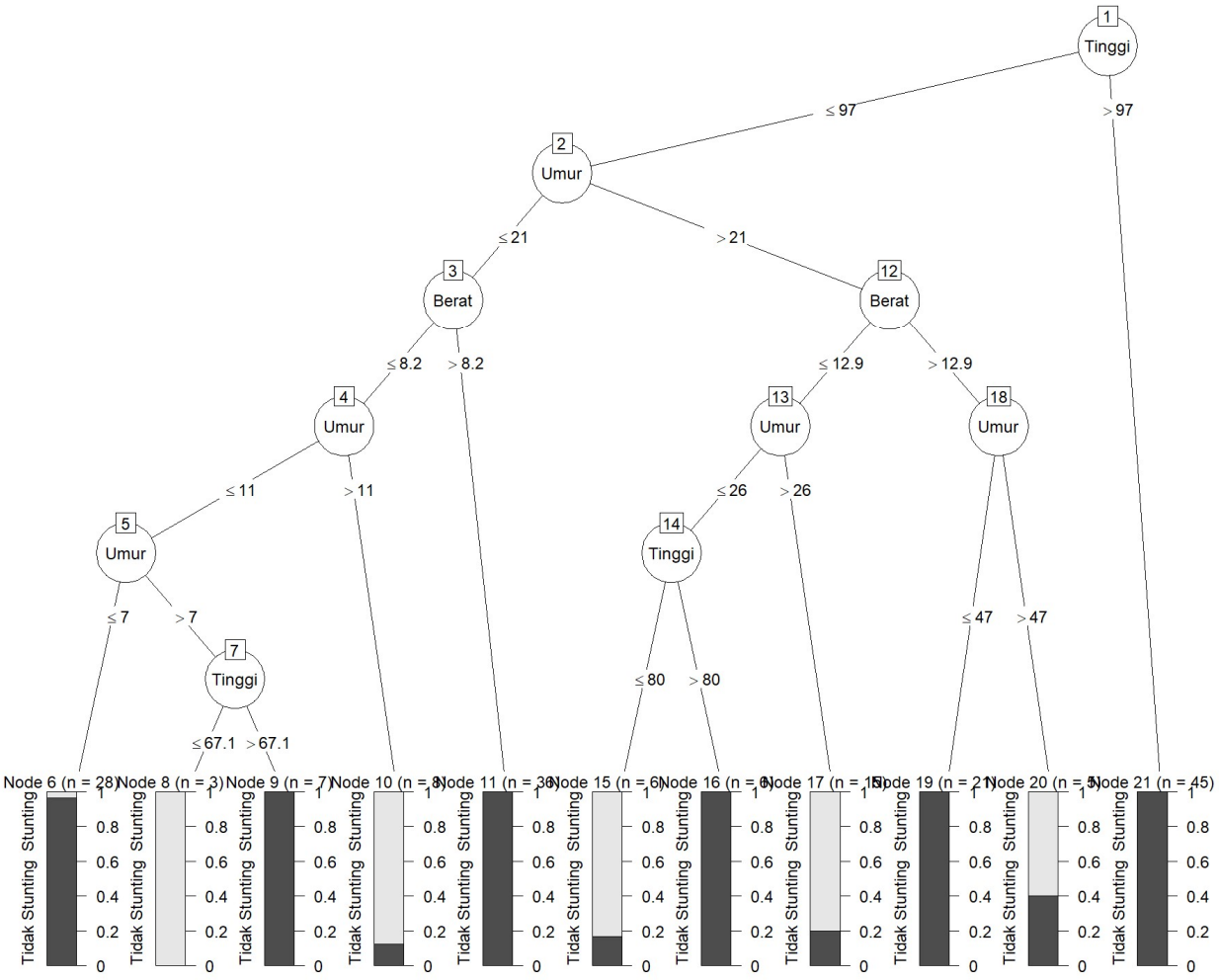
Pengujian	Algoritma C4.5	Algoritma C5.0
K1	85%	90%
K2	75%	75%
K3	80%	65%
K4	70%	80%
K5	75%	90%
K6	85%	90%
K7	100%	100%
K8	85%	85%
K9	95%	100%
K10	40%	55%
Mean	**79%**	**83%**

Dari hasil analisis yang telah dilakukan berdasarkan Gambar 2, perbandingan hasil klasifikasi antara algoritma C4.5 dan C5.0 menunjukkan bahwa akurasi algoritma C5.0 secara signifikan lebih tinggi daripada algoritma C4.5 dalam kasus balita *stunting*. Algoritma C5.0 memberikan tingkat akurasi sebesar 83%, sementara algoritma C4.5 hanya memberikan tingkat akurasi sebesar 79%. Dari hasil analisis didapatkan bahwa model terbaik untuk mengklasifikasikan status gizi balita *stunting* adalah algoritma C5.0.

Algortima C5.0 menghitung dan mengklasifikasi sebanyak 200 data balita yang direkap oleh Puskesmas IV Koto Mudik Kabupaten Pesisir Selatan tahun 2022. Analisis algoritma C5.0 menggunakan *software R studio* dimana data *training* dan data *testing* dibagi menggunakan *k fold cross validation* dengan nilai k = 10. Data dibagi sebanyak nilai k = 10, lalu setiap nilai k akan melakukan analisis dan memberikan akurasi sebanyak 10. Setelah data *training* dan data *testing* dibagi lalu dalam pembentukan model dilakukan dengan meneliti nilai *entropy, gain,* dan *gain ratio* yang ada pada persamaan 1,2 dan 8, nilai *gain ratio* tertinggi akan menjadi akar pada pohon keputusan. Perhitungan akan terus dilakukan sampai semua simpul dari pohon sudah memiliki keputasan atau simpul sudah tidak mempunyai cabang. Setelah semua keputusan sudah didapatkan, selanjutnya adalah melakukan pengujian dari algoritma C5.0 terhadap data *testing* menggunakan *k-fold cross validation*. Pada Gambar 1 menampilkan hasil pohon keputusan menggunakan algoritma C5.0.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/172

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X



Gambar 1. Hasil Algoritma C5.0

Berdasarkan Gambar 1 hasil pohon keputusan menunjukkan ada 1 simpul cabang yaitu variabel tinggi, 9 simpul internal dan 11 simpul daun yang merupakan nilai kelas atau hasil keputusan. Berbeda dengan algortima C4.5 dimana variabel berat menjadi simpul akar. Pada algoritma C5.0 variabel tinggi merupakan simpul akar dari pohon keputusan. Hasil dari pohon keputusan algoritma C5.0 adalah balita yang memiliki tinggi besar dari 97 Cm maka tidak berisiko *stunting*, balita yang berumur besar dari 7 bulan memiliki berat kurang sama dari 8.2 kg dan memiliki tinggi badan kurang sama dari 67.1 cm maka berpotensi *stunting.* Balita yang memili tinggi kurang sama dari 97 cm memiliki berat badan kurang sama 12.9 kg maka berpotensi *stunting.* Dapat disimpulkan pada algoritma C5.0 variabel tinggi badan, berat badan, dan umur balita berpengaruh terhadap klasifikasi status gizi balita *stunting*.

IV. KESIMPULAN

Berdasarkan hasil dan pembahasan dari klasifikasi status gizi balita *stunting* menggunakan algoritma C4.5 dan C5.0 dengan metode evaluasi *k-fold cross-validation*, didapatkan bahwa algoritma C5.0 menghasilkan akurasi klasifikasi yang lebih tinggi dibandingkan algoritma C4.5. Algoritma C5.0 memberikan akurasi rata-rata sebesar 83%, sedangkan algoritma C4.5 memberikan akurasi rata-rata sebesar 79%. Dapat disimpulkan bahwa algoritma C5.0 lebih efektif dalam mengklasifikasikan status gizi balita *stunting*. Saran untuk penelitian selanjutnya, menambahkan variabel prediktor tambahan yang mungkin berkontribusi pada akurasi klasifikasi seperti menambahkan variabel status sosial ekonomi, tingkat pendidikan orang tua, dan kondisi lingkungan.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/172

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

DAFTAR PUSTAKA

Almatsier, S. (2001). *Prinsip Dasar Ilmu Gizi*. Gramedia Pustaka Utama: Jakarta.

Dewi, D. A. W., Cholissodin, I., & Sutrisno, S. (2019). Klasifikasi Penyimpangan Tumbuh Kembang Anak Menggunakan Algoritme C5. 0. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *3*(10), 10258-10265.

Kemenkes, R. I. (2022). Buku Saku Hasil Studi Status Gizi Indonesia (SSGI) Tingkat Nasional, Provinsi, Dan Kabupaten/Kota Tahun 2021. *Kemenkes RI. Jakarta. Retrieved from https://www. litbang. kemkes. go. id/buku-saku-hasil-studi-status-gizi-indonesia-ssgi-tahun-2021*.

Kohavi, R. (1995). A Study Of Cross-Validation And Bootstrap For Accuracy Estimation And Model Selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Ilanda, Y., Vionanda, D., Kurniawati, Y., & Fitria, D. (2023). Perbandingan Metode Prediksi Laju Galat dalam Pemodelan Klasifikasi Algoritma C4. 5 untuk Data Tidak Seimbang. *UNP Journal of Statistics and Data Science*, 1(4), 240-247.

Pandya, R., dkk .(2015). *C5.0* Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications*, vol. 117, no. 16, pp. 18-21.

Prasetyo, E. (2014). *Data Mining: Mengolah Data Menjadi Informasi Menggunakan Matlab*.CV. Andi Offset: Yogyakarta.

Proverawati. (2009). *Gizi Untuk Kebidan*an. Nuha Medika: Yogyakarta.

Putri, Y. R., Mukhlash, I., & Hidayat, N. (2016). Prediksi pola kecelakaan kerja pada perusahaan non ekstraktif menggunakan algoritma decision tree: C4. 5 dan C5. 0. *Jurnal Sains Dan Seni Pomits*, 2(1), 1-6.

Raschka, Sebastian. (2018). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, *arXiv*:*1811.12808.*

Wood, I., Vixxcher, P., & Mengersen, K. (2007). Classification Based Upon Gene Expression Data: Bias and Precision of Error Rates. *Bioinformatics,* 23(11), 1363-1370.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/172