UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 2 Tahun 2024 Halaman 159-164.
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Handling Multiclass Imbalance in the Sample Area Sampling Frame Survey Dataset using the SCUT Method

Wilia Sondriva, Yenni Kurniawati*, Nonong Amalita, Admi Salma
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: yennikurniawati@fmipa.unp.ac.id

Submitted : 04 Mei 2024
Revised : 28 Mei 2024
Accepted : 29 Mei 2024

ABSTRACT

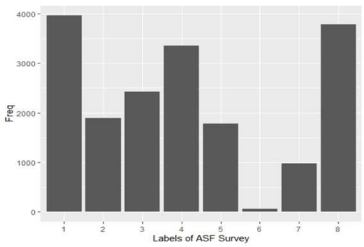
Area Sampling Frame (ASF) is a survey used by the Indonesian government to measure rice productivity in Indonesia. ASF survey is important data because accurate and high-quality rice productivity data is highly needed. There is extreme imbalance in the ASF survey data, thus requiring handling of this imbalance. SMOTE and Cluster-based Undersampling Technique (SCUT) is a method that can be used to address the dataset imbalance. SCUT combines oversampling using SMOTE and undersampling using CUT. The results from SCUT show that the number of data points in each class becomes balanced. Subsequently, a two-sample mean test is conducted to observe the mean differences between the original dataset and the dataset after handling. The results show that in the early vegetative, late vegetative, and harvest phases, the means are significantly similar between the original dataset and the dataset after handling, but in the generative phase, the means are not significantly similar. Therefore, synthetically generated data using the SCUT method generally exhibit similar mean characteristics.

Keywords: *Imbalance, Multiclass, SMOTE and Cluster-based Undersampling Technique (SCUT)*

I. PENDAHULUAN

Ketidakseimbangan *multiclass* adalah situasi dimana jumlah data dalam satu kelas secara signifikan lebih banyak atau lebih sedikit dibandingkan dengan kelas lainnya (Dharmawan, 2023). Kelas data yang amatannya lebih banyak disebut kelas mayoritas, sedangkan kelas yang amatannya lebih sedikit disebut kelas minoritas. Dalam konteks data yang mengalami ketidakseimbangan *multiclass*, proses pengolahan dan analisis data seperti klasifikasi, pengklusteran, prediksi, dan sebagainya menjadi lebih sulit dilakukan (Alqaida dkk, 2022). Oleh karena itu, penanganan ketidakseimbangan *multiclass* ini perlu dilakukan.

Penelitian oleh Kurniawati (2023) menyatakan bahwa adanya masalah ketidakseimbangan ekstrim pada data *multiclass* hasil pengamatan lahan padi dari survei Kerangka Sampel Area (KSA). KSA tersebut merupakan salah satu survei yang digunakan oleh pemerintah Indonesia untuk mengukur produktivitas padi di Indonesia (BPS, 2018). Data KSA adalah data penting karena data ketersediaan pangan yang berkualitas dan akurat sangat dibutuhkan. Pengamatan KSA mengelompokkan kondisi lahan tanaman padi menjadi delapan label KSA yakni fase vegetatif awal, vegetatif akhir, generatif, panen, persiapan lahan, puso, sawah bukan padi, dan bukan sawah (BPS, 2018). Ketidakseimbangan *multiclass* pada data survei KSA dapat dilihat pada Gambar1.



Sumber : Kurniawati (2023)

Gambar 1. Ilustrasi Ketidakseimbangan *Multiclass* Data Survei KSA

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/163

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

Berdasarkan Gambar 1 dapat terlihat bahwa label KSA memiliki jumlah data yang berbeda pada setiap label KSA. Terlihat bahwa label 1 (vegetatif awal) adalah label yang memiliki jumlah data paling banyak dan label 6 (puso) memiliki jumlah data paling sedikit. Namun, pada artikel ini hanya terfokus pada fase pertumbuhan padi yakni fase vegetatif awal, vegetatif akhir, generatif, dan panen. Hal ini dikarenakan keempat fase tersebut merupakan tahapan utama dalam siklus pertumbuhan padi. Fase vegetatif awal yakni ketika pada berumur 1-35 hari, fase vegetatif akhir yakni ketika padi berumur 35-55, fase generatif yakni ketika padi berumur 55-105 hari, dan fase panen yakni ketika padi sedang atau sudah dipanen (BPS, 2018). Pada keempat fase tersebut terdapat ketidakseimbangan *multiclass* pada masing-masing label dimana fase vegetatif awal memiliki jumlah data paling banyak dan fase vegetatif akhir memiliki jumlah data paling sedikit. Suatu data dikatakan seimbang jika memiliki nilai IR (*Imbalance Rate)* mendekati 1 (Khan dkk, 2021). Nilai IR dihitung dengan jumlah kelas mayoritas dibagi dengan setiap kelas. Label vegetatif akhir memiliki nilai IR sebesar 2.75, label generatif sebesar 1.4, dan label panen sebesar 1.3. Sehingga dapat dikatakan terdapat ketidakseimbangan *multiclass.*

Penelitian yang dilakukan oleh Kurniawati (2023), Marsuhandi dkk. (2019), dan Triscowati (2019) memanfaatkan amatan citra satelit LANDSAT 8 untuk mengamati kelas atau memodelkan kelas yang ada pada KSA. Indeks citra satelit LANDSAT 8 yang akan digunakan adalah *Enhanced Vegetation Index* (EVI). EVI merupakan Indeks vegetasi pada data citra satelit yang sering digunakan untuk mendeteksi fase pertumbuhan padi karena mampu meningkatkan sinyal vegetasi dengan sensitivitas di daerah biomassa tinggi dan meningkatkan pemantauan vegetasi dengan melakukan pengurangan pengaruh atmosfer (Domiri, 2011). Setiap kelas pada keempat fase pertumbuhan padi pada label KSA memiliki nilai EVI yang berbeda. Fase vegetatif awal memiliki nilai EVI pada rentang 0.2-0.3, fase vegetatif akhir pada rentang 0.3-0.4, fase generatif pada rentang 0.4-0.5, dan fase panen pada rentang 0.5-0.6.

Metode yang dapat digunakan untuk menangani ketidakseimbangan *multiclass* adalah dengan menggunakan metode *SMOTE and Cluster-based Undersampling Technique* (SCUT). Metode SCUT pertama kali diajukan oleh (Agrawal dkk, 2015). SCUT mengombinasikan SMOTE dan *Cluster-based Undersampling Technique* (CUT) untuk mengatasi ketidakseimbangan data pada klasifikasi *multiclass*. Ide pokok dari CUT yaitu masing-masing kelas dalam kumpulan data dikelompokkan secara individual. Pada SMOTE akan dibangkitkan data sintetik atau data buatan untuk kelas yang memiliki jumlah data kurang dari rata-rata kelas. Agrawal dkk. (2015) membandingkan SCUT dengan SMOTE, *random undersampling*, dan CUT. Hasil penelitian tersebut menunjukkan SCUT dan SMOTE secara konsisten menghasilkan kinerja terbaik. Penelitian lainnya juga dilakukan oleh Kurniawati (2023) menyatakan bahwa SCUT mampu mengatasi masalah ketidakseimbangan kelas yang ekstrim pada data *multiclass*.

Setelah melakukan penanganan ketidakseimbangan, selanjutnya dilakukan pengujian terhadap data yang sudah dibangkitkan. Pengujian dilakukan untuk membandingkan karakteristik dari data yang dibangkitkan terhadap data asli. Perbandingan ini penting dilakukan karena setiap nilai memiliki kategori yang berbeda pada label EVI.

II. METODE PENELITIAN

A. Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder. Data diperoleh dari penelitian Kurniawati (2023) yang menggunakan dataset survei Kerangka Sampel Area dan Citra Satelit LANDSAT 8. Label KSA yang akan digunakan adalah fase pertumbuhan padi yakni vegetatif awal, vegetatif akhir, generatif, dan panen. Sedangkan indeks citra satelit yang digunakan adalah EVI pada periode April 2018.

B. Teknik Analisis Data

Adapun langkah-langkah analisis yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Melakukan eksplorasi data dengan menggunakan *barchart* untuk melihat distribusi kelas.

Eksplorasi dilakukan untuk melihat karakteristik gugus data, khususnya kriteria ketidakseimbangan data dan kondisi tumpang tindih (*overlapping*) pada kelas setiap gugus data. Ketidakseimbangan data ditunjukkan oleh proporsi setiap kelas.

2. Penanganan data tidak seimbang menggunakan SCUT.

Algoritma pada SCUT mengombinasikan teknik *undersampling* berupa *Cluster-based Undersampling Technique* dan teknik *oversampling* berupa SMOTE. Langkah-langkah melakukan SCUT adalah sebagai berikut.

a. Membagi gugus data menjadi n bagian yaitu D_1, D_2, D_3, $\cdots$, D_n, di mana n adalah banyaknya kelas dan D_i merupakan kelas tunggal.
b. Hitung rata-rata anggota kelas (m)

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/163

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

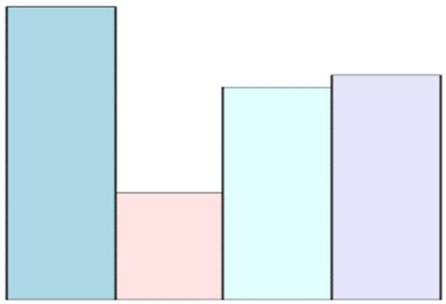
c. Untuk setiap kelas D_i, i = 1,2, ⋯ , n, jika jumlah anggota kelasnya lebih besar dari m, lakukan *cluster-based undersampling* dengan cara berikut. Jika gerombolkan D_i (*cluster within each class*) menggunakan EM *algorithm*. Sedangkan untuk setiap kluster C_j, j = 1,2, ⋯ , k, Memilih amatan secara acak dari kluster C_j sehingga total amatan dari seluruh kluster C_j sama dengan m
d. Untuk setiap D_i, i = 1,2, ⋯ , n, jika jumlah anggota kelasnya kurang dari m, lakukan *oversampling* dengan menerapkan SMOTE pada Di untuk mendapatkan data D_i
e. Untuk setiap D_i, i = 1,2, ⋯ , n, jika jumlah anggota kelasnya sama dengan m, maka data $D_i' = D_i$
f. Gabungkan setiap gugus data baru yang diperoleh pada tahap c-e.

3. Uji Normalitas
Uji Normalitas adalah sebuah uji yang dilakukan dengan tujuan untuk menilai sebaran data pada sebuah kelompok data atau variabel. Uji normalitas yang umum digunakan jika jumlah sampel lebih dari 50 adalah Uji Kolmogorov Smirnov. Konsep dasar dari uji normalitas Kolmogorov Smirnov adalah dengan membandingkan distribusi data dengan distribusi normal baku.
4. Lakukan pengujian dua sampel untuk melihat perbedaan rataan antara dataset asli dengan dataset yang telah ditangani menggunakan metode *SMOTE and Cluster-based Undersampling Technique* sebagai berikut.
   a. Uji Wilcoxon
      Uji Wilcoxon adalah sebuah uji nonparametrik yang digunakan untuk mengukur signifikansi perbedaan antara dua kelompok data berpasangan. Uji ini digunakan untuk menganalisis hasil pengamatan yang berpasangan dari dua data apakah terdapat perbedaan atau tidak.
   b. Uji Scheffe
      Uji Scheffe adalah uji parametrik yang digunakan untuk menganalisis hasil pengamatan yang berpasangan dari dua data apakah terdapat perbedaan atau tidak.
5. Interpretasi hasil.

III. HASIL DAN PEMBAHASAN

A. Eksplorasi Data

Eksplorasi data dilakukan untuk melihat ketidakseimbangan pada pada masing-masing kelas label KSA. Gambaran ketidakseimbangan tersebut dapat dilihat pada Gambar 2.



Gambar 2. *Barchart* Ketidakseimbangan pada Label KSA

Pada Gambar 2 dapat dilihat bahwa Label 1 atau vegetatif awal memiliki jumlah data yang lebih banyak dibandingkan kelas lainnya sehingga disebut sebagai kelas mayor. Sedangkan label 2 atau vegetatif akhir memiliki jumlah kelas yang jauh lebih sedikit dibandingkan kelas lainnya sehingga disebut kelas minor. Sehingga dapat dikatakan bahwa data tersebut mengalami ketidakseimbangan *multiclass.* Selanjutnya akan dilakukan penanganan menggunakan metode *SMOTE and Cluster-based Undersampling Technique*.

B. *SMOTE and Cluster-based Undersampling Technique* (SCUT)

SMOTE and Cluster-based Undersampling Technique merupakan gabungan dari *oversampling* menggunakan SMOTE dan *undersampling* menggunakan CUT. SMOTE akan membangkitkan data sintetik pada kelas yang memiliki jumlah data kurang dari rata-rata kelas. Hasil penanganan ketidakseimbangan pada dataset KSA menggunakan metode SCUT dapat dilihat pada Tabel 1.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/163

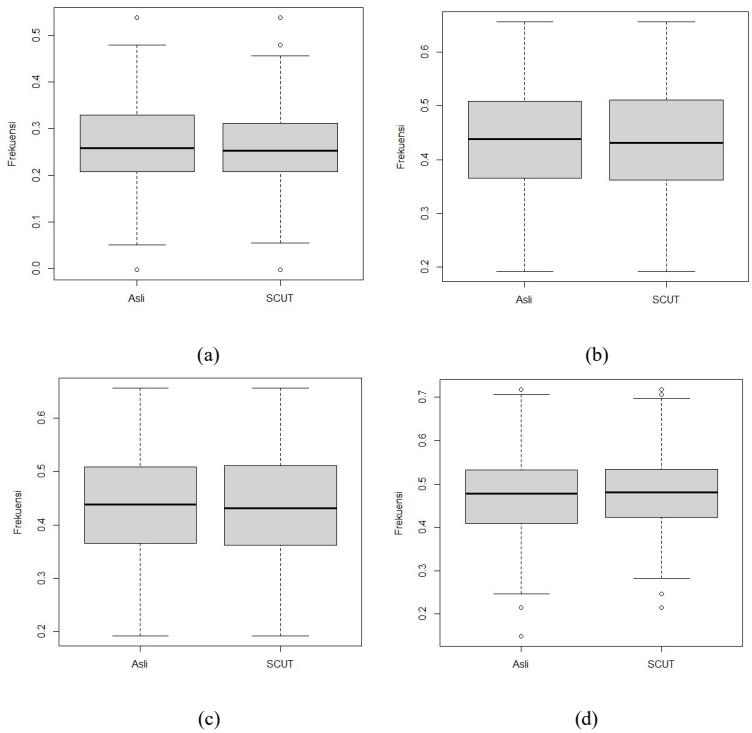
ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Tabel 1. Dataset KSA Sebelum dan Sesudah Dilakukan Penanganan SCUT

Label KSA	Dataset Asli	SCUT
Vegetatif Awal	154	110
Vegetatif Akhir	56	110
Generatif	112	110
Panen	118	110

Berdasarkan Tabel 1 dapat dilihat bahwa setelah dilakukan penanganan menggunakan SCUT, jumlah data setiap kelas pada Label KSA menggunakan SCUT masing-masing menjadi 110. Fase vegetatif awal, vegetatif akhir, dan panen dilakukan *Cluster-based Undersampling Technique* karena jumlah data dalam setiap kelas lebih dari rata-rata anggota kelas. Jumlah rata-rata kelas yakni jumlah seluruh data dibagi jumlah kelas dan diperoleh bahwa rata-rata kelas adalah 110. Sehingga dilakukan CUT pada ketiga label KSA tersebut. Sedangkan pada fase generatif (kelas minor) dilakukan *Synthetic Minority Oversampling Technique* (SMOTE) karena jumlah data dalam setiap kelas kurang dari rata-rata anggota kelas. Dengan menggunakan metode SCUT tersebut data KSA sudah seimbang.

Selanjutnya dilakukan perbandingan distribusi antara dataset asli dan dataset setelah dilakukan penanganan menggunakan SCUT menggunakan *boxplot*. *Boxplot* bisa digunakan untuk melihat perbandingan nilai *maximum*, minimum, kuartil pertama, kuartil kedua (*mean*), kuartil ketiga, dan *outlier* dari kedua dataset. Ilustrasi *boxplot* dapat dilihat pada Gambar 3.



Gambar 3. (a) *Boxplot* Perbandingan Label 1 Data Asli dan SCUT, (b) *Boxplot* Perbandingan Label 2 Data Asli dan SCUT, (c) *Boxplot* Perbandingan Label 3 Data Asli dan SCUT, dan (d) *Boxplot* Perbandingan Label 2 Data Asli dan SCUT,

Berdasarkan Gambar 3 dapat dilihat bahwa distribusi antara dataset asli dan SCUT pada Label 1 (vegetatif awal), label 2 (vegetatif akhir), dan label 3 (generatif) relatif sama, hal ini dapat dilihat dari kedua dataset yang memiliki garis sejajar atau sama antara dataset sebelum dilakukan penanganan dan setelah dilakukan penanganan. Namun pada label 4 (panen) terlihat sedikit perbedaan distribusi data pada kedua boxplot. Hal tersebut perlu dibuktikan dengan

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/163

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

pengujian signifikansi dua sampel untuk melihat perbedaan dari dataset dengan penanganan SCUT terhadap dataset asli.

C. Uji Normalitas

Uji normalitas digunakan untuk melihat apakah data berdistribusi normal atau tidak. Pengujian dilakukan untuk setiap label KSA yang sudah dilakukan penanganan menggunakan metode SCUT. Pengujian dilakukan dengan uji Kolmogorov Smirnov. Jika signifikansi di atas 0.05 maka berarti tidak terdapat perbedaan yang signifikan antara data yang akan diuji dengan data normal baku. Hasil uji normalitas Kolmogorov *Simirnov* dapat dilihat pada Tabel 2.

Tabel 2. Hasil Uji Normalitas pada SCUT

Label KSA	*P-Value*
Vegetatif Awal	0.1187
Vegetatif Akhir	0.0743
Generatif	0.0172
Panen	0.1187

Berdasarkan Tabel 2 fase vegetatif awal, vegetatif akhir, dan panen memiliki data berdistribusi normal karena *p-value* > 0.05. Sedangkan fase generatif memiliki data tidak berdistribusi normal karena *p-value* ≤ 0.05. Selanjutnya dilakukan uji untuk melihat perbedaan rataan dari kedua data. Data yang berdistribusi normal akan dilakukan uji Scheffe dan data yang tidak berdistribusi normal akan dilakukan uji Wilcoxon.

D. Uji Scheffe dan Uji Wilcoxon

Uji yang digunakan untuk melihat perbandingan rataan pada artikel ini adalah Uji Scheffe dan Uji Wilcoxon. Uji Scheffe digunakan untuk melihat perbedaan rataan dari dua sampel yang berbeda untuk data berdistribusi normal. Jika *p-value* > 0.05 maka dapat dikatakan tidak terdapat perbedaan rataan yang signifikan antara kedua sampel. Uji Scheffe dilakukan pada fase vegetatif awal, vegetatif akhir, dan panen. Serta juga dilakukan Uji Wilcoxon untuk fase generatif. Uji wilcoxon berguna untuk melihat perbedaan rataan antara dua sampel non parametrik. Hasil Uji Scheffe dan Uji Wilcoxon yang diperoleh dapat dilihat pada Tabel 3.

Tabel 3. Hasil Scheffe dan Wilcoxon Dataset Asli dan SCUT

Uji	Label KSA	*P-Value*
Uji Scheffe	Vegetatif Awal	0.7653
	Vegetatif Akhir	0.9965
	Panen	0.1331
Uji Wilcoxon	Generatif	0.0014

Berdasarkan Tabel 3 uji scheffe dapat dilihat bahwa fase vegetatif awal, vegetatif akhir, dan panen memiliki *p-value* > 0.05. Sehingga dapat dikatakan bahwa rataan antara dataset asli dengan dataset yang sudah dilakukan penanganan menggunakan SCUT tidak memiliki perbedaan rataan yang signifikan. Pada uji Wilcoxon diperoleh hasil *p-value* pada uji wilcoxon untuk fase generatif yakni sebesar 0.0014< 0.05. Sehingga dapat dikatakan terdapat perbedaan rataan yang signifikan antara dataset asli dan dataset setelah penanganan SCUT.

IV. KESIMPULAN

Berdasarkan hasil penelitian ini dengan melakukan penanganan ketidakseimbangan dataset menggunakan SCUT diperoleh bahwa jumlah data pada setiap kelas menjadi seimbang. Selanjutnya dilakukan visualisasi menggunakan *boxplot* untuk melihat distribusi data sintetik yang diperoleh menggunakan SCUT terhadap dataset asli. Diperoleh hasil bahwa distribusi data antara dataset asli dan SCUT relatif sama. Kemudian dilakukan pengujian rataan dua sampel untuk melihat perbedaan rataan antara dataset asli dan dataset setelah penanganan. Hasil yang diperoleh adalah pada fase vegetatif awal, vegetatif akhir, dan panen memiliki rataan yang signifikan sama antara dataset asli dan dataset setelah dilakukan penanganan menggunakan SCUT, namun fase generatif memiliki rataan yang tidak sama antara dataset asli dan setelah dilakukan penanganan menggunakan SCUT. Oleh karena itu, metode SCUT mampu menciptakan data sintetik yang memiliki karakteristik yang sama dengan dataset asli pada label vegetatif awal, vegetatif akhir, dan panen.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/163

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

Sedangkan pada label generatif ditemukan perbedaan rataan antara dataset asli dan SCUT yang berarti terdapat perbedaan karakteristik dari data sintetik yang dihasilkan menggunakan metode SCUT.

DAFTAR PUSTAKA

Agrawal, A., Viktor, H. L., & Paquet, E. (2015), "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling", *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Manag*, hal. 226–234.

Alqaida, R. A., Ngurah, G., Wibawa, A., Yahya, I., Abapihi, B., Laome, L., & Oleo, U. H. (2022), "Combine Undersampling untuk Menangani Data Tidak Seimbang", *Prosiding Seminar Nasional Sains dan Terapan*, hal. 71–78.

BPS. (2018), "Luas Panen dan Produksi Beras di Indonesia 2018", *Badan Pusat Statistik,* hal. 26-29.

Dharmawan, H. (2023), "Perbandingan Ukuran Kepentingan Peubah dari Berbagai Algoritme Pembelajaran Mesin untuk Kondisi Data Tidak Seimbang dengan Berbagai Jenis Perlakuan", *Institut Pertanian Bogor*, hal. 35-38.

Domiri, D. D. (2011), "Aplikasi Simulasi Model Dinamis Pertumbuhan Tanaman Untuk Menduga Produksi Tanaman Padi", *Jurnal Penginderaan Jauh*, hal. 35–49.

Kurniawati, Y. (2023), "Penduga Area Kecil Berhirarki untuk Luas Panen Padi Berbasis Survei KSA-BPS dengan Memanfaatkan Citra Satelit LANDSAT 8", *Institut Pertanian Bogor,* hal. 50-52

Marsuhandi, A. H., Soleh, A. M., Wijayanto, H., & Domiri, D. D. (2019), "Pemanfaatan Ensemble Learning dan Penginderaan Jauh untuk Pengklasifikasian Jenis Lahan Padi", *Institut Pertanian Bogor,* hal. 20-22.

Triscowati, D. W. (2019), "Klasifikasi Fase Pertumbuhan Padi Menggunakan Random Forest Berdasarkan Data Multitemporal Landsat-8", *Institute Pertanian Bogor*, hal. 38-40.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/163