Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2022, V (3): 595 – 601 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

KOMPARASI RANDOM FOREST DAN LOGISTIC REGRESSION DALAM KLASIFIKASI PENDERITA COVID-19 BERDASARKAN GEJALANYA

Ichsan Firmansyah[1], Jaka Tirta Samudra[2], Doughlas Pardede[3], Zakarias Situmorang[4]
Universitas Potensi Utama, Medan
e-mail: ichsanfirmansyah1989@gmail.com[1], jakatirta135@gmail.com[2], doug.pardede@gmail.com[3], zakarias65@yahoo.com[4]

Abstract: *In data mining, we can use symptoms suffered by patients for a reference in classifying positive and negative Covid-19 patients using data mining. Random Forest and logistic regression are two data mining algorithms with high accuracy, precision, and sensitivity in data classification. This study compares the random forest and the logistic regression algorithm - where we use the lasso and ridge regulations - on classifying positive and negative Covid-19 patients based on their symptoms. From 5434 data used in the data set, the evaluation results show that the random forest algorithm is the best in terms of accuracy, precision, and sensitivity compared to other algorithms, while the logistic regression algorithm with ridge regulation is the worst. The random forest algorithm is the most reliable in classifying patients with positive Covid-19, while the logistic regression algorithm with ridge regulation is the least reliable. Also, the random forest algorithm is the most reliable in classifying patients with negative Covid-19, while the logistic regression algorithm with lasso regulation is the least reliable.*

Keywords: *classification;covid-19;data mining;logistic regression;random forest.*

Abstrak: Dalam data mining, kita dapat menggunakan gejala yang diderita pasien sebagai acuan dalam mengklasifikasikan pasien positif dan negatif Covid-19 menggunakan data mining. Random forest dan logistic regression adalah dua algoritma data mining yang memiliki akurasi (accuracy), presisi (precision), dan sensitivitas (recall) tinggi dalam klasifikasi data. Penelitian ini membandingkan algoritma random forest dan logistic regression - di mana kami menggunakan regulasi lasso dan ridge - dalam mengklasifikasikan pasien positif dan negatif Covid-19 berdasarkan gejalanya. Dari 5434 data yang digunakan dalam data set, hasil evaluasi menunjukkan bahwa algoritma random forest adalah yang terbaik dalam hal akurasi, presisi, dan sensitivitas dibandingkan dengan algoritma lainnya, sedangkan algoritma logistic regression dengan regulasi ridge adalah yang terburuk. Algoritma random forest paling andal dalam mengklasifikasikan pasien positif Covid-19, sedangkan algoritma logistic regression dengan regulasi ridge merupakan algoritma yang paling tidak tidak dapat diandalkan. Selain itu, algoritma random forest paling andal dalam mengklasifikasikan pasien dengan Covid-19 negatif, sedangkan algoritma logistic regresssion dengan regulasi lasso merupakan yang paling tidak dapat diandalkan.

Kata kunci: covid-19;data mining;klasifikasi;logistic regression;random forest.

PENDAHULUAN

Covid-19, penyakit yang disebab-kan oleh virus SARS-Cov-2, sejak pertama kali ditemukan pada tanggal 29 Juni 2021 telah menjadi pandemi di seluruh belahan dunia, khususnya di Indonesia (Nurjannah, Dar, Bangun, 2021). Gejala yang umum dialami oleh pasien Covid-19 adalah demam, batuk, bersin, sesak nafas, batuk kering, kelelahan, sakit tenggorokan, nyeri

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2022, V (3): 595 – 601 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

kepala, pilek (Susilo et. al, 2020). Data gejala yang dialami pasien ini dapat ditambang menggunakan teknik data mining sehingga dapat membantu diagnosa awal positif Covid-19 pada pasien.

Data mining merupakan teknik untuk menganalisa sebuah basis data dan menemukan pengetahuan yang berguna di dalamnya (*knowledge discovery in database*) (Erwansyah et. al, 202). Pencarian pengetahuan dalam data mining berfungsi untuk menyelesaikan masalah seperti asosiasi (*association*), pengklusteran (*clustering*), prediksi (*prediction*), estimasi (*estimation*) dan klasifikasi (*classification*) (Boy, 2020).

Random forest, sebagai salah satu algoritma klasifikasi yang mudah untuk diaplikasikan, beban komputasinya rendah namun tetap memiliki akurasi yang tinggi (Rianto & Yunis, 2021), mengaplikasikan bagging dan pemilihan fitur secara acak pada setiap iterasinya untuk membentuk pohon-pohon yang menghasilkan klasifikasi data dari pohon dengan nilai voting tertinggi (Triscowati et. al, 2019). Hasil klasifikasi algoritma ini cukup baik saat dievaluasi menggunakan teknik k-fold cross validation, dengan nilai k sebesar 10, mencapai nilai akurasi 76.2% dan presisi 77,3% (Yusuf et. al, 2020).

Algoritma logistic regression juga memiliki akurasi dan ketepatan yang cukup baik dalam hal akurasi dan ketepatan, dengan capaian nilai akurasi sebesar 75,97% dan presisi sebesar 76,92% pada klasifikasi penyakit diabetes (Kurniadi, & Putri, 2018). Algoritma ini menggunakan distribusi Bernoulli untuk pengamatan variabel respon, berupa kategori biner (0 dan 1), yang dipengaruhi oleh variabel prediktor dalam sebuah model regresi pada proses klasifikasinya (Purwa, 2019).

Penelitian ini membandingkan algoritma *random forest* dengan *logistic regression* dalam mengklasifikasikan penyakit Covid-19 berdasarkan gejala pasien dan mengevaluasi hasilnya dengan k-Fold Cross Validation, nilai k sebesar 10. Performa kedua algoritma dievaluasi menggunakan *confusion metrix* untuk melihat algoritma mana yang lebih baik dalam mengklasifikasikan kelas *true positive, true negative, false positive* dan *false negative*.

METODE

Tabel 1. Sampel Data Set

SB	D	BK	ST	P	A	PK	SK	PJ	D	HT	K	PP	Covid
Yes	Yes	Yes	No	No	Yes	No	No	No	No	No	Yes	No	Yes
Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	No	No	No	Yes
Yes	Yes	Yes	Yes	No	Yes	No	No	No	No	Yes	No	Yes	Yes
Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
Yes	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes	No	No	Yes	No
Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	No
Yes	Yes	No	No	No	Yes	Yes	Yes	No	No	No	Yes	Yes	No
Yes	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	No
Yes	No	No	Yes	No	No	No	No	No	No	Yes	Yes	Yes	No
Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	Yes	Yes	No

Keterangan : SB = Susah Bernafas, D = Demam, BK = Batuk Kering, ST = Sakit Tenggorokan, P = Pilek, A = Asma, PK = Penyakit Paru Kronis, SK = Sakit Kepala, PJ = Penyakit Jantung, D = Diabetes, HT = Hipertensi, K = Kelelahan, PP = Penyakit Pencernaan, Covid = Menderita Covid-19

Tabel 2. Sampel Hasil Normalisasi

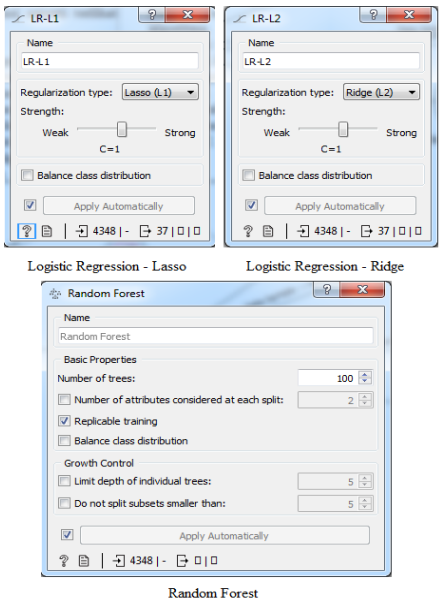
SB	D	BK	ST	P	A	PK	SK	PJ	D	HT	K	PP	Covid
1	1	1	0	0	1	0	0	0	0	0	1	0	Yes
1	1	1	0	1	0	1	0	1	1	0	0	0	Yes
1	1	1	1	0	1	0	0	0	0	1	0	1	Yes
1	1	1	0	1	1	1	1	0	0	0	1	1	Yes
1	1	1	0	0	1	0	1	1	1	0	0	1	No
1	0	1	1	1	0	1	0	1	0	1	1	1	No
1	1	0	0	0	1	1	1	0	0	0	1	1	No
1	0	0	0	0	0	0	0	1	1	1	1	1	No
1	0	0	1	0	0	0	0	0	0	1	1	1	No
1	0	1	1	1	0	1	0	1	0	1	1	1	No

Data set kemudian dinormalisasi dengan mengubah nilai Yes pada masing-masing fitur dengan nilai 1 dan dengan nilai 0, dengan sampel hasil normalisasi seperti terlihat pada Tabel 2.

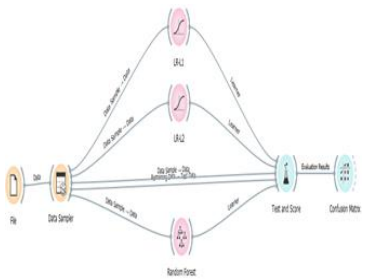
Klasifikasi data hasil normalisasi ini menggunakan aplikasi *Orange 3.30*,

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2022, V (3): 595 – 601 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

dengan memanfaatkan *widget* File, Data Sampler, Learner, serta Test and Score untuk membentuk model klasifikasi seperti terlihat pada Gambar 1. *Widget* File berfungsi untuk membaca file data set yang digunakan. *Widget* Data Sampler untuk memilih secara acak data *training* dan data *test* yang digunakan dalam proses klasifikasi. *Widget* Learner terdiri dari LR-L1 sebagai *classifier* algoritma *logisctic regression* dengan regulasi *lasso*, LR-L2 sebagai *classifier* algoritma *logisctic regression* dengan regulasi *ridge* dan Random Forest sebagai *classifier* algoritma *random forest*, dengan konfigurasi masing-masing Learner seperti telihat pada Gambar 2. *Widget* Test and Score sebagai evaluasi akurasi (*accuracy*), presisi (*precision*) dan sensitifitas (*recall*) algoritma. *Widget* Confusion Matrix sebagai evaluasi performa kedua algoritma.



Gambar 1. Konfigurasi *Widget Learner*



Gambar 2. Model Klasifikasi

Hasil klasifikasi kedua algoritma kemudian dievaluasi pada widget Test and Score untuk melihat nilai accuracy, precision dan recall kedua algoritma.
Widget Confusion Matrix mengevaluasi keandalan kedua algoritma berdasarkan nilai false positive rate dan false negative rate.

Nilai *accuracy*, *precision* dan *recall* dihitung menggunakan persamaan (1) sampai (3) sedangkan nilai false positive rate dan false negative rate dihitung menggunakan persamaan (4) dan (5) (Markoulidakis et. al, 2021):

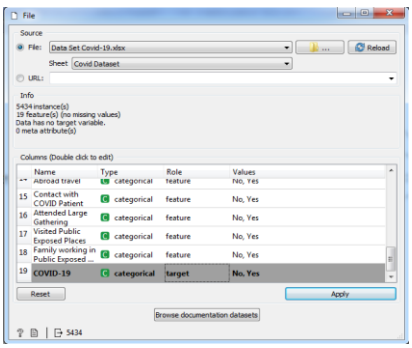
$$NA = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$
$$NP = TP/(TP+FP) \quad (2)$$
$$NR = TP/(TP+FN) \quad (3)$$
$$FPR = FP/(TN+FP) \quad (4)$$
$$FNR = FN/(TP+FN) \quad (5)$$

Keterangan:

NA	=	Nilai accuracy
NP	=	Nilai precision
NR	=	Nilai recall
FPR	=	False Positive rate
FNR	=	False Negative rate
TP	=	True Positive
TN	=	True Negative
FP	=	False Positive
FN	=	False Negative

HASIL DAN PEMBAHASAN

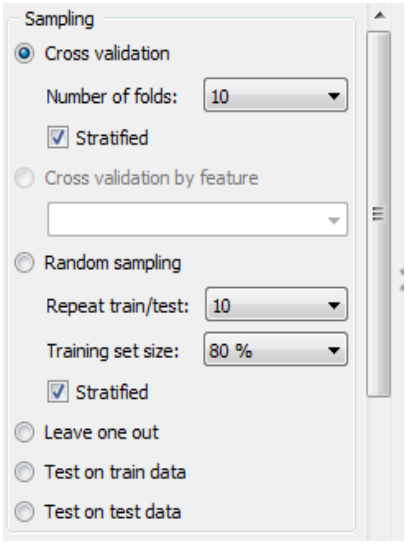
Data set dibuka menggunakan *widget* File, dan dipilih Covid-19 sebagai target sehingga diperoleh informasi fitur dan target di dalam data set seperti terlihat pada Gambar 3.



Gambar 3. Informasi Data Set

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2022, V (3): 595 – 601 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Klasifikasi dilakukan secara otomatis oleh aplikasi *Orange 3.30* dengan hasil yang terlihat pada *widget* Test and Score dengan menggunakan konfigurasi *10-fold cross validation* sebagai metode evaluasi, seperti terlihat pada Gambar 4.



Gambar 4. Konfigurasi *Widget* Test and Score

Hasil pada *widget* Test and Score ini merupakan nilai *accuracy*, *precision* dan *recall* kedua algoritma seperti terlihat pada Gambar 5.

Evaluation Results

Model	CA	Precision	Recall
LR-L1	0.970	0.969	0.970
LR-L2	0.969	0.968	0.969
Random Forest	0.982	0.982	0.982

Gambar 5. Hasil *Cross Validation*

Dari Gambar 5, terlihat bahwa algoritma *random forest* memiliki nilai *accuracy*, *precision* dan *recall* tertinggi dibandingkan algoritma *logistic regression* dengan regulasi *lasso* dan *logistic regression* dengan regulasi *ridge*, sedangkan nilai *accuracy*, *precision* dan *recall* terendah dimiliki oleh algoritma *logistic regression* dengan regulasi *ridge*.

Performa kedua algoritma dalam mengklasifikasikan data set yang digunakan terlihat pada *widget* Confusion Matrix, dimana diperoleh nilai prediksi untuk kelas *true negative*, *false positive*, *false negative* dan *true positive* kedua algoritma seperti terlihat pada Gambar 6 sampai Gambar 8.

		Predicted No	Predicted Yes	Σ
Actual	No	758	79	837
	Yes	53	3458	3511
	Σ	811	3537	4348

Gambar 6. Hasil *Confusion Matrix Logistic Regression-Lasso*

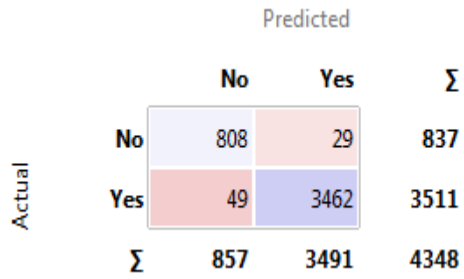
Dari hasil *confusion matrix* algoritma *logistic regression* dengan regulasi *lasso* pada Gambar 6, terlihat bahwa dari 837 data status pasien bukan Covid-19, berhasil diprediksi dengan benar 758 data sedangkan 79 data sisanya gagal. Sedangkan dari 3511 data status pasien penderita Covid-19, berhasil diprediksi 3458 data dengan benar sedangkan 53 data sisanya gagal.

		Predicted No	Predicted Yes	Σ
Actual	No	751	86	837
	Yes	50	3461	3511
	Σ	801	3547	4348

Gambar 7. Hasil *Confusion Matrix Logistic Regression-Ridge*

Dari hasil *confusion matrix* algoritma *logistic regression* dengan regulasi *ridge* pada Gambar 7, terlihat bahwa dari 837 data status pasien bukan Covid-19, berhasil diprediksi dengan benar 751 data sedangkan 86 data sisanya gagal. Sedangkan dari 3511 data status pasien penderita Covid-19, berhasil diprediksi 3461 data dengan benar sedangkan 50 data sisanya gagal.

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
Oct 2022, V (3): 595 – 601 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR



Gambar 8. Hasil *Confusion Matrix Random Forest*

Dari hasil *confusion matrix* algoritma *random forest* pada Gambar 8, terlihat bahwa dari 837 data status pasien bukan Covid-19, berhasil diprediksi dengan benar 808 data sedangkan 29 data sisanya gagal. Sedangkan dari 3511 data status pasien penderita Covid-19, berhasil diprediksi 3462 data dengan benar sedangkan 49 data sisanya gagal.

Berdasarkan nilai dari Gambar 6 sampai Gambar 8, dibentuk tabel untuk masing-masing nilai *true negative*, *false positive*, *false negative* dan *true positive* kedua algoritma, seperti terlihat pada Tabel 3.

Tabel 3. Perbandingan Hasil Confusion Matrix

Algoritma	TN	FP	FN	TP
LR-L1	758	79	53	3458
LR-L2	751	86	50	3461
Random Forest	808	29	49	3462

Keterangan: LR-L1 = *Logistic Regression* dengan regulasi *Lasso*, LR-L2 = *Logistic Regression* dengan regulasi *Ridge*, TN = Nilai *True Negative*, FP = Nilai *False Positive*, FN = Nilai *False Negative*, TP = Nilai *True Positive*

Dengan menggunakan data pada Tabel 3, dihitung nilai *false positive rate* dan *false negative rate* untuk mengevaluasi keandalan masing-masing algoritma sebagai berikut:

$$FPR_{LRL1} = \frac{79}{758+79}\ X\ 100\% = 9{,}438\%$$

$$FPR_{LRL2} = \frac{86}{751+86}\ X\ 100\% = 10{,}274\%$$

$$FPR_{RF} = \frac{29}{808+29}\ X\ 100\% = 3{,}464\%$$

$$FNR_{LRL1} = \frac{53}{3458+53}\ X\ 100\% = 1{,}509\%$$

$$FNR_{LRL2} = \frac{50}{3461+50}\ X\ 100\% = 1{,}424\%$$

$$FNR_{RF} = \frac{49}{3462+49}\ X\ 100\% = 1{,}395\%$$

Keterangan:

FPR_{LRL1} = *False Positive rate* algoritma *logistic regression* dengan regulasi *lasso*

FPR_{LRL2} = *False Positive rate* algoritma *logistic regression* dengan regulasi *ridge*

FPR_{RF} = *False Positive rate* algoritma *random forest*

FNR_{LRL1} = *False Negative rate* algoritma *logistic regression* dengan regulasi *lasso*

FNR_{LRL2} = *False Negative rate* algoritma *logistic regression* dengan regulasi *ridge*

FNR_{RF} = *False Negative rate* algoritma *random forest*

Dari hasil perhitungan di atas, dibentuk tabel untuk membandingkan nilai *false positive rate* dan *false negative rate* masing-masing algoritma, seperti terlihta pada Tabel 4.

Tabel 4. Perbandingan Hasil Confusion Matrix

Algoritma	FPR (%)	FNR (%)
LR-L1	9,438	1,509
LR-L2	10,274	1,424
Random Forest	3,464	1,395

Berdasarkan nilai pada Tabel 4, terlihat bahwa algoritma memiliki nilai *false positive rate* dan *false negative rate* terendah dibandingkan algoritma *logistic regression* dengan regulasi *lasso* dan *logistic regression* dengan regulasi *ridge*. Algoritma *logistic regression* dengan regulasi *ridge* memiliki nilai *false positive rate* tertinggi dibandingkan algoritma *logistic regression* dengan regulasi *lasso*

dan *random forest*. Algoritma *logistic regression* dengan regulasi *lasso* memiliki nilai *false negative* tertinggi dibandingkan algoritma *logistic regression* dengan regulasi *ridge* dan *random forest.*

SIMPULAN

Dari hasil klasifikasi 5434 data pasien gengan gejala Covid-19 menggunakan *algoritma logistic regression dengan regulasi lasso, logistic regression* dengan regulasi *ridge* dan *random forest*, diperoleh bahwa dalam mengklasifikasi data set yang digunakan, algoritma dengan akurasi, ketepatan dan sensitifitas tertinggi adalah algoritma random forest, sedangkan algoritma *logistic regression* dengan regulasi *ridge* memiliki nilai akurasi, ketepatan dan sensitifitas terendah.

Dari segi keandalan, algoritma random forest lebih unggul dibandingkan *algoritma logistic regression* dengan regulasi lasso maupun algoritma *logistic regression* dengan regulasi ridge. *Algoritma logistic regression* dengan regulasi ridge memiliki keandalan terendah dibandingkan algoritma logistic regression dengan regulasi lasso dan random forest dalam mengklasifikasikan data pasien penderita Covid-19. *Algoritma logistic regression* dengan regulasi lasso memiliki keandalan terendah dibandingkan algoritma logistic regression dengan regulasi ridge dan random forest dalam mengklasifikasikan data pasien bukan penderita Covid-19.

DAFTAR PUSTAKA

Nurjannah, Dar, M. H., Bangun, B. (2021) Sistem Pelacakan Kontak Covid-19 Menggunakan Teknologi QR Code Berbasis Web. *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*. 7(3): 283-292.

Susilo, A., Rumende, C. M., Pitoyo, C. W., Santoso, W. D., Yulianti, M., Herikurniawan, Sinto, R., Singh, G., Nainggolan, L., Nelwan, E. J., Chen, L. K., Widhani, A., Wijaya, E., Wicaksana, B., Maksum, M., Annisa, F., Jasirwan, C. O. M., Yunihastuti, E. (2020) Coronavirus Disease 2019: Tinjauan Literatur Terkini. *Jurnal Penyakit Dalam Indonesia*. 7(1): 45-67.

Erwansyah, K. Purwadi, Saniman, Syahputra, T. (2021) Penerapan Data Mining Untuk Mendapatkan Paket Promo Perlengkapan Pesta Menggunakan Algoritma Apriori Di Celebration Peak. *Journal of Science and Social Research*. 4(2): 96-105.

Boy, A. F. (2020) Implementasi Data Mining Dalam Memprediksi Harga Crude Palm Oil (CPO) Pasar Domestik Menggunakan Algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatera Utara). *Journal of Science and Social Research*. 3(2): 78-85.

Rianto, M. & Yunis, R. (2021). Analisis Runtun Waktu Untuk Memprediksi Jumlah Mahasiswa Baru Dengan Model Random Forest. *Cyberspace: Paradigma*. 23(1): 70-74.

Triscowati, D. W., Sartono, B., Kurnia, A., Dirgahayu, D. & Wijayanto, A. W. (2019). Classification Of Rice-Plant Growth Phase Using Supervised Random Forest Method Based On Landsat-8 Multitemporal Data. *International Journal of Remote Sensing and Earth Sciences*. 16(2): 187-196.

Yusuf, B., Qalbi, M., Basrul, Dwitawati, I., Malahayati & Ellyadi, M. (2020). Implementasi Algoritma Naive Bayes Dan Random Forest Dalam Memprediksi Prestasi Akademik Mahasiswa Universitas Islam Negeri Ar-Raniry Banda Aceh. *Cyberspace: Jurnal Pendidikan Teknologi Informasi*. 4(1): 50-58.

Kurniadi, F. I., Putri, V. K. (2018). Perbandingan Regresi Linear dengan Heaviside Activation Function dengan Logistic Regression untuk Klasifikasi Diabetes. *ULTIMATICS*. 10(1): 7-10.

Purwa, T. (2019). Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017). *Jurnal Matematika, Statistika Dan Komputasi (Jmsk)*. 16(1): 58-73.

Markoulidakis, I., Rallis, I., Georgoulas, I., Kopsiaftis, G., Doulamis, A., Doulamis, N. (2021). Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies*. 9(81): 58-73.