

KOMPARASI ALGORITMA NAIVE BAYES, RANDOM FOREST DAN SVM UNTUK MEMPREDIKSI NIAT PEMBELANJA ONLINE

Cucu Ika Agustyaningrum¹, Windu Gata², Ridan Nurfalalah³, Ummu Radiyah⁴,
Mawadatul Maulidah⁵

^{1,2,5}Program Studi Ilmu Komputer, STMIK Nusa Mandiri

³Program Sistem Informasi, STMIK Nusa Mandiri

⁴Program Teknik Informatika, STMIK Nusa Mandiri

Jl. Kramat Raya No. 18, Jakarta Pusat, Jakarta

Telp. (021) 31908575

e-mail :

¹14002365@nusamandiri.ac.id, ²windu@nusamandiri.ac.id, ³ridan.rlh@nusamandiri.ac.id,

⁴ummu.urd@nusamandiri.ac.id, ⁵14002373@nusamandiri.ac.id

ABSTRACT

In recent years, the use of e-commerce or online shops has increased considerably. Various online stores have sprung up on the internet, both small and large-scale. This has a very important effect on the effective use of time and the level of sales figures. Therefore, e-commerce or online stores must have the ability to assess the means used to identify and classify online shopping intentions to generate profits for the store. Classification of online shopper intentions can be carried out using several algorithms, such as Naive Bayes, Random Forest, and Support Vector Machine. In this study, the comparison of algorithms was carried out using the WEKA application by knowing the value of the F1-Score, Accuracy, Kappa Statistic, and Mean Absolute Error. There is a difference between the test results, for the F1-Score, Accuracy, Kappa Statistic results in the Random Forest algorithm test which is the best compared to Naive Bayes and Support Vector Machine. While the Mean Absolute Error test results for the Support Vector Machine algorithm are the best values than Naive Bayes and Random Forest. So based on this research the Random Forest Algorithm is the best and appropriate algorithm to be applied as a classification of online shopper intentions because the Random Forest algorithm dominates in knowing the value of criteria such as F1-Score, Accuracy, Kappa Statistic, and Mean Absolute Error.

Keywords — Algorithm, E-Commerce, Machine Learning, Shopping Intention, Weka.

ABSTRAK

Beberapa tahun terakhir ini, penggunaan *e-commerce* atau toko *online* sangat meningkat. Berbagai macam toko *online* yang bermunculan di internet, baik berskala kecil maupun yang berskala besar. Hal ini memiliki pengaruh yang sangat penting pada penggunaan waktu yang efektif dan tingkat angka penjualan. Maka dari itu *e-commerce* atau toko *online* harus mempunyai kemampuan menilai sarana yang digunakan untuk mengetahui dan mengklasifikasikan niat pembelian *online* sehingga menghasilkan keuntungan bagi toko tersebut. Niat pembelian *online* dapat dilakukan pengklasifikasian menggunakan beberapa algoritma, seperti *Naive Bayes*, *Random Forest* dan *Support Vector Machine*. Dalam penelitian ini perbandingan algoritma dilakukan menggunakan aplikasi WEKA dengan mengetahui nilai *F1-Score*, *Akurasi*, *Kappa Statistic* dan *Mean*

Absolute Error. Terdapat perbedaan antara hasil pengujian, untuk nilai *F1-Score*, *Akurasi*, *Kappa Statistic* menghasilkan pengujian algoritma *Random Forest*-lah yang paling baik dibandingkan *Naive Bayes* dan *Support Vector Machine*. Sedangkan pada nilai *Mean Absolute Error* hasil pengujian algoritma *Support Vector Machine* merupakan nilai terbaik dari pada *Naive Bayes* dan *Random Forest*. Sehingga berdasarkan penelitian ini Algoritma *Random Forest* merupakan algoritma yang paling baik dan tepat untuk diterapkan sebagai pengklasifikasian niat pembelanja *online*, karena algoritma *Random Forest* yang paling mendominasi dalam mengetahui nilai kriteria seperti *F1-Score*, *Akurasi*, *Kappa Statistic* dan *Mean Absolute Error*.

Kata Kunci—Algoritma, *E-Commerce*, *Machine Learning*, Niat Belanja, Weka.

I. PENDAHULUAN

Beberapa tahun terakhir ini, penggunaan *e-commerce* atau toko *online* sangat meningkat. Berbagai macam toko *online* yang bermunculan di internet, baik berskala kecil maupun yang berskala besar. Bahkan toko *offline* pun yang dulu hanya membuka di toko/ruko sekarang beralih ke toko *online* dan menutup toko/ruko nya. Sehingga konsumen dapat mencari barang selama 24 jam, bahkan dari luar kota maupun luar negeri, begitu juga hal nya penjual mendapatkan konsumen di seluruh dunia.

Pengalaman ini memiliki pengaruh yang sangat penting pada penggunaan waktu yang efektif dan tingkat angka penjualan. Banyak perusahaan *e-commerce* berinventasi dalam deteksi dini (Sakar, S. Olcay Polat, Katircioglu, & Kastro, 2018). Walaupun *e-commerce* ini banyak bermunculan dan sedang tenar, akan tetapi banyak juga masyarakat tidak semuanya tertarik. Banyak toko *online*

yang gagal dan tidak berhasil untuk memperoleh keuntungan, karena konsumen merasa toko *online* tersebut menjadi sarana jual/beli yang kurang menarik (Christian, 2019).

Dapat disimpulkan bahwa kenyamanan layanan toko *online* menjadi penentu perilaku konsumen (Khan & Khan, 2018). Dan juga rekomendasi verbal dari keluarga dekat, kerabat, teman atau bahkan penilaian pada bintang di toko tersebut (Liao, Lin, Luo, & Chea, 2016).

Toko *online* harus mempunyai kemampuan untuk menilai sarana yang digunakan, sehingga dapat mengetahui dan mengklasifikasikan niat pembelanja *online*, maka diperoleh keuntungan dari hasil penjualan. Kemampuan ini dapat di aplikasikan pada toko *online* dengan cara menggunakan algoritma *machine learning*.

Dengan menggunakan dua modul jenis algoritma *machine learning* dan prediksi yang paling baik, dalam hal ini algoritma yang digunakan yaitu algoritma *multilayer perceptron* (Sakar et al., 2018).

Pada penelitian kali ini fokus dalam perbandingan algoritma *machine learning* manakah yang paling baik, untuk menilai niat dan potensi belanja konsumen dengan menggunakan *tools* WEKA, sehingga perbandingan algoritma mendapatkan hasil akurat dan yang paling baik, dan toko tersebut memperoleh pendapatan yang sesuai.

Data Mining

Data mining merupakan salah satu cabang ilmu komputer yang populer. Menurut David Hand, Heikki Mannila dan Padharic Smyth dari MIT dalam jurnal (Nurachim, 2019) *Data Mining* merupakan analisa terhadap data (biasanya data yang berukuran besar) untuk menemukan hubungan yang jelas serta menyimpulkan yang belum diketahui sebelumnya dengan cara dipahami dan berguna bagi pemilik data tersebut.

Machine Learning

Istilah Machine Learning mengacu pada deteksi otomatis dari pola yang bermakna dalam data. *Machine learning* adalah salah satu cara manusia untuk mengajarkan sesuatu kepada computer. Tanpa melakukan pemrograman tertentu secara eksplisit komputer dapat belajar untuk mengolah data-data yang diberikan kepadanya. Algoritma *machine learning* lah yang berperan untuk mengajarkan pengolahan data kepada computer

(Christian, 2019).

Algoritma Klasifikasi

Naive Bayes adalah Algoritma sederhana dan lugas yang didasarkan pada teorema Bayes untuk probabilitas bersyarat. Hal ini untuk mengklasifikasi data tergantung pada frekuensi terjadinya deskriptor data dari set pelatihan. Asumsi algoritma *Naive Bayes* adalah semua data sama-sama independen. Dengan asumsi ini, algoritma menemukan ketergantungan antara fitur dari pelatihan (Kabir, Ashraf, & Ajwad, 2019).

Support Vektor Machine (SVM) merupakan metode klasifikasi jenis terpadu (supervised) karena ketika proses pelatihan perlu target pembelajaran tertentu (Nurachim, 2019).

Random Forest yaitu konsep untuk membuat sejumlah besar pohon keputusan berkolerasi dimana semua pohon keputusan bertindak sebagai model ansambel. Setiap pohon keputusan meletakkan prediksi kelas dan keputusan akhir didasarkan pada hasil maksimum (Kabir et al., 2019).

Beberapa penelitian terkait yang dilakukan dalam penerapan data mining menggunakan *multilayer perceptron* yang pernah dilakukan menggunakan dua modul (Sakar et al., 2018). Ada juga Perbandingan Model Algoritma Machine Learning dengan Menggunakan WEKA

dan *Sci-Kit Learn* Dalam Mengklasifikasikan Niat Pembelanja Online (Christian, 2019). Service Convenience and Post-Purchase Behaviour of Online Buyers : An Empirical Study (Khan & Khan, 2018). Factors Influencing Online Shoppers' Repurchase Intentions: The Roles of Satisfaction and Regret (Liao et al., 2016). Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data (Kabir et al., 2019). The Effect of Product Review Balance and Volume on Online Shoppers' Risk Perception and Purchase Intention (Yang, Sarathy, & Lee, 2016). Real-Time Prediction of Online Shoppers' Purchasing Intention Using *Random Forest* (Baati & Mohsil, 2020).

II. METODE PENELITIAN

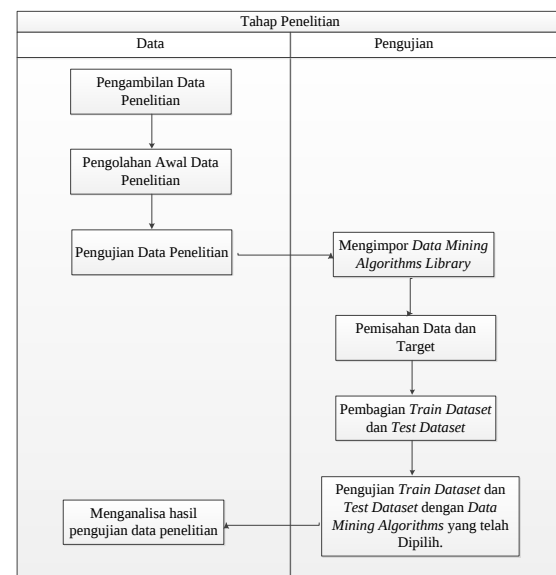
2.1 Tahap Penelitian.

Pada penelitian ini diawali dengan pengambilan data dari website *UCI Machine Learning Repository*. Selanjutnya data diolah menjadi data awal yang akan di proses menggunakan *tools WEKA* 3.8.2. Lalu diuji dengan algoritma dengan nilai objek yang sama.

Terdapat tiga buah algoritma yang akan dibandingkan menggunakan *tools* tersebut yakni algoritma *Naive Bayes*,

Support Vector Machine dan *Random Forest*. Setelah data diproses dan diuji, selanjutnya yaitu menganalisa data yang dihasilkan dari *tools WEKA* dengan tiga buah algoritma *machine learning*. Sehingga menghasilkan algoritma yang paling baik untuk diterapkan pada data tersebut.

Tahapan Penelitian yang dilakukan dijelaskan pada Gambar 1 berikut dibawah ini (Christian, 2019) :



Gambar 1. Tahapan Penelitian

2.2. Data Yang Digunakan

Data yang dipakai dalam penelitian ini yaitu dataset “*Online Shoppers Purchasing Intention Dataset*” yang diambil dari halaman website *UCI Machine Learning Repository*. C. Okan Sakar, Mete Alpaslan Katircioglu, S. Olcay Polat dan Yomi Kastro pada tahun 2018 menggunakan penelitian dataset ini.

nilai objek menggunakan algoritma *Naive Bayes*.

Sesuai dengan algoritma yang akan digunakan. Melalui aplikasi WEKA dapat dihasilkan data pengujian pada Gambar 4.

[illegible]

The screenshot shows the 'About' tab of the 'weka.gui.GenericObjectEditor' dialog for the 'weka.classifiers.bayes.NaiveBayes' class. The 'About' section contains a text box with the description 'Class for a Naive Bayes classifier using estimator classes.' and two buttons: 'More...' and 'Capabilities...'. Below this, several configuration options are listed with their current values: 'batchSize' is 100, 'debug' is False, 'displayModelInOldFormat' is False, 'doNotCheckCapabilities' is False, 'numDecimalPlaces' is 2, 'useKernalEstimator' is False, and 'useSupervisedDiscretization' is False. At the bottom, there are four buttons: 'Open...', 'Save...', 'OK', and 'Cancel'.

Gambar 3. Penentuan Nilai Objek algoritma *Naive Bayes* pada WEKA

The screenshot shows the Intel® Performance Analyzer (ITA) interface. The 'Statistics' tab is active, displaying the 'Cache Hit Ratio' for the 'Cache' component. The main pane shows a value of 99.99% for the 'Cache' component. Below this, a table lists the 'Cache Hit Ratio' for various components, with the 'Cache' component having the highest ratio of 99.99%.

Component	Cache Hit Ratio
Cache	99.99%
Cache (L1)	99.99%
Cache (L2)	99.99%
Cache (L3)	99.99%
Cache (L4)	99.99%
Cache (L5)	99.99%
Cache (L6)	99.99%
Cache (L7)	99.99%
Cache (L8)	99.99%
Cache (L9)	99.99%
Cache (L10)	99.99%
Cache (L11)	99.99%
Cache (L12)	99.99%
Cache (L13)	99.99%
Cache (L14)	99.99%
Cache (L15)	99.99%
Cache (L16)	99.99%
Cache (L17)	99.99%
Cache (L18)	99.99%
Cache (L19)	99.99%
Cache (L20)	99.99%

Gambar 4. Hasil Data Menggunakan Algoritma *Naive Bayes* pada WEKA

Pada pengujian data set *online shoppers purchasing intention* yang diatas dengan menggunakan metode *Naive Bayes* menghasilkan tingkat keberhasilan 80,9% dan *error* 19,1%.

Pengujian menggunakan aplikasi WEKA dilakukan dengan melihat dan memperhatikan nilai-nilai objek pada masing-masing algoritma yang digunakan, dalam penelitian ini terdapat tiga algoritma yang digunakan yaitu *Naive Bayes*, *Random Forest* dan *Support Vector Machine* (SVM). Gambar 3 merupakan

Pengujian berikutnya yaitu menggunakan metode *Random Forest* yang dilakukan dengan melihat dan memperhatikan nilai-nilai objek pada Gambar 5.



Gambar 5. Penentuan Nilai Objek algoritma *Random Forest* pada WEKA

Dengan algoritma yang digunakan yaitu *Random Forest* melalui pengujian di aplikasi WEKA, maka dapat dihasilkan pada Gambar 6.

Sehingga dari pengujian tersebut dengan menggunakan metode *Random Forest* menghasilkan tingkat keberhasilan 89,9% dan *error* 10,1% dapat dilihat pada Gambar 7.

Gambar 6. Hasil Data Menggunakan Algoritma *Random Forest* pada WEKA

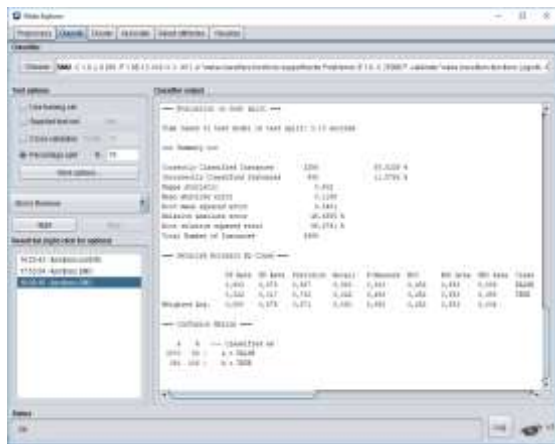
Selanjutnya pengujian menggunakan algoritma *Support Vector Machine* (SVM), nilai-nilai objek dapat dilihat Gambar 7.



Gambar 7. Penentuan Nilai Objek algoritma SVM pada WEKA

Algoritma SVM dapat menghasilkan nilai pengujian tingkat keberhasilan

88,02% dan *error*nya yaitu 11,97% dapat dilihat pada Gambar 8.



Gambar 6. Hasil Data Menggunakan Algoritma SVM pada WEKA

Dari data hasil yang diperoleh pengujian menggunakan aplikasi WEKA berisi nilai-nilai tertentu dan untuk membandingkan algoritma klasifikasi sebaik apa data yang akan dipilih dari hasil nilai *f1-score*, akurasi, *Kappa Statistic*, *mean absolute error* dari setiap algoritma yang diterapkan. Hasil ringkasan pengujian data tercantum dalam tabel 1 dibawah ini.

Tabel 1. Hasil Pengujian Data pada Aplikasi WEKA

Classifier	F1-Score	Akurasi	<i>Kappa Statistic</i>	MAE
<i>Naive Bayes</i>	0,825	80,9%	0,41	0,234
Random Forest	0,894	89,9%	0,581	0,144
SVM	0,858	88,02%	0,119	0,119

3. 2. Hasil F1-Score

F1- skor adalah ukuran yang paling banyak digunakan dalam evaluasi berbasis batas, memperkenalkan kerangka presisi-recall (Avola et al., 2019).

Dari hasil data yang diuji menggunakan aplikasi WEKA, algoritma *Random Forest* menghasilkan nilai 0,894. Untuk nilai *F1-Score* itu sendiri, nilai yang paling baik yaitu 1. Sehingga nilai 0,894 merupakan nilai yang paling mendekati 1 dari ke tiga algoritma yang lain. Jadi algoritma *Random Forest* merupakan algoritma yang paling baik pada nilai *F1-Score*. Sedangkan algoritma *Naive Bayes* merupakan nilai *F1-Score* yang terburuk karena mendekati nilai 0 dengan hasil 0,825.

3. 3. Hasil Akurasi

(Baati & Mohsil, 2020) menyimpulkan nilai yang menyajikan akurasi rata-rata yang tertinggi semakin baik algoritmanya.

Hasil dari data yang diuji menggunakan aplikasi WEKA, maka pengklasifikasian data menggunakan algoritma *Random Forest* menghasilkan nilai akurasi tertinggi yaitu dengan nilai 89,9%, dibandingkan nilai akurasi yang lain yang paling rendah adalah *Naive Bayes* dengan nilai 80,9%.

3. 4. Hasil *Kappa Statistic*

(Kavitha & Manikandan, 2019) menyimpulkan untuk menemukan derajat kesesuaian yaitu menggunakan *Kappa Statistic*, yang menghasilkan interval 0 dan 1. Maka semakin tinggi nilai *Kappa Statistic* semakin baik pula algoritma yang dipakai.

Dari pengujian data menggunakan aplikasi WEKA, hasil yang digunakan dalam pengklasifikasian data yang paling tinggi menggunakan algoritma *Random Forest*, dengan nilai *Kappa Statistic* 0,581. Sedangkan untuk nilai *Kappa Statistic* yang terendah dihasilkan dari algoritma *Naive Bayes* yakni 0,41.

3. 5. Hasil *Mean Absolute Error*

Mean Absolute Error adalah salah satu metode yang digunakan untuk mengukur tingkat keakuratan model peramalan. Nilai MAE menunjukkan rata-rata kesalahan (*error*) absolut antara hasil peramalan/prediksi dengan nilai riil (Suryanto & Muqtadir, 2019).

Dari data yang diuji menggunakan aplikasi WEKA, maka hasil mengklasifikasian data menggunakan algoritma *SVM*-lah yang merupakan nilai *mean absolut error* nya paling rendah dengan nilai 0,11 dan untuk nilai *mean absolut error* tertinggi algoritma *Naive Bayes* 0,234.

IV. SIMPULAN

Dari hasil pengujian pada penjelasan sebelumnya terdapat hasil yang berbeda pada *mean absolut error*.

Untuk nilai *F1-Score*, Akurasi dan *Kappa Statistic*, menghasilkan posisi kedudukan nilai yang sama, dan sesuai dengan kriteria dari masing-masing algoritma. Pada ketiga penilaian tersebut, algoritma *Random Forest*-lah yang merupakan algoritma paling baik dibandingkan algoritma *SVM* dan *Naive Bayes*.

Maka dari itu dapat diambil kesimpulan bahwa algoritma *Naive Bayes* dan *Support Vector Machine* merupakan algoritma yang tidak direkomendasikan untuk penilaian data *Online Shoppers Purchasing Intention Dataset*.

Sehingga dalam penelitian kali ini, algoritma *Random Forest* yang paling baik untuk digunakan dalam pengklasifikasian data *Online Shoppers Purchasing Intention Dataset*. Karena algoritma *Random Forest* yang paling mendominasi dalam mengetahui nilai kriteria seperti *F1-Score*, Akurasi, *Kappa Statistic* dan *Mean Absolute Error*.

DAFTAR PUSTAKA

- [1] Avola, D., Cinque, L., Foresti, G. L., Lamacchia, F., Marini, M. R., Perini,

- L., ... Telesca, G. (2019). A Shape Comparison Reinforcement Method Based on Feature Extractors and F1-Score, 2155–2159.
- [2] Baati, K., & Mohsil, M. (2020). *Real-Time Prediction of Online Shoppers* ' (Vol. 1). Springer International Publishing.
<https://doi.org/10.1007/978-3-030-49161-1>
- [3] Christian, Y. (2019). JITE (Journal of Informatics and Telecommunication Engineering), 3(1), 58–66.
- [4] Kabir, R., Ashraf, F. Bin, & Ajwad, R. (2019). Analysis of Different Predicting Model for *Online Shoppers* ' Purchase Intention from Empirical Data, (March 2020).
<https://doi.org/10.1109/ICCIT48885.2019.9038521>
- [5] Kavitha, S., & Manikandan, S. (2019). Improving the Automobile Purchasing Behavior of Customer : Classification Techniques, (2), 2219–2223.
<https://doi.org/10.35940/ijeat.B2924.129219>
- [6] Khan, M. A., & Khan, S. (2018). Service Convenience and Post-Purchase Behaviour of *Online Buyers* : An Empirical Study, 167–188. <https://doi.org/10.1007/s12927-018-0006-x>
- [7] Liao, C., Lin, H., Luo, M. M., & Chea, S. (2016). Factors influencing *Online Shoppers* ' Repurchase Intentions: The Roles of Satisfaction and Regret. *Information & Management*.
<https://doi.org/10.1016/j.im.2016.12.005>
- [8] Nurachim, R. I. (2019). Pemilihan Model Prediksi Indeks Harga Saham Yang Dikembangkan Berdasarkan Algoritma *Support Vector Machine* (Svm) Atau Multilayer Perceptron (Mlp) Studi Kasus : Saham Pt Telekomunikasi Indonesia Tbk Jurnal Teknologi Informatika & Komputer | Vol . 5 , No . 1 , Maret 2019, 5(1), 29–35.
- [9] Sakar, C. O., S. Olcay Polat, Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of *online shoppers* ' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 0.
<https://doi.org/10.1007/s00521-018-3523-0>
- [10] Suryanto, A. A., & Muqtadir, A. (2019). Penerapan Metode Mean Absolute Error (Mea) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi, (1), 78–83.
- [11] Yang, J., Sarathy, R., & Lee, J. (2016). AC Management Science and

- Information Systems. *Decision Support Systems*.
<https://doi.org/10.1016/j.dss.2016.06.009>
- [12] Fitria, Y. A. (2019). Visualization of data on earthquake prone areas from the analysis of earthquake data vibrations. *Test Engineering and Management*, 81(11-12).
- [13] Yulmaini, Y., Sanusi, A., & Yusendra, M. A. E. (2018). The Implementation of AHP for Determining Dominant Criteria in Higher Education Competitiveness Development Strategy Based on Information Technology. *International Journal of Artificial Intelligence Research*, 3(1).
- [14] Yuliawati, D., Saleh, S., & Indera, I. (2018). Prototype Pengadaan Dan Distribusi Barang Pada Waralaba Fried Chicken dan Burger lampung. *SIMADA (Jurnal Sistem Informasi dan Manajemen Basis Data)*, 1(1), 61-70.
- [15] Azhari, A., & Lestari, S. (2019, December). Forecasting The Number Of Customers Using The Box-Jenkins (Arima) And Linear Regression method (Case Study: PT. AIA Financial LPG Sunrise Agency). In *Prosiding International conference on Information Technology and Business (ICITB)* (pp. 206-217).
- [16] Purbo, O. W., Sriyanto, S., Irianto, S. Y., Aziz, R. Z., & Herwanto, R. (2020). Benchmark and comparison between hyperledger and MySQL. *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, 18(2), 705-715.
- [17] Jeprianto, J., & Aziz, R. A. (2020). Implementasi Algoritma Rough Set Dan Naive Bayes Untuk Mendapatkan Rule Dalam Menyeleksi Pemohon Bantuan Fasilitas Rumah Ibadah (Studi Kasus: Pemerintah Kabupaten Pringsewu). *JTKSI (Jurnal Teknologi Komputer dan Sistem Informasi)*, 3(2), 74-83.
- [18] Fitria, S. Y., & SeptiliaArfida, H. Green Cloud Computing Ideas with Security Issues in Setting of Distributed Computing and Cloud Framework.