

COMPARATIVE STUDY OF SPAM EMAIL CLASSIFICATION DECISION TREE BETWEEN USING CART AND J48

Ihsan Fauzi, Hilmy Sukma Nurmakarim, Ramadhan Islami Pasha, Chaerur Rozikin

Informatika, Universitas Singaperbangsa Karawang

Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat

Ihsanfauzi200.if47@gmail.com

ABSTRAK

Email merupakan salah satu bentuk komunikasi digital yang paling populer dan efisien di era modern. Namun, seiring dengan meningkatnya penggunaan email. Email spam merupakan pesan yang tidak diinginkan yang dikirimkan secara massal ke sejumlah besar penerima dengan tujuan mengiklankan produk atau layanan, menyebarkan *malware*. Spam mengganggu pengguna dan mengonsumsi sumber daya jaringan dan server. Solusi yang ditawarkan dalam mengatasi spam menggunakan *decision tree*. Decision tree merupakan teknik pembelajaran mesin klasifikasi data berdasarkan fitur-fitur tertentu. Dalam konteks penyaringan email spam, decision tree dapat digunakan untuk mengidentifikasi dan memisahkan email spam dari email yang valid berdasarkan analisis pola dan karakteristik isi email. Metode decision tree bekerja dengan membangun model dari data latih yang berisi contoh-contoh email yang sudah diklasifikasikan sebagai spam atau bukan spam. Model ini kemudian digunakan untuk mengklasifikasikan email baru dengan mengikuti alur keputusan yang ditentukan oleh pohon keputusan. Keunggulan utama dari metode ini adalah kemampuannya untuk menangani data yang kompleks dan memberikan hasil yang akurat dengan interpretasi yang mudah dipahami. Hasil penelitian menunjukkan bahwa metode decision tree J48 memiliki akurasi yang lebih baik yaitu 98,5% sedangkan metode CART memiliki akurasi sebesar 89%. Meskipun demikian kedua metode mampu menghasilkan prediksi yang akurat dan sesuai Ketika dimasukkan data baru kedalam modelnya.

Kata kunci : email, spam, CART, J48, Decision Tree

1. PENDAHULUAN

Pada era digital saat ini, email merupakan alat komunikasi yang penting baik dalam kehidupan pribadi maupun profesional. Penggunaannya yang luas serta fleksibilitasnya membuat email menjadi pilihan utama untuk berkomunikasi, berbagi informasi, dan menjalankan berbagai aktivitas bisnis. Namun, dengan semakin banyaknya pengguna email, muncul pula tantangan besar berupa email spam. Email spam adalah pesan yang tidak diinginkan yang sering kali dikirim secara massal dengan tujuan komersial, penyebaran malware, atau bahkan penipuan. Fenomena ini tidak hanya mengganggu penerima, tetapi juga mengonsumsi sumber daya jaringan dan membebani server email. Masalah spam menjadi semakin kompleks seiring dengan perkembangan teknologi dan meningkatnya volume email yang dikirim setiap hari [2][3].

Upaya untuk menyaring dan mengurangi spam memerlukan metode yang efektif dan efisien. Salah satu pendekatan yang banyak digunakan adalah melalui teknik Jurnal Machine Learning (ML) pembelajaran mesin, khususnya algoritma Decision tree. Algoritma ini menawarkan cara yang sistematis dan akurat untuk mengklasifikasikan bahwa email merupakan spam atau bukan spam berdasarkan fitur-fitur tertentu yang terdapat dalam email tersebut. Decision tree bekerja dengan membangun model prediktif dari data latih yang mengandung contoh-contoh email yang sudah diklasifikasikan. Model ini menggunakan serangkaian aturan keputusan yang disusun dalam bentuk pohon, dimana setiap cabang

merepresentasikan pilihan berdasarkan fitur email, dan setiap daun merepresentasikan klasifikasi akhir (spam atau non-spam[7][11].

Penelitian ini akan berfokus pada penerapan algoritma decision tree dalam penyaringan email spam[10].

Bertujuan untuk membandingkan dua varian utama dari algoritma ini, yaitu CART (Classification and Regression Tree) dan J48. Kedua algoritma ini memiliki pendekatan yang sedikit berbeda dalam membangun pohon keputusan, dan tujuan dari penelitian ini adalah untuk menentukan mana yang memberikan tingkat akurasi lebih tinggi dalam mendeteksi spam. Dengan mengidentifikasi algoritma yang lebih efektif, penelitian ini diharapkan dapat memberikan kontribusi nyata dalam upaya memerangi spam. Hasil yang diperoleh dapat digunakan untuk meningkatkan sistem penyaringan spam yang ada, sehingga pengalaman pengguna email menjadi lebih baik dan sumber daya sistem dapat digunakan dengan lebih efisien.

2. TINJAUAN PUSTAKA

Pada bagian ini akan dijelaskan landasan teori teori yang digunakan dalam penelitian ini.

2.1. Email Spam

Email spam merupakan pesan email yang tidak diminta oleh pemilik dari email. Email spam terdiri dari beberapa tipe seperti iklan, Nigerian spam, dan phishing. Spam email dikirimkan biasanya dikirimkan dalam jumlah yang banyak hal tersebut dapat

menimbulkan masalah seperti *space* pada *mail box* dan sulit untuk di *maintain* [8].

2.2. Decision Tree

Decision tree merupakan sebuah algoritma yang bekerja dengan mengubah kriteria menjadi *node* yang terhubung menjadi struktur pohon. Metode ini merupakan pemodelan prediktif yang dapat digunakan untuk klasifikasi dan prediksi tugas. Struktur daripada *decision tree* terdiri dari *node* dan *edge*. *Node* dibagi menjadi akar, cabang dan daun[9].

2.3. J48

J48 adalah salah satu jenis dari metode *decision tree* yang merupakan algoritma klasifikasi. Metode J48 bekerja dengan mengambil analisis keputusan berdasarkan perbandingan dari probabilitas dari *leaf* dan *node* yang berbeda kemudian dilakukan evaluasi untuk diambil sebuah keputusan[1].

2.4. Cart

CART merupakan singkatan/ *classification and regression tree* pada metode ini membangun pohon klasifikasi atau pohon regresi. Pada pohon klasifikasi digunakan untuk klasifikasi secara biner. Pohon regresi dapat digunakan untuk *forecasting* variabel yang dependen[6].

2.5. Penelitian Terdahulu

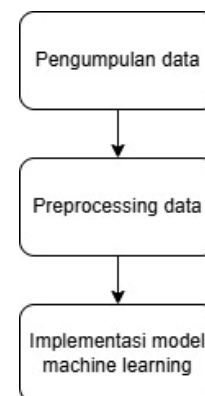
Menurut penelitian sebelumnya yang dilakukan oleh Putta Charan et al. (2022) dalam penelitian yang dilakukan dengan judul, "Higher Accuracy of Spam Mail Prediction using Decision Tree Algorithm Comparing with Random Forest Algorithm", membahas tentang meningkatnya volume email spam yang menyumbang lebih dari 77% dari total lalu lintas email global. Mereka menyoroti perlunya metode deteksi yang lebih canggih dan efisien untuk mengatasi masalah ini. Masalah utama yang dihadapi adalah akurasi yang rendah dari metode deteksi spam yang ada, yang sering kali menghasilkan false positives dan false negatives. Penelitian tersebut mengusulkan penggunaan algoritma Machine Learning yang lebih canggih, khususnya Decision Tree dan Random Forest, untuk meningkatkan akurasi prediksi spam. Mereka mengembangkan dan mengevaluasi kedua algoritma tersebut berdasarkan kemampuan mereka dalam mengklasifikasikan email secara efektif dan efisien [4].

Penelitian yang dilakukan oleh Sulaeman et al. (2022) dalam penelitian yang dilakukan dengan judul, "Perbandingan Kinerja Algoritma Naïve Bayes Dan C.45 Dalam Klasifikasi Spam Email", membahas tentang perbandingan antara algoritma Naïve Bayes dan C.45 dalam klasifikasi spam email. Mereka menyoroti pentingnya pemilihan algoritma yang tepat dalam memfilter email yang masuk ke kotak masuk pengguna. Hasil penelitian menunjukkan perbedaan kinerja yang signifikan antara Naïve Bayes dan C.45 dalam hal akurasi[2].

Penelitian yang dilakukan oleh Rizka Safitri Lutfiyani et al. (2021) dalam penelitian yang dilakukan dengan judul, "Implementasi Pendeteksian Spam Email menggunakan Metode Text Mining dengan Algoritma Naive Bayes dan Decision Tree J48", membahas tentang solusi dalam menyortir email spam dan bukan spam menggunakan text mining dan algoritma Naïve Bayes serta Decision Tree J48. Penelitian ini memperkuat urgensi untuk mengembangkan solusi yang efektif dalam menangani masalah spam email. Dengan mengintegrasikan temuan-temuan dari ketiga penelitian tersebut, penelitian ini bertujuan untuk memberikan wawasan baru dalam memahami dan membandingkan kinerja algoritma decision tree (CART, J48, dan Random Forest) dalam konteks penyaringan email spam. Dengan demikian, melalui pendekatan inovatif dan evaluasi yang cermat terhadap berbagai algoritma, diharapkan penelitian ini dapat memberikan kontribusi Jurnal Machine Learning (ML) yang signifikan dalam peningkatan efektivitas dan efisiensi deteksi spam email[5].

3. METODE PENELITIAN

Pada penelitian ini menggunakan tahapannya dibagi menjadi 3 tahapan. Tahapan tahapan penelitian yang dilakukan seperti pengumpulan data, preprocessing data, dan implementasi model machine learning.



Gambar 1. Flowchart langkah-langkah metode penelitian

3.1. Pengumpulan Data

Pada tahapan ini peneliti melakukan pengumpulan email spam dan bukan spam, Melakukan preprocessing data, implementasi algoritma decision tree CART dan J48 untuk penklasifikasian email. Data yang digunakan pada penelitian ini adalah email berkategori spam dan bukan spam. Proses pengumpulan data dilakukan secara manual yaitu dengan mengambil email sejumlah 500 dengan email berkategori spam sejumlah 250 dan email berkategori bukan spam berjumlah 250 yang dikumpulkan dari kotak masuk dan folder spam dari beberapa akun email. Setiap email yang dikumpulkan kemudian diberi label 'spam' untuk email yang berkategori spam dan 'not spam' untuk email yang berkategori bukan

spam. Pengumpulan data dilakukan secara manual untuk memastikan bahwa email yang dikumpulkan dapat merepresentasi dan mencakup semua kategori email. Dengan metode pengumpulan data ini dapat dipastikan bahwa email yang digunakan sama rata antara email yang berkategori spam dan email yang tidak berkategori spam untuk melatih dan menguji model sehingga prediksi yang dihasilkan dapat menjadi lebih akurat. Tabel 1. merupakan sampel dari data yang dikumpulkan:

Tabel 1. Dataset sampel email

text	spam
Subject : [Bangkit 2023 H2] Invitation to capstone project closing session	0
Subject : Invitation: [Bangkit 2023 H2] Capstone Project Closing Session @ Mon Jan 8, 2024 2pm - 2:45pm (WIB) (Ihsan Fauzi C314BSY3236)	0
Subject : Updated invitation: [Bangkit 2023 H2] Capstone Project Closing Session @ Mon Jan 8, 2024 2pm - 3pm (WIB) (Ihsan Fauzi C314BSY3236)	0
Subject : Bangkit Fourth Month Milestone Rewards	0
Subject : Updated invitation: Weekly Consultation 21 - CC-54 @ Tue Jan 9,	0
Subject : New Year, New Skills: \$10 Each or Bundle Up for More Savings!	1
Subject : Postman's biggest API conference ever is here—save the date!	1
Subject : Elevate Your Skills: Year-End Specials on Tech eBooks and Video Courses!	1
Subject : Unlock Exclusive Packt Content for Only \$10/Month!	1
Subject : NBA 2K24 Winter Holiday Sale! * Don't Miss Out!	1

3.2. Preprocessing Data

Setelah data yang dibutuhkan terkumpul tahapan selanjutnya dalam penelitian ini adalah dengan melakukan preprocessing data. Langkah Langkah preprocessing yang dilakukan adalah menkonversi teks menjadi Lowercase yang bertujuan untuk mengurangi variasi huruf, setelah itu dilakukan penghapusan tanda baca untuk mengurangi alfa numerik yang tidak relevan dan dapat mengganggu akurasi dari model. Langkah Langkah tersebut akan diterapkan kepada setiap email berkategori spam maupun bukan spam. Preprocessing ini dilakukan untuk memastikan bahwa data teks yang digunakan bersih dan juga konsisten sehingga memungkinkan model untuk menghasilkan prediksi yang lebih efektif dan akurat dalam melakukan klasifikasi *email* spam dan non-spam.

3.3. Implementasi Model Machine Learning

Dalam penelitian ini dua algoritma pohon Keputusan akan digunakan untuk membangun model pengklasifikasian email spam. Jenis algoritma yang digunakan adalah CART (Classification dan regression

tree) dan algoritma J48 (implementasi dari algoritma C4.5). CART algorithm membuat pohon keputusan dengan mempartisi data secara rekursif berdasarkan nilai fitur yang menghasilkan pembagian murni di setiap node, baik untuk classification atau regresi. Algoritma ini menggunakan metrik pengotor atau entropi Gini untuk memilih distribusi terbaik. selanjutnya, algoritma J48 membuat pohon keputusan, tetapi menggunakan entropy dan gain informasi untuk memilih pemisahan terbaik di satu node, as well as mekanisme pruning untuk mengurangi pohon kompleksitas dan meningkatkan generalisasi model.

Untuk setiap algoritma, data yang telah diproses dan dikonversi ke format ARFF digunakan untuk melatih model. Setelah pelatihan, model evaluasi menggunakan data pengguna untuk menilai akurasi dan efisiensi dalam mengklasifikasikan email sebagai spam atau bukan spam. Implementasi dan perbandingan kedua algoritma memberikan wawasan tentang kemanjuran masing-masing metode dalam mendeteksi spam email. Selanjutnya model akan dilatih menggunakan dataset yang sudah dibuat sebelumnya. Model akan dilatih untuk menemukan pattern dari email spam yang berlabel 1 dan not spam yang berlabel 0.

4. HASIL DAN PEMBAHASAN

Tujuan dari penelitian ini adalah untuk membandingkan kinerja dua algoritma pohon keputusan, yaitu CART (Klasifikasi dan Regresi Pohon) dan J48 (Implementasi C4.5 di Weka), dalam klasifikasi email spam. Data yang dikumpulkan secara manual untuk penelitian ini diberi label "spam" dan "tidak spam" berdasarkan isi dan karakteristik email.

Data dibagi menjadi dua kelompok: data latih (80 persen) dan data uji (20 persen). setelah praproses data yang mencakup penghapusan tanda baca dan normalisasi teks. Metode CountVectorizer mengubah teks menjadi representasi vektor numerik untuk mengekstraksi fitur dari teks.

Subject: Hi Team, Just a reminder about our meeting tomorrow.
Predicted label (J48): Not Spam

Gambar 4. Hasil prediksi terhadap data baru pada algoritma j48

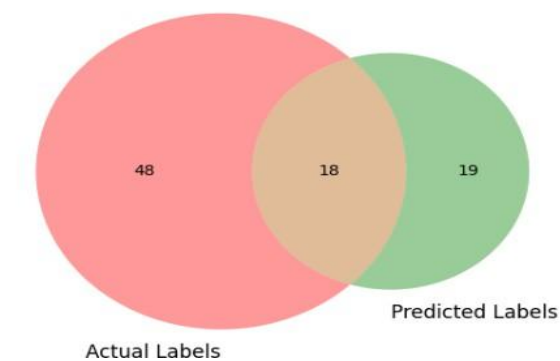
Gambar 4. diatas menunjukkan kinerja model Ketika sebuah data baru dimasukkan kedalam model J48. Dapat dilihat bahwa model mampu memprediksi secara tepat email baru yang dimasukkan kedalam model.

4.1. Metode J48

Gambar 2 adalah distribusi dari actuals label dan predicted labels di testing set pada model yang mengimplementasikan metode Decision Tree J48. Distribusi actuals dan predicted labels dalam tugas klasifikasi memberikan wawasan tentang performa dan perilaku model pembelajaran mesin. Dari diagram venn yang terdapat di dalam gambar actual labels

memiliki ratio 48, predicted labels memiliki ratio 19 dan 18 merupakan keduanya. Berdasarkan diagram diatas Model tepat tetapi tidak terlalu sensitif. Ini cenderung memprediksi lebih sedikit positif atau email spam, dan ketika itu memprediksi positif, biasanya benar. Namun, gagal mengidentifikasi banyak kasus positif yang sebenarnya, yang menunjukkan bahwa model tersebut mungkin terlalu konservatif dalam prediksinya. Ini bisa disebabkan oleh berbagai alasan seperti ambang keputusan yang tinggi, pelatihan yang tidak memadai tentang kasus positif, atau ketidakseimbangan kelas dalam data pelatihan.

Distribution of Actual and Predicted Labels in Testing Set



Gambar 2. Distribusi label aktual dan prediksi pada algoritma J48

Subject: Congratulations! You've won a free vacation.
 Predicted label (Decision Tree): Spam
 Subject: Hi Team, Just a reminder about our meeting tomorrow.
 Predicted label (Decision Tree): Not Spam

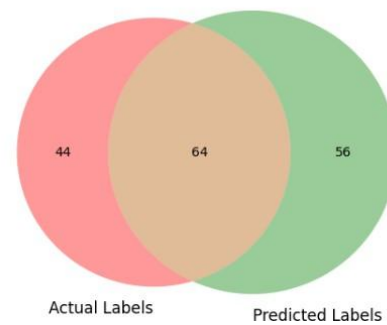
Gambar 3. Akurasi algoritma j48

Gambar 3. menunjukkan akurasi yang dihasilkan oleh metode J48. Hasil penelitian menunjukkan bahwa model J48 mendeteksi email spam dengan sangat baik dengan tingkat akurasi 98.5%. Prediksi dari beberapa email baru menunjukkan bahwa model dapat menggeneralisasi pola yang telah dipelajari dengan baik. Visualisasi hasil menawarkan informasi tambahan tentang distribusi label yang diprediksi serta kinerja model secara keseluruhan.

4.2. Metode CART

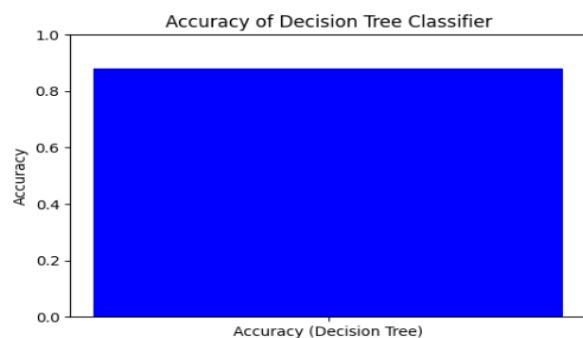
Gambar 5. adalah distribusi dari actuals label dan predicted labels di testing set pada model yang mengimplementasikan metode Decision Tree CART. Distribusi actuals dan predicted labels dalam tugas klasifikasi memberikan wawasan tentang performa dan perilaku model pembelajaran mesin. Label actual dan label prediksi digunakan untuk evaluasi model dengan membandingkan nilai actual dan prediksi untuk menilai performa dari model yang dihasilkan dalam bentuk metrik seperti akurasi, presisi, recall, dan F1-Score.

Distribution of Actual and Predicted Labels in Testing Set



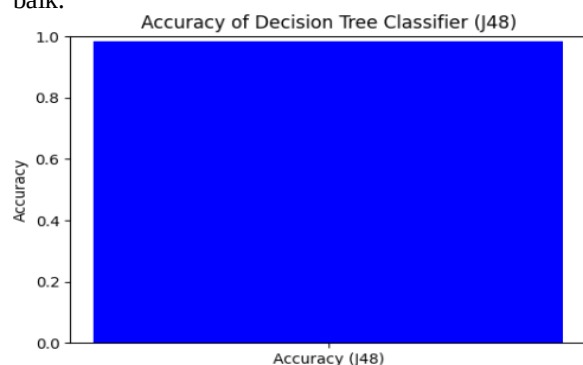
Gambar 5. Distribusi label aktual dan prediksi pada algoritma CART

Dari diagram venn yang terdapat di dalam gambar actual labels memiliki ratio 44, predicted labels memiliki ratio 56 dan 64 merupakan keduanya. Hal ini berarti model sangat efektif dalam mendeteksi kasus positif atau email spam yang ada.



Gambar 6. Akurasi algoritma CART

Gambar 6. menunjukkan akurasi yang dihasilkan oleh metode CART. Hasil penelitian menunjukkan bahwa model CART mendeteksi email spam dengan sangat baik dengan tingkat akurasi 89%. Dengan presentase tersebut menunjukkan bahwa model ini juga dapat melakukan pengklasifikasian email dengan baik.



Gambar 7. Hasil prediksi algoritma CART terhadap data baru

Gambar 7. menunjukkan kinerja model Ketika sebuah data baru dimasukkan kedalam model CART. Dapat dilihat bahwa model mampu memprediksi secara tepat email baru yang dimasukkan kedalam model.

5. KESIMPULAN DAN SARAN

Email spam merupakan permasalahan yang marak menimpa pengguna internet serta dapat menimbulkan resiko penipuan dan mengurangi sumber daya jaringan dan server. Dalam mengatasi masalah tersebut *decision tree* merupakan metode yang dapat digunakan sebagai klasifikasi email spam atau non-spam berdasarkan subjek dari email. Hasil penelitian ini menunjukkan bahwa metode *decision tree* J48 memiliki tingkat akurasi yang lebih tinggi sebesar 98,5% dibandingkan dengan metode *decision tree* CART sebesar 89% dalam memprediksi email spam dan non-spam. Kedua metode yang digunakan mampu memberikan prediksi yang akurat dalam menyaring email spam atau non-spam. Dalam penelitian selanjutnya dapat menggunakan metode lain dalam klasifikasi email spam atau non-spam. Dan dapat membandingkan metode CART serta J48 dalam kasus lainnya.

DAFTAR PUSTAKA

- [1] Lutfiyani, R. S., & Retnowati, N. (2021). IMPLEMENTASI PENDETEKSIAN SPAM EMAIL MENGGUNAKAN METODE TEXT MINING DENGAN ALGORITMA NAÏVE BAYES DAN DECISION TREE J48. *Jurnal Komputer Dan Informatika*, 9(2), 244–252. <https://doi.org/10.35508/jicon.v9i2.5304>
- [2] Sulaeman, N., Suarna, N. N., Ajiz, N. A., Bahtiar, N. A., & Fathurrohman, N. (2022). Perbandingan kinerja algoritma Naïve Bayes dan C.45 dalam klasifikasi spam email. *Kopertip*, 6(1), 8–14. <https://doi.org/10.32485/kopertip.v6i1.130>
- [3] Sharma T, Aarthy SL. An automatic attendance monitoring system using RFID and IOT using Cloud. In *Green Engineering and Technologies (IC-GET)*, 2016 Online International Conference on 2016 Nov 19 (pp. 1-4). IEEE.
- [4] Charan, Putta. (2022). Higher Accuracy of Spam Mail Prediction using Decision Tree Algorithm Comparing with Random Forest Algorithm. *Journal of Pharmaceutical Negative Results*, 13(4). <https://doi.org/10.47750/pnr.2022.13.s04.006>
- [5] Lutfiani. R.S. Retnowati. N. (2021). Implementasi Pendeteksian Spam Email menggunakan Metode Text Mining dengan Algoritma Naive Bayes dan Decision Tree J48 | <https://doi.org/10.35508/jicon.v9i2.5304>
- [6] Adhitya. R. R., Witanti. W, Yuniarti. R. (2023). PERBANDINGAN METODE CART DAN NAÏVE BAYES UNTUK KLASIFIKASI CUSTOMER CHURN, <https://doi.org/10.31949/infotech.v9i2.5641>
- [7] Ramadhan, N. G., Wibowo, M., Rosely, N. F. L. M., & Quix, C. (2022). Opinion mining indonesian presidential election on twitter data based on decision tree method. *Jurnal Infotel/Jurnal Infotel*, 14(4), 243–248. <https://doi.org/10.20895/infotel.v14i4.832>
- [8] Shrivastava, A. K., Dewangan, A. K., Ghosh, S. M., & Singh, D. (2021). Development of Proposed Ensemble Model for Spam e-mail Classification. *Informacinės Valdymas, Technologijos Ir* 50(3). <https://doi.org/10.5755/j01.itc.50.3.27349>
- [9] Nuryawan, A. T. O., Hasbullah, M., Rizal, M., Rajab, M. F., & Agustina, N. (2023). ALGORITMA DECISION TREE UNTUK ANALISIS SENTIMEN PUBLIC TERHADAP MARKETPLACE DI INDONESIA. *Naratif*, 5(1), 18–25. <https://doi.org/10.53580/naratif.v5i1.186>
- [10] Čavor, X. (2021). "Decision Tree Model for Email Classification." In *25th International Conference on Information Technology (IT)* (pp. 1-4). Zabljak, Montenegro. <https://doi.org/10.1109/IT51528.2021.9390143>
- [11] Arnomo, S. A. (2021). Analisa Decision Tree untuk Kepuasan Penggunaan Sinyal dari Base Transceiver Station (BTS). *Jurnal Sistem Dan Teknologi Informasi*, 9(2), 199. <https://doi.org/10.26418/justin.v9i2.43425>