Universitas Ahmad Dahlan

Jurnal Ilmiah Matematika
Vol. 10, No. 1, Mei 2023, pp. 34-42
ISSN 2774-3241
http://journal.uad.ac.id/index.php/Konvergensi

Optimasi Metode CART Menggunakan Metode Bagging Pada Studi Kasus Data Imbalance Berbasis Metode Adasyn

Khana Pujiyanti [a,1,*]

[*a] Universitas Ahmad Dahlan, Yogyakarta, Indonesia;
[1] khana2000015035@webmail.uad.ac.id
*Correspondent Author

Received: *Revised:* *Accepted:*

KATAKUNCI

Data *Imbalance*
CART
ADASYN
BAGGING

ABSTRAK

Penelitian ini membahas mengenai permasalahan data *imbalance* yang menyebabkan kinerja dari model klasifikasi menjadi tidak optimal. Dalam penelitian ini menerapkan metode *Adaptive Synthetic* Sampling (ADASYN) untuk menangani permasalahan data *imbalance*, metode Classification and Regression Tree (CART) diterapkan sebagai metode klasifikasi pada dataset penyakit stroke, dan metode Boostrap Agregating (*Bagging*) untuk mengoptimalkan metode Cart. Penelitian ini bertujuan untuk mengetahui cara kerja dan performa dari penerapan metode Adasyn, Cart, dan Bagging dengan membangun tiga model klasifikasi yaitu model Cart, model Cart Adasyn, dan model Cart Adasyn Bagging. Hasil penelitian menunjukan model Cart menghasilkan nilai akurasi sebesar 94%, G-mean sebesar 0%, dan AUC sebesar 50%. Model Cart Adasyn menghasilkan nilai akurasi 78%, G-Mean 74% dan AUC 74%. Model Cart Adasyn Bagging menghasilkan nilai akurasi 78%, G-mean 76%, dan AUC 76%. Oleh karena itu, dapat disimpulkan bahwa kombinasi metode Cart, Adasyn, dan Bagging memberikan performa terbaik dalam mengatasi data tidak seimbang untuk klasifikasi penyakit stroke. Model Cart Adasyn Bagging terbukti lebih baik dalam memprediksi kedua kelas mayoritas dan kelas minoritas.

KEYWORDS

Imbalance dataset
CART
ADASYN
BAGGING

ABSTRACT

This research discusses the problem of imbalance dataset which causes the performance of the classification model to be not optimal. In this study, the Adaptive Synthetic Sampling (ADASYN) method is applied to handle imbalance dataset problems, the Classification and Regression Tree (CART) method is applied as a classification method on stroke disease datasets, and the Boostrap Aggregating (Bagging) method to optimize the Cart method. This study aims to determine the workings and performance of the application of the Adasyn, Cart, and Bagging methods by building three classification models, namely the Cart model, the Cart Adasyn model, and the Cart Adasyn Bagging model. The results showed that the Cart model produced an accuracy value of 94%, G-mean of 0%, and AUC of 50%. The Cart Adasyn model produces an accuracy value of 78%, G-Mean 74% and AUC 74%. The Cart Adasyn Bagging model produces an accuracy value of 78%, G-mean 76%, and AUC 76%. Therefore, it can be concluded that the combination of Cart, Adasyn, and Bagging methods provides the best performance in overcoming unbalanced data for stroke disease classification. The Cart

 ISSN 2774-3241

Adasyn Bagging model proved to be better at predicting both the majority class and the minority class.

Pendahuluan

Klasifikasi merupakan salah satu metode pembelajaran mesin yang digunakan untuk memprediksi label atau kategori suatu data, dan banyak diaplikasikan dalam berbagai bidang, seperti perbankan dan medis. Tantangan umum dalam klasifikasi adalah ketidakseimbangan data (*imbalanced* dataset), yaitu kondisi di mana jumlah data dalam satu kelas jauh lebih banyak dibandingkan kelas lainnya. Ketidakseimbangan ini menyebabkan model cenderung bias terhadap kelas mayoritas sehingga menurunkan performa klasifikasi [1]. Untuk mengatasi hal tersebut, [2] mengembangkan metode *Adaptive Synthetic Sampling* (ADASYN), yaitu teknik oversampling yang menghasilkan data sintetis berdasarkan tingkat kesulitan dalam mempelajari kelas minoritas.

Beberapa penelitian telah mengkaji efektivitas ADASYN dalam menangani data tidak seimbang. [3] membandingkan metode ADASYN-SVM dan SMOTE-SVM dalam deteksi Diabetes Mellitus Tipe 2, dan menemukan bahwa ADASYN-SVM memiliki akurasi lebih tinggi (87,3%) dibanding SMOTE-SVM (85,4%). [4] menggabungkan ADASYN dengan CART untuk prediksi curah hujan di Kabupaten Bandung, dan hasilnya menunjukkan performa lebih baik dibanding CART tanpa ADASYN.

Metode *Classification and Regression Tree* (CART) merupakan metode statistik nonparametrik berbasis pohon keputusan yang digunakan untuk klasifikasi dan regresi. Namun, model CART sering kali tidak stabil sehingga akurasinya mudah terpengaruh oleh perubahan kecil pada data latih. Untuk mengatasi hal ini, metode *Bagging* (*Bootstrap Aggregating*) digunakan untuk meningkatkan stabilitas dan akurasi model.

[5] menerapkan ensemble CART dengan *bagging* untuk menilai pencapaian siswa dalam matematika, dan memperoleh kesesuaian model hingga 96% terhadap data aktual. [6] menggunakan CART dan bagging untuk klasifikasi multi-label topik hadits, dan menunjukkan bahwa *bagging* dapat menurunkan nilai *hamming loss* menjadi 0,1914 atau meningkatkan akurasi sebesar 5%. Sementara itu, [7] menggunakan CART ensemble dan bagging untuk memodelkan pengaruh strategi pemasaran terhadap penjualan furnitur, menghasilkan koefisien determinasi sebesar 96% dan MAPE di bawah 5%. [8] menerapkan CART dan CART-Bagging untuk klasifikasi kelulusan mahasiswa, dan membuktikan bahwa CART-Bagging dapat meningkatkan akurasi dari 70% menjadi 85,71%.

 ISSN 2774-3241

Selain itu, stroke merupakan salah satu penyakit kronis utama yang memiliki tingkat kematian tinggi di Indonesia, dengan prevalensi meningkat dari 7 per 1000 penduduk (2013) menjadi 10,9 per 1000 penduduk (2018) menurut Riskesdas. Oleh karena itu, klasifikasi yang akurat terhadap risiko stroke sangat penting.

Berdasarkan penelitian terdahulu, studi ini akan mencoba menerapkan metode Adasyn untuk mengatasi data tidak seimbang kemudian diuji menggunakan metode Decision Tree khususnya Cart untuk klasifikasi penyakit Stoke. Namun, Decision Tree masih terbatas untuk data yang 6 terlalu kompleks, sehingga diperlukan metode ensembel untuk meningkatkan akurasi Decision Tree. Dengan tujuan dapat mengetahui tingkat akurasi algoritma Cart Adasyn Bagging pada klasifikasi penyakit stroke.

Metode

Penelitian ini dimulai dengan melakukan tahap pengumpulan data yang akan digunakan sebagai objek kajian. Selanjutnya, dilakukan analisis deskriptif untuk memahami struktur dan karakteristik data secara menyeluruh. Data kemudian diproses melalui tahap *preprocessing* yang berguna untuk memastikan kualitas dan kesiapan data sebelum dimodelkan. Untuk mengatasi permasalahan ketidakseimbangan kelas pada data latih, digunakan pendekatan ADASYN (Adaptive Synthetic Sampling). Setelah proses penyeimbangan data, dilakukan pelatihan dan pengujian terhadap beberapa model klasifikasi, yaitu CART, CART dengan ADASYN, serta CART ADASYN yang dikombinasikan dengan metode *bagging*. Evaluasi kinerja dari masing-masing model dilakukan menggunakan metrik *confusion matrix*, dan hasil evaluasi tersebut menjadi dasar dalam penarikan simpulan akhir dari penelitian ini.

Hasil dan Pembahasan

1. Pengolahan Data

Penelitian ini menggunakan kumpulan data pasien penyakit stroke yang terdiri dari 4981 data dengan 11 variabel yang terdiri dari gander, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status, stroke. Selanjutnya proses pengolahan data meliputi proses prepropcessing untuk membersihakan, mengubah serta menyiapkan data sehingga proses klasifikasi dapat berjalan lebih mudah. Adapun tahapan preprocesing yang dilakukan yaitu data cleaning, transformasi data, pemilihan fitur, split dataset.

a. Data cleaning

Setelah dilakukan pemerikasaan pada dataset diketahui bahwa dataset yang digunakan tidak memiliki missing value dan tidak ada data yang terduplikasi sehingga dataset yang

 ISSN 2774-3241

digunakan siap untuk diolah.

b. Transformasi data

Transformasi data merupakan proses untuk mengganti format type data dengan menyesuaikan pada type data yang lainnya. Type data pada setiap variabel terbagi menjadi dua yaitu type data numerik dan type data kategorik. Variabel dengan type data kategorikal dikonversi menjadi 0,1, dts. Berikut tabel yang menampilkan dataset dari variabel yang telah melalui proses transformasi data.

Tabel 1. Data dengan Variabel yang Melalui Proses Transformasi Data

	Gender	***Age***	***...***	***Bmi***	***Smoking_status***
0	1	2	...	1	1
1	1	2	...	1	2
2	0	1	...	1	3
3	0	2	...	2	2
4	1	2	...	1	1
:	:	:	:	:	:
4976	1	1	...	1	1
4977	1	1	...	1	3
4978	0	1	...	1	3
4979	1	1	...	1	3
4980	0	2	...	1	2

c. Pemilihan fitur

Penelitian ini memiliki 10 variabel kemudian membuang fitur ever_married, work_type, dan residence_type karena memberikan pengaruh yang lebih sedikit dibandingkan faktor yang lainnya. Sehingga penelitian ini menggunakan 7 fitur sebagai variabel independen.

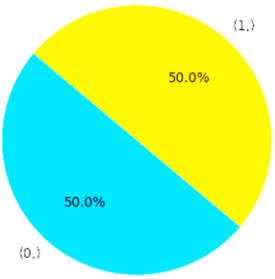
d. Split dataset

Pada penelitian ini terdapat 7 variabel independen dan 1 variabel dependen. Kedua variabel tersebut masing-masing dibagi menjadi data training dan data testing. Pembagian ini dilakukan dengan membagi data training sebanyak 90% dan data testing sebanyak 10%, sehingga diperoleh data training sebanyak 4482 dataset dan data testing sebanyak 499 dataset.

2. Penyeimbangan Data

Proses penyeimbangan data dilakukan dengan menerapkan metode ADASYN pada data training. Jumlah dataset pada data training sebanyak 4482 dataset dengan 4265 data pasien tidak menderitaa stroke, dan 217 data pasien penderita stroke. Jadi dapat diketahui bahwa jumlah kelas minoritas sebanyak 217 data pasien penderita stroke, dan jumlah kelas mayoritas sebanyak 4265 data pasien tidak menderita stroke. Berikut adalah hasil proses penyeimbangan data :

 ISSN 2774-3241



Gambar 1. Presentase Kelas Setelah Menerapkan Metode Adasyn

Berdasarkan Gambar 1 menunjukkan distribusi kelas setelah dilakukan resampling menggunakan metode ADASYN. Terlihat bahwa kedua kelas, yaitu kelas (0,) dan kelas (1,), memiliki proporsi yang seimbang masing-masing sebesar 50%. Hal ini menunjukkan bahwa ADASYN berhasil mengatasi permasalahan ketidakseimbangan kelas pada dataset.

3. Penerapan Model

Pada tahapan penerapan model ini dataset akan dibangun dengan beberapa model klasifikasi. Pada penelitian ini akan membangun tiga model klasifikasi yaitu, model CART, model CART ADASYN, dan model CART ADASYN BAGGING.

a. CART

CART merupakan model klasifikasi yang pertama dibangun pada penelitian ini. Berikut merupakan perolehan nilai index gini dan gini gain untuk setiap variabel yang digunakan :

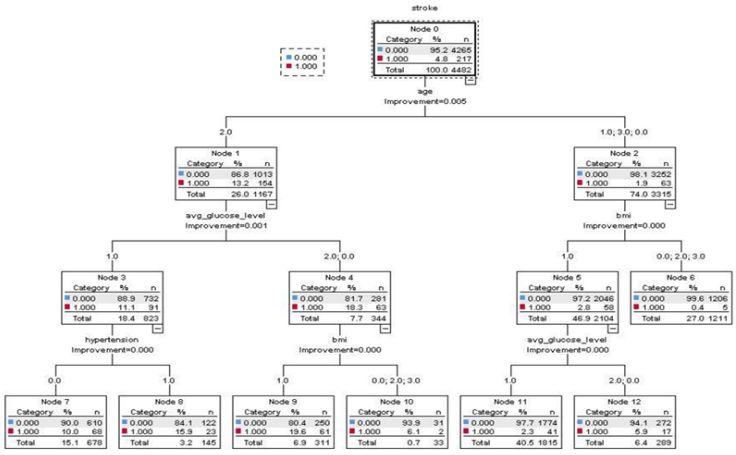
Tabel 2. Nilai Index Gini Dan Gini Gain Pada Setiap Variabel

No	*Variabel*	*Kategori*	*Jumlah*	*Index Gini*	*Gini Gain*
1	Stroke		4482	0.0921	
2	Gender	0	2606	0.0889	0.0015
		1	1876	0.0935	
3	Age	0	442	0.0022	0.0049
		1	2487	0.0479	
		2	1167	0.2309	
		3	386	0.0052	
4	Hypertension	0	4061	0,0781	0,0004
		1	421	0,2224	
5	Heart_disease	0	4237	0,0802	0,0027
		1	245	0,2512	
6	Avg_glucose_level	0	372	0,2375	0,0004
		1	3776	0,0711	
		2	334	0,1635	
7	bmi	0	682	0,0490	0,0007
		1	3064	0,1122	
		2	443	0,0667	
		3	293	0,0068	
8	Smoking_status	0	1339	0,0636	0,0002

 ISSN 2774-3241

	1	794	0,1395
	2	1652	0,0909
	3	697	0,0928

Berdasarkan tabel tersebut dapat diketahui bahwa variabel dengan nilai gini gain terbesar yaitu variabel age sebesar 0,0049. Sehingga dapat dipastikan bahwa variabel age dipilih sebagai node akar atau node 1. Kemudian melakukan percabangan kanan dan kiri untuk dilakukan proses klasifikasi dengan pohon klasifikasi seperti pada Gambar 2.



Gambar 2. Pohon Klasifikasi Pada Model CART

b. CART ADASYN

CART ADASYN merupakan model klasifikasi yang kedua dibangun pada penelitian ini. Berikut merupakan perolehan nilai index gini dan gini gain untuk setiap variabel yang digunakan

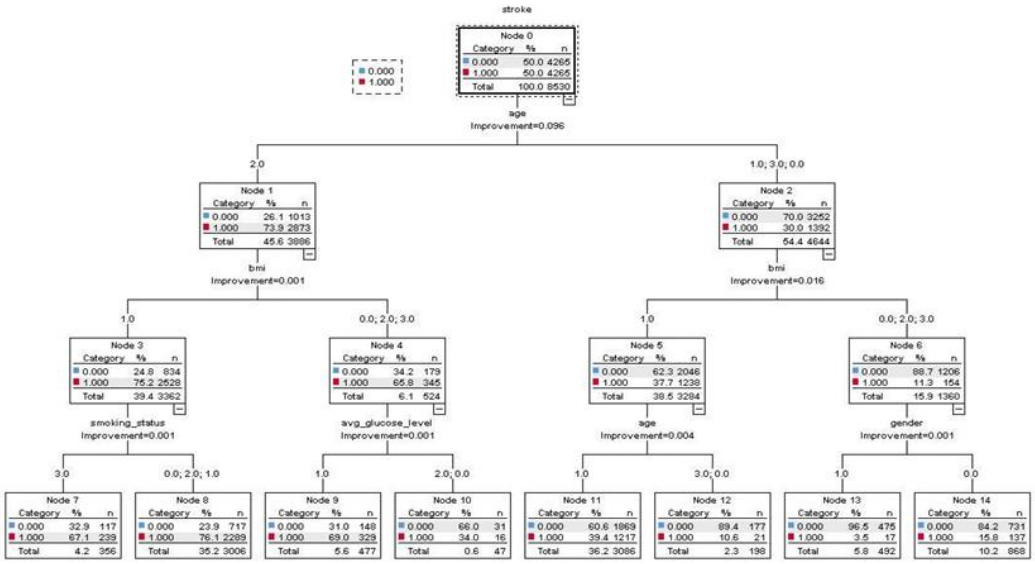
Tabel 3. Nilai Index Gini Dan Gini Gain Pada Setiap Variabel

No	*Variabel*	*Kategori*	*Jumlah*	*Index Gini*	*Gini Gain*
1	Stroke		8530	0,5	
2	Gender	0	4909	0,4999	0,0001
		1	3621	0,4998	
3	Age	0	462	0,0867	0,0962
		1	3754	0,3972	
		2	3886	0,4724	
		3	428	0,1807	
4	Hypertension	0	7471	0,4990	0,0067
		1	1059	0,4529	
5	Heart_disease	0	7941	0,4998	0,0031
		1	589	0,4578	
6	Avg_glucose_level	0	994	0,4373	0,0083
		1	6894	0,4984	
		2	642	0,4986	
7	bmi	0	892	0,3736	0,0373
		1	6646	0,4911	

 ISSN 2774-3241

		2	685	0,4688	
		3	307	0,0929	
8	Smoking_status	0	2172	0,4815	0,0109
		1	1916	0,4727	
		2	3166	0,4999	
		3	1276	0,4992	

Berdasarkan tabel tersebut dapat diketahui bahwa variabel dengan nilai gini gain terbesar yaitu variabel age sebesar 0,0962. Sehingga dapat dipastikan bahwa variabel age dipilih sebagai node akar atau node 1. Kemudian melakukan percabangan kanan dan kiri untuk dilakukan proses klasifikasi dengan pohon klasifikasi seperti pada Gambar 3.



Gambar 3. Pohon klasifikasi pada model CART ADASYN

c. CART ADASYN BAGGING

Model yang terakhir dibangun yaitu model CART ADASYN BAGGING. Pada model ini, data hasil metode ADASYN akan di latih pada metode BAGGING dengan metode klasifikasi dasar yang digunakan yaitu metode CART. Tahapan yang dilalui yaitu tahap bootstrap dan tahap aggregating. Berikut contoh penerapan tahap aggregating dengan aturan suara terbanyak pada 3 dataset:

Tabel 4. Hasil Klasifikasi Pada Pohon Boostrap

Pohon	***Data***		
Bootstrap	***1***	***2***	***3***
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0
5	1	0	0
6	1	0	0
7	1	0	0

 ISSN 2774-3241

8	1	0	0
9	1	0	0
10	1	0	0
11	1	0	0
12	1	0	0

Berdasarkan tabel 4 dapat diketahui bahwa data 1 diklasifikasikan pada kelas 1 karena pada ke 12 pohon tersebut mengklasifikasikan pada kelas 1. Data 2 dan 3 akan diklasifikasikan pada kelas 0 karena pada ke 12 pohon tersebut mengklasifikasikan pada kelas 0. Dengan cara yang sama terapkan pada data lainnya sehingga diperoleh data prediksi pada tahap aggregating.

4. Perbandingan Performa Model

Setelah dilakukan perhitungan ketepatan klasifikasi pada model Cart, Cart, dan Cart Adasyn Bagging diatas, berikut perolehan nilai ketepatan klasifikasi pada data imbalance:

Tabel 5. Hasil Perbandingan Performa Model

	Model		
	CART	***CART ADASYN***	***CART ADASYN BAGGING***
Akurasi	94%	78%	78%
G-Mean	0%	74%	76%
AUC	50%	74%	76%

Berdasarkan tabel 5 dapat diketahui bahwa nilai akurasi pada model Cart 94%, Cart Adasyn 78%, dan Cart Adasyn Bagging 78%. Terjadi penurunan nilai akurasi pada model yang melalui teknik oversampling. Nilai G-mean pada model CART sebesar 0%, model Cart Adasyn sebesaar 74% dan Cart Adasyn Bagging sebesar 76%. Terjadi peningkatan nilai G mean pada model yang menerapkan teknik oversampling. Perolehan nilai AUC pada model Cart sebesar 50%, model Cart Adasyn 74% dan Cart Adasyn Bagging sebesar 76%. Terjadi peningkatan pada model yang menerapkan teknik oversampling.

Simpulan

Berdasarkan penelitian yang telah dilakukan, metode Adaptive Synthetic Sampling (ADASYN) diterapkan untuk menangani ketidakseimbangan kelas pada data pelatihan dengan menghasilkan 4048 sampel sintetik melalui proses perhitungan rasio ketidakseimbangan kelas, penentuan jumlah sampel sintetis, dan interpolasi antara sampel minoritas dengan tetangga terdekatnya menggunakan faktor acak. Metode Classification and Regression Tree (CART) digunakan sebagai model klasifikasi dengan membangun pohon keputusan berdasarkan indeks gini dan gini gain, sedangkan metode Bootstrap Aggregating (Bagging) diterapkan untuk mengoptimalkan performa model CART Adasyn melalui pembentukan 12 sampel bootstrap dan penggabungan hasil prediksi berdasarkan mayoritas suara. Hasil pengujian menunjukkan bahwa model CART tanpa oversampling menghasilkan akurasi tinggi

 ISSN 2774-3241

sebesar 94% tetapi memiliki G-Mean 0% dan AUC 50%, yang menunjukkan kegagalan dalam memprediksi kelas minoritas. Setelah diterapkan ADASYN, akurasi menurun menjadi 78%, namun G-Mean dan AUC meningkat signifikan menjadi 74%, yang menunjukkan model lebih seimbang dalam memprediksi kedua kelas. Penambahan Bagging pada model CART Adasyn meningkatkan G-Mean menjadi 76% dan AUC menjadi 76%, meskipun akurasi tetap 78%. Dengan demikian, kombinasi metode CART, ADASYN, dan Bagging terbukti memberikan performa terbaik dalam klasifikasi data tidak seimbang penyakit stroke dibandingkan model CART maupun CART ADASYN secara terpisah.

Daftar Pustaka

[1] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016, doi: 10.1007/s13748-016-0094-0.

[2] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," Proceedings of the International Joint Conference on Neural Networks, 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.

[3] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Sci. J. Informatics*, vol. 8, no. 2, pp. 276–282, 2021, doi: 10.15294/sji.v8i2.32484.

[4] S. N. Lathifah, F. Nhita, A. Aditsania, and D. Saepudin, "Rainfall forecasting using the classification and regression tree (CART) algorithm and adaptive synthetic sampling (study case: Bandung regency).," in *2019 7th International Conference on Information and Communication Technology (ICoICT)*, 2019, pp. 1–5. doi: 10.1109/ICoICT.2019.8835308.

[5] S. Gocheva-Ilieva, H. Kulina, and A. Ivanov, "Assessment of students' achievements and competencies in mathematics using cart and cart ensembles and bagging with combined model improvement by mars," *Mathematics*, vol. 9, no. 1, pp. 1–17, 2021, doi: 10.3390/math9010062.

[6] R. Kustiawan, A. Adiwijaya, and M. D. Purbolaksono, "A Multi-label Classification on Topic of Hadith Verses in Indonesian Translation using CART and Bagging," *J. Media Inform. Budidarma*, vol. 6, no. 2, p. 868, 2022, doi: 10.30865/mib.v6i2.3787.

[7] H. Kulina and S. Gocheva-Ilieva, "CART Ensemble and Bagging Algorithm for Estimating of FactorsInfluencing the Furniture Market - Pending to add," no. August, 2023.

[8] W. Imtiyaz, N. Satyahadewi, and H. Perdana, "Application of Bagging Cart in the Classification of on-Time Graduation of Students in the Statistics Study Program of Tanjungpura University," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 4, pp. 2243–2252, 2023, doi: 10.30598/barekengvol17iss4pp2243-2252.