Published online at: **https://journal.rekayasainformatika.com/**

JURNAL REIN
(REKAYASA INFORMATIKA)

Vol. 1 No. 1 (2024) hal 22 - 27

Komparasi Metode KNN Imputation Dan Random Forest Untuk Hasil Klasifikasi Data UMKM

Antonius Wahyu Sudrajat [1], Idham Cholid[2]
Informatics Management Study Programme, Faculty of Computer Science and Engineering [1],
Management Study Programme, Faculty of Economics and Business [2]
Universitas Multi Data Palembang
Palembang, Indonesia
wahyu.sudrajat@mdp.ac.id[1], idham@mdp.ac.id[2]

Abstract

MSMEs play an important role in economic growth in Indonesia. The improvement of MSMEs carried out by the government is based on precise data. Data incompleteness is a problem in managing MSME data. Handling Missing valus in MSME data is important. The imputation method is a method taken in handling missing data. Many researchers have handled missing data with various methods. The purpose of this study is to compare or compare the K-Nearest Neihbor method with Imputation (KNNI with Random Forest) in overcoming missing data on MSME datasets in one of the districts in South Sumatra. The evaluation is done using the score accuracy and mean absolute percentage error (MAPE) methods. Our results show that Random Forest imputation consistently outperforms KNN imputation across various scenarios. Specifically, the Random Forest approach achieved an accuracy score of 0.9958, while the KNN score achieved an accuracy of 0.9916. In addition, using MAPE, Random Forest has a lower average error rate of 0.41%. In future research, it is necessary to further improve the accuracy results by optimizing each method..

Keywords: KNN Imputation, Missing Value, Metode Imputation, Random Forest;

1. Introduction

Data merupakan kumpulan angka, teks, simbol yang belum memiliki makna bagi pemiliknya. Data harus melalui sebuah proses untuk dapat digunakan dalam pengambilan keputusan yang selanjutnya disebut sebagai informasi. Artinya data adalah sumber dan dasar dalam menen tukan kualitas informasi. Missing value dalam sebuah data masih menjadi permasalahan hingga saat ini. Missing data atau missing value merupakan kondisi dimana terdapat nilai yang tidak lengkap atau kosong pada satu atau beberapa kriteria[1]. Menangani data yang hilang merupakan tantangan umum dalam analisis data, sangat penting untuk menanganinya dengan tepat untuk memastikan kualitas dan keandalan hasil analisis.

Penopang perekonomian di Indonesia salah satunya adalah Usaha Menengah, Kecil dan Mikro (UMKM) dengan menyerap tenaga kerja dan nilai produksi yang tinggi [2]. Data UMKM merupakan data penting yang harus dikelola oleh pemerintah sehingga dapat digunakan dalam pengambilan keputusan. Data UMKM dikumpulkan melalui beberapa metode salah satunya adalah dengan menyebarkan kuisioner. Dalam proses pengisi ada yang dilakukan oleh para pelaku UMKM seringkali tidak lengkap, hal ini dapat disebabkan oleh ketidak pahaman akan data yang harus diisi atau pelaku UMKM tidak mau memberikan informasi secara tepat karena dianggap sebagai informasi pribadi[3]. Hal tersebut mengakibatkan data yag dikumpulkan menjadi tidak lengkap dan mengakibatkan informasi dan keputusan yang diambil dapat menjadi salah.

Mengatasi ketidak lengkapan data dapat dilakukan dengan beberapa cara diantaranya adalah menghapus data dan melakukan inputasi. Menghapus data akan mengakibatkan data berkurang dan pelaku UMKM menjadi tidak memiliki kesempat dalam mendapatkan manfaat. Metodel imputasi merupakan solusi yang dapat dilakukan dalam hal ini, terdapat banyak metode imputasi yang dapat digunakan dan telah diteliti oleh banyak para peneliti.

Beberapa penelitian telah menyarankan strategi untuk menangani nilai yang tidak lengkap dalam kumpulan data. Imputasi adalah cara umum untuk menangani kumpulan data yang tidak lengkap. Imputasi data yang hilang berarti mengganti atau memperbaiki kekurangan tersebut dengan nilai yang masuk akal untuk mencapai kelengkapan[4]. Imputasi data yang hilang merupakan langkah penting karena kesalahan pengambilan keputusan akan terjadi ketika didukung oleh data set

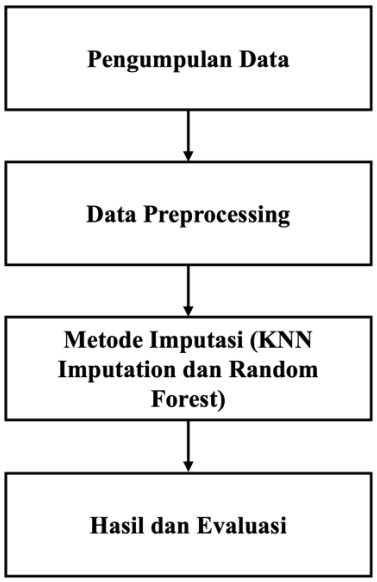
Received: 19-06-2024 | Accepted: 29-06-2024 | Published Online: 25-07-2024

yang tidak lengkap.[5] Pembelajaran mesin (ML) merupakan salah satu model yang diusulkan untuk imputasi data.

Tujuan dalam penelitian ini adalah untuk melakukan komparasi antara metode K-Nearest Neihbor dengan Imputation (KNNI dengan Random Forest dalam mengatasi data yang hilang pada dataset UMKM di salah satu kabupaten di Sumatera Selatan. Selanjutnya evaluasi dilakuka dengan menggunakan metode akurasi skor dan *Mean Absolute Percentage Error* (MAPE). Penelitian ini diharapkan dapat memberikan gambaran bagi peneliti yang lain dalam melakukan penanganan data yang hilang dan khusus nya dalam pemilihan metode ataupun upaya dalam meningkatkan metode yang ada.

2. Research Methods

Dataset yang akan digunakan dalam penelitian ini adalah dataset UMKM di salah satu kabupaten di Sumatera Selatan. Beberapa tahapan pada penelitian ini ditunjukkan pada gambar 1.



Gambar 1 Langkah Penelitian

Langkah awal dari penelitian ini adalah melakukan pengumpulan data. Selanjutnya data preprocessing, langkah ini melakukan perubahan dataset sehingga dapat digunakan dalam proses data mining. Tahapan selanjutnya adalah melakukan imputisasi nilai yang hilang dengan menggunakan metode imputasi KNN Imputation dan Random Forest pada dataset UMKM. Dataset yang telah dilakukan imputisasi dari data yang hilang, selanjutnya dilakukan klasifikasi menggunakan metode KNN, Navie Bayes, SVM. Langkah terakhir adalah melakukan evaluasi menggunakan Accuracy score.

1. Pengumpulan Data

   Tahapan ini merupakan tahapan yang dilakukan setelah memahami kebutuhan bisnis. Data yang dikumpulkan merupakan data UMKM yang dikumpulkan di tahun 2017 oleh salah satu dinas di Provinsi Sumatera Selatan. Data UMKM yang dikumpulkan terdiri dari 798 record, dengan atribut nama UMKM, kategori industri, alamat, kecamatan, NPWP, bidang usaha, acal bahan baku, hasil produksi, nama PIC, No Telp, jumlah tenaga kerja, modal usaha, omset, dan aset yang dimiliki.

2. Data preprocessing

   Preprocessing dilakukan untuk menghilangkan data yang duplikat. Pada penelitian ini yang menjadi fokus adalah data set UMKM yang telah dikumpulkan. Karakteristik dari data set UMKM yang memiliki nilai yang hilang diantaranya adalah omset dan aset, maka data yang digunakan dalam penelitian ini diantaranya adalah atribut nama UMKM, jenis usaha, omzet, aset dan jenis industri. Dataset UMKM yang akan dilakukan proses dapat dilihat pada tabel 1.

Tabel 1 Data UMKM

No	Nama	Jenis Usaha	Omset	Aset	Jensi Industri
1	Pabrik Cincau	Produksi cincau	1.080000e+09	60000000.0	Kecil
2	Pabrik tahu	Produksi tahu	5.000000e+07	20000000.0	Mikro
3	Pabrik roti	Produksi roti	2.400000e+07	NaN	Mikro
4	Warung Kecil	Manisan	1.440000e+08	NaN	Mikro
5	Warung timah	Makanan kecil	1.260000e+07	NaN	Mikro
…	…	…	…	…	…

Berdasarkan data UMKM kemudian dilakukan identifikasi terhadap nilai dari setiap atribut untuk mengetahui karakteristik data khususnya untuk melihat jumlah data yang hilang pada atribut omzet dan azet. Karakteristik data UMKM dapat dilihat pada gambar 2.

Tabel 2 karakteristik Data UMKM

No.	Atribut	Missing	persentase (%)
1	Nama	0	0%
2	Jenis Usaha	0	0%
3	Omzet	22	2,8%
4	Azet	63	7,9%
5	Jenis Industri	0	0%

3. Metode Imputation

   Untuk megatasi nilai yang hilang, dalam penelitian ini digunakan dua metode, yaitu K-Nearest Neighbor Imputation (KNNI) dan Random Forest. Metode tersebut telah banyak digunakan dalam mengatasi nilai yang hillang.

   a) K-Nearest Neighbor Imputation (KNNI)

      Metode ini merupakan metode penanganan missing value yang populer. Salah satu

kelebihan dari metode ini adalah tidak membutuhakan pembentukan model prediksi untuk setiap item data yang memiliki nilai NaN atau hilang. Sementara kelemahan dari metode ini adalah pada saat mencari atau menentukan nilai k yang paling sesuai[6].

Tahapan dalam metode ini adalah sebagai berikut:

1) Menentukan nilai K
2) Menghitung jarak antara observasi yang mengandung mising value pada variabel ke-j dengan observasi lainnya yang tidak mengandung mising value dengan menggunakan rumus :

$$d(X_a, X_b) = \sqrt{\sum_{j=1}^{m}(X_{aj} - X_{bj})^2} \quad (1)$$

3) Mencari K observasi terdekat berdasarkan nilai jarak terkecil
4) Menghitung bobot (weight) pada setiap[K observasi terdekat
5) Menghitung nilai rata-rata pada K obesrvasi terdekat yang tidak mengandung missing value dengan menggunakan prosedur weighted mean estimation yaitu dengan rumus:

$$\widehat{X_j} = \frac{1}{KW} \sum_{K=1}^{K} W_k \, V_{kj} \quad (2)$$

6) Melakukan proses imputasi missing value pada data observasi yang mengandung nilai yang hilang dengan rata-rata yang diperoleh pada langkah sebelumnya.

b) Random Forest

Metode ini dapat meningkatkan hasil akurasi jika terdapat data yang hilang dan untuk resisting outlier[7]. Metode ini merupakan metode yang berbasis klasifikasi regresi dimana terdapat progres agregasi pohon keputusan[8]. Random Forest adalah algoritma supervised learning yang dikeluaran oleh Breiman pada tahun 2001. Rumus random forest adalah sebagai berikut:

4. Evaluasi

a) Acuraccy scor

Merupakan proporsi jumlah prediksi yang benar[9]. Rumus persamaan dari metode ini adalah sebagai berikut[10]:

$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

b) MAPE

Merupakan prosentase kesalahan dari model peramalan atau prediksi. Semakin rendah nilai presentase MAPE, maka semakin rendah pula kesalahan peramalan atau prediksi. Berikut ini adalah rumus untuk mendapatka[11]n MAPE sebagai berikut:

$$MAPE = \frac{\sum_{i=1}^{n} \frac{x_{i-} F_i}{x_i}}{n} 100\% \quad (3)$$

3. Results and Discussions

Dataset dalam penelitian ini terdiri dari data dari 798 bisnis UMKM. Dataset ini memiliki nilai yang hilang dari dua kolom; aset dan omzet. Aset memiliki 63 (7,9%) nilai yang hilang dan omzet memiliki 22 (2,8). Untuk mengimputasi nilai yang hilang tersebut (NaN), penulis menggunakan Imputasi KNN dan Random Forest.

Langkah pertama yang dilakukan adalah dengan menggunakan split train-test 70/30. Metode yang digunakan untuk membagi data adalah dengan menggunakan train test split dari library scikit-learn. Kemudian, menerapkan kedua metode tersebut untuk mengisi nilai yang hilang. Setelah nilai yang hilang terisi, akurasi dari kedua metode tersebut dihitung dengan menggunakan nilai akurasi dan MAPE (Mean Absolute Percentage Error.

3.1 Imputasi

1. Metode Imputasi KNN Imputation

Nilai yang hilang dalam dataset diatasi dengan menggunakan imputasi K-Nearest Neighbors (KNN). Metode ini memanfaatkan konsep kemiripan, mengisi titik data yang hilang berdasarkan informasi dari tetangga terdekatnya. Untuk dapat melakukan proses tersebut, penulis menggunakan library imputer KNN dari scikit-learn. Daftar tetangga terdekat (k) yang digunakan adalah 5, 10, dan 12.

Tabel 1 dataset

	No.	Nama	Jenis Usaha	Omzet	Asset	Jenis Industri
0	1	Pabrik Cincau	Produksi Cincau	1.0800 00e+09	60000 000.0	Kecil
1	2	Pabrik Tahu	Produksi Tahu	5.0000 00e+07	20000 000.0	Micro
2	3	Pabrik Roti	Produksi Roti	2.4000 00e+07	NaN	Micro
3	4	Warung Kecil	Dagang manisan, Sayur, Gas, Makanan	1.4400 00e+08	NaN	Micro
4	5	Warung Timah	Dagang makanan kecil	1.2600 00e+07	NaN	Micro

KNN imputation (k=5) was applied to address missing values in the dataset.

Tabel 2. K=5

	No.	Nama	Jenis Usaha	Omzet	Asset	Jenis Industri
0	1	Pabrik Cincau	Produksi Cincau	108000 0000	60000 0000	Kecil
1	2	Pabrik Tahu	Produksi Tahu	5.0000 000	20000 0000	Micro
2	3	Pabrik Roti	Produksi Roti	240000 000	93000 00	Micro

3	4	Warung Kecil	Dagang manisan, Sayur, Gas, Makanan	144000000	10600000	Micro
4	5	Warung Timah	Dagang makanan kecil	126000000	8560000	Micro

Tabel 3. K=10

	No.	Nama	Jenis Usaha	Omzet	Asset	Jenis Industri
0	1	Pabrik Cincau	Produksi Cincau	1080000000	60000000	Kecil
1	2	Pabrik Tahu	Produksi Tahu	5.0000000	20000000	Micro
2	3	Pabrik Roti	Produksi Roti	240000000	4120000	Micro
3	4	Warung Kecil	Dagang manisan, Sayur, Gas, Makanan	144000000	6690000	Micro
4	5	Warung Timah	Dagang makanan kecil	126000000	6080000	Micro

Tabel 4. K=12

	No.	Nama	Jenis Usaha	Omzet	Asset	Jenis Industri
0	1	Pabrik Cincau	Produksi Cincau	1080000000	60000000	Kecil
1	2	Pabrik Tahu	Produksi Tahu	5.0000000	20000000	Micro
2	3	Pabrik Roti	Produksi Roti	240000000	7391666	Micro
3	4	Warung Kecil	Dagang manisan, Sayur, Gas, Makanan	144000000	7658333	Micro
4	5	Warung Timah	Dagang makanan kecil	126000000	7150000	Micro

Tabel 5 menunjukkan hasil perbandingan inputasi menggunakan metode KNNI untuk masing-masing k, yaitu 5, 10 dan 12. Tabel ini menunjukkan semakin besar jumlah k, maka semakin kecil nilai imputasinya.

Tabel 5. Perbandingan Nilai K

	Asset	Asset K=5	Asset K=10	Asset K=12
0	600000000	600000000	600000000	600000000
1	200000000	200000000	200000000	200000000
2	7391666	9300000	4120000	7391666
3	7658333	10600000	6690000	7658333
4	7150000	8560000	6080000	7150000

2. Metode Imputation Random Forest

Metode lain yang digunakan untuk mengimputasi nilai yang hilang adalah Random Forest. Di sini, penulis juga menggunakan pustaka scikit-learn. Berikut ini adalah perbandingan antara data saat ini dan data setelah komputasi:

Tabel 6. Sebelum imputasi

	No.	Nama	Jenis Usaha	Omzet	Asset	Jenis Industri
0	1	Pabrik Cincau	Produksi Cincau	1.080000e+09	60000000.0	Kecil
1	2	Pabrik Tahu	Produksi Tahu	5.000000e+07	20000000.0	Micro
2	3	Pabrik Roti	Produksi Roti	2.400000e+07	NaN	Micro
3	4	Warung Kecil	Dagang manisan, Sayur, Gas, Makanan	1.440000e+08	NaN	Micro
4	5	Warung Timah	Dagang makanan kecil	1.260000e+07	NaN	Micro

Tabel 7. Setelah imputasi

	No.	Nama	Jenis Usaha	Omzet	Asset	Jenis Industri
0	1	Pabrik Cincau	Produksi Cincau	108000000	60000000	Kecil
1	2	Pabrik Tahu	Produksi Tahu	500000000	20000000	Micro
2	3	Pabrik Roti	Produksi Roti	240000000	17540000	Micro
3	4	Warung Kecil	Dagang manisan, Sayur, Gas, Makanan	144000000	14382500	Micro
4	5	Warung Timah	Dagang makanan kecil	126000000	14930000	Micro

Tabel 8 menunjukkan nilai inputan ketika menggunakan metode random forest.

Tabel 8. Perbandingan nilai pada Random Forest

	Asset	Asset
0	60000000.0	60000000
1	20000000.0	20000000
2	NaN	17540000
3	NaN	14382500
4	NaN	14930000

3. Perbandingan Metode KNN Imputation dan Random Forest

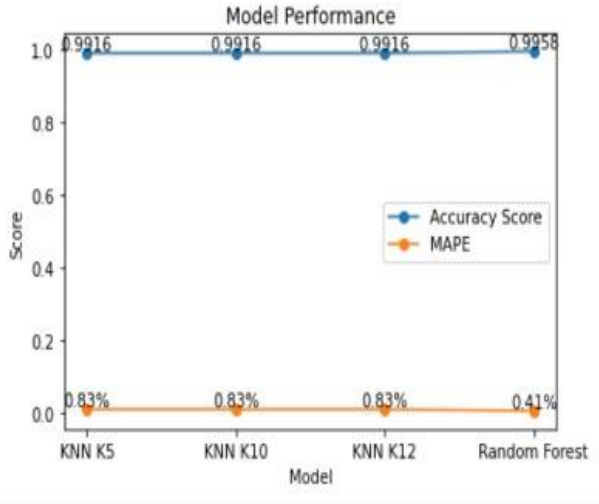
Untuk menilai efektivitas strategi imputasi yang berbeda, penulis membandingkan kinerja Random Forest dan K-Nearest Neighbors (KNN) imputation

dengan menggunakan skor akurasi dan MAPE sebagai metrik.

Akurasi mencerminkan proporsi nilai yang dihitung dengan benar dalam set data. Skor akurasi untuk KNNI dengan jumlah tetangga terdekat 5, 10, dan 12 adalah 0.9916. Di sisi lain, nilai akurasi untuk Random Forest adalah 0.9958.

Kemudian, kinerja K-Nearest Neighbors (KNN) Imputation dan Random Forest dievaluasi dengan menggunakan Mean Absolute Percentage Error (MAPE). Model KNN dengan k=5, k=10, dan k=12 menunjukkan MAPE yang sama yaitu sebesar 0.83%. Hal ini menunjukkan bahwa meningkatkan nilai k dalam rentang ini tidak berdampak signifikan terhadap akurasi model dalam skenario spesifik ini.

Random Forest mengungguli model KNNI dengan mencapai MAPE yang lebih rendah, yaitu 0,41%. Hal ini mengindikasikan bahwa, secara rata-rata, prediksi model Random Forest lebih mendekati nilai aktual dibandingkan dengan model KNN.



Gambar 1 Perbandingan

3.2 Discussions

Setelah dilakukan proses imputasi, data UMKM yang mengalami missing values berhasil diimputasi, baik dengan menggunakan metode KNNI maupun Random Forest. Hasil penelitian kami menunjukkan bahwa imputasi Random Forest secara konsisten mengungguli imputasi KNN di berbagai skenario. Secara khusus, pendekatan Random Forest mencapai skor akurasi 0,9958, sementara skor KNN mencapai akurasi 0,9916. Selain itu, dengan menggunakan MAPE, Random Forest memiliki tingkat kesalahan rata-rata yang lebih rendah, yaitu sebesar 0.41%.

4. Conclusions

Berdasarkan proses uji coba terhadap data dengan menggunakan metode yang telah ditentukan sebelumnya maka peneliti menyimpulkan bahwa penelitian ini menguji efektivitas algoritma random forest dan K-Nearest Neighbours (KNN) untuk mengimplikasikan data yang hilang. Setelah mengevaluasi kedua metode tersebut dengan menggunakan nilai akurasi dan Mean Absolute Percentage Error (MAPE), hasilnya menunjukkan bahwa pendekatan random forest mengungguli KNN dalam konteks ini. Temuan ini menunjukkan bahwa imputasi random forest mungkin merupakan pilihan yang lebih sesuai ketika bertujuan untuk menjaga akurasi data dan meminimalkan kesalahan.

Acknowledgements

Penulis mengucapkan terima kasih kepada Kepala LPPM Universitas Multi Data Palembang yang telah memberi dukungan sehingga penelitian ini dapat terlaksana.

References

[1] W. Sudrajat and I. Cholid, “K-NEAREST NEIGHBOR (K-NN) UNTUK PENANGANAN MISSING VALUE PADA DATA UMKM,” 2023.

[2] idham C. ermatita Wahyu Sudrajat, “Application of the Apriori Algorithm and FP-Growth to tind out the Association Rule between Gender, Education level on wages of SMEs workers in palembang City,” *Journal of Small Business and Entrepreneurship Development*, vol. 4, no. 1, 2016, doi: 10.15640/jsbed.v4n1a3.

[3] G. Doquire and M. Verleysen, “Feature selection with missing data using mutual information estimators,” *Neurocomputing*, vol. 90, pp. 3–11, Aug. 2012, doi: 10.1016/j.neucom.2012.02.031.

[4] T. Thomas and E. Rajabi, “A systematic review of machine learning-based missing value imputation techniques,” *Data Technologies and Applications*, vol. 55, no. 4, pp. 558–585, 2021, doi: 10.1108/DTA-12-2020-0298.

[5] A. R. Ismail, N. Z. Abidin, and M. K. Maen, “Systematic Review on Missing Data Imputation Techniques with Machine Learning Algorithms for Healthcare,” Mar. 01, 2022, *Department of Electrical Engineering, Universitas Muhammadiyah Yogyakarta*. doi: 10.18196/jrc.v3i2.13133.

[6] S. Y. Siregar, S. St, T. Toharudin, B. Tantular, S. Si, and M. Si, “PERFORMA METODE K NEAREST NEIGHBOR IMPUTATION (KNNI) UNTUK MENANGANI MULTIVARIATE MISSING DATA.”

[7] R. Supriyadi, W. Gata, N. Maulidah, A. Fauzi, I. Komputer, and S. Nusa Mandiri Jalan Margonda Raya No, “Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah,” vol. 13, no. 2, pp. 67–75, 2020, [Online]. Available: http://journal.stekom.ac.id/index.php/E-Bisnis■page67

[8] M. Dan *et al.*, “Application of Random Forest Method to Identify Food and Beverage Industries Experiencing Raw Material Difficulties Penerapan Metode Random Forest untuk Mengidentifikasi Industri,” *Indonesian Journal of Statistics and Its Applications*, vol. 8, no. 01, pp. 37–46, 2024, doi: 10.29244/ijsa.v8i1p37-46.

[9] F. Yulian Pamuji, Ahmad Rofiqul Muslikh, Rizza Muhammad Arief, and Delviana Muti, "Komparasi Metode Mean dan KNN Imputation dalam Mengatasi Missing Value pada Dataset Kecil," *Jurnal Informatika Polinema*, vol. 10, no. 2, pp. 257–264, Feb. 2024, doi: 10.33795/jip.v10i2.5031.

[10] L. Amatullah, Y. Widiastiwi, and N. Chamidah, "Penerapan Klasifikasi Random Forest Terhadap Data Gangguan Spektrum Autisme (ASD) Pada Anak-Anak Menggunakan Seleksi Fitur Principal Component Analysis".

[11] A. Fadlil, Herman, and D. Praseptian M, "K Nearest Neighbor Imputation Performance on Missing Value Data Graduate User Satisfaction," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 4, pp. 570–576, 2022, doi: 10.29207/resti.v6i4.4173.

