PERBANDINGAN 3 ALGORITMA KLASIFIKASI DATA MINING DALAM PRO-KONTRA BAHAYA ROKOK ELEKTRIK

Eva Argarini Pratama[1)], Corie Mei Hellyana[2)], Nuzul Imam Fadlilah[3)]

[1]*Sistem Informasi, Universitas Bina Sarana Informatika*
[2] *Sistem Informasi Akuntansi, Universitas Bina Sarana Informatika*
[3]*Teknologi Komputer, Universitas Bina Sarana Informatika*
[1,2,3]*Jl. Kramat Raya No. 98 Kwitang, Senen, Jakarta Pusat*
Email: [1]eva.eap@bsi.ac.id, [2]corie.cma@bsi.ac.id, [3]nuzul.nfh@bsi.ac.id

Abstract

Many studies and facts exist in society that smoking is bad for health. This causes smokers to look for alternatives from traditional cigarettes to e-cigarettes (vaping) to help quit smoking slowly. However, recently the use of e-cigarettes has become its own pros and cons in society because of the acceptance of information from the community itself. This study tries to compare the use of 3 data mining classification methods, namely Decision Tree, Naïve Bayes and Logistic Regression to get the most accurate algorithm and find out the results of the classification of dangers or not using e-cigarettes based on the views of the community itself using 4 comparison factors, namely how often. see how people use/offer/advertise/promote e-cigarettes, how often they see shows about the dangers of e-cigarettes, have close friends who smoke vaping, and have family (father/mother/siblings) who smoke. From the results of data processing and testing by measuring the performance of the three algorithms using the confusion matrix procedure, operator cross validation and the ROC curve, the decision tree algorithm produces the highest level of accuracy value of 81.00% and based on the decision tree algorithm graph it can also be seen that you have Close friends who smoke e-cigarettes are the main factors that people think vaping is dangerous.
Keyword: *confusion matrix*, *cross validation, e-cigarette, datamining, classification*

Abstrak

Banyak penelitian dan fakta yang ada di masyarakat bahwa merokok itu buruk bagi kesehatan. Hal ini menyebabkan perokok mencari alternatif dari rokok tradisional hingga rokok elektrik (vaping) untuk membantu berhenti merokok secara perlahan. Namun akhir-akhir ini penggunaan rokok elektrik menjadi pro dan kontra tersendiri di masyarakat karena faktor penerimaan informasi dari masyarakat itu sendiri. Penelitian ini mencoba membandingkan penggunaan 3 metode klasifikasi data *mining* yaitu *Decision Tree*, *Naïve Bayes* dan *Logistic Regression* untuk mendapatkan algoritma yang paling akurat dan mengetahui hasil klasifikasi bahaya atau tidaknya penggunaan rokok elektrik berdasarkan pandangan masyarakat itu sendiri menggunakan 4 faktor pembanding yaitu seberapa sering. melihat bagaimana orang menggunakan/ menawarkan/ mengiklankan/ mempromosikan rokok elektrik, seberapa sering mereka melihat tayangan tentang bahaya rokok elektrik, memiliki teman dekat yang merokok vape, dan memiliki keluarga (ayah/ibu / saudara kandung) yang merokok. Dari hasil pengolahan data dan pengujian dengan melakukan pengukuran performansi ketiga algoritma tersebut menggunakan prosedur *confusion matrix*, *operator cross validation* dan kurva ROC, algoritma *decision tree* menghasilkan tingkat nilai akurasi yang tertinggi sebesar 81,00% dan berdasarkan grafik algoritma *decision tree* juga terlihat bahwa Anda memiliki teman dekat yang merokok de ngan rokok elektrik menjadi faktor utama seseorang menganggap vaping berbahaya.
Kata Kunci: *confusion matrix*, *cross validation*, rokok elektrik, *datamining*, klasifikasi

1. Pendahuluan

Sekitar 1,1 miliar orang di dunia berusia 15 tahun ke atas memiliki kebiasaan merokok, dengan 80% perokok adalah orang yang tinggal di negara berkembang atau negara dengan pendapatan rendah dan menengah. Pertumbuhan konsumsi tembakau di negara berkembang menjadi beban negara di bidang kesehatan, ekonomi dan sosial dan ini merupakan masalah yang serius dan ada kecenderungan meningkat [1]. Pada dasarnya, efek berbahaya dari merokok tembakau sudah diketahui sebelumnya. Penyebab kanker paru-paru adalah 90% karena merokok, karena perokok memiliki kemungkinan empat kali lebih tinggi untuk mengembangkan penyakit ini daripada non-perokok [2].

Dalam beberapa tahun terakhir, rokok elektronik (EC) telah banyak diiklankan sebagai perangkat merokok alternatif yang diklaim dapat membantu perokok berhenti merokok [3]. Rokok elektrik pertama kali diperkenalkan ke pasar pada tahun 2003, mereka dipromosikan sebagai cara untuk mengurangi konsumsi rokok. Jika seseorang mengidap penyakit paru-paru, tentunya kebiasaan berhenti merokok merupakan salah satu upaya terpenting

untuk menjaga kesehatan. Namun, bukti yang mendukung penggunaan rokok elektrik sebagai strategi efektif untuk membantu perokok berhenti merokok masih kurang. Selain itu, dampak penggantian rokok tradisional atau temabakau dengan rokok elektrik masih belum jelas [4].

Faktanya, rokok elektrik tidak membakar tembakau, sehingga dianggap memiliki risiko/ dampak buruk yang jauh lebih rendah daripada rokok tradisional. Selain hanya menghasilkan uap yang secara visual menyerupai asap, untuk alasan yang sama, efek uap inilah yang diduga dapat digunakan sebagai pengganti rokok tembakau. Selain itu, beberapa survei di internet dan uji klinis menunjukkan bahwa rokok elektrik dapat membantu perokok berhenti merokok atau mengurangi bahaya dengan merokok lebih sedikit, tanpa risiko bahaya yang luar biasa bagi pengguna atau pengguna pasif. Efek kesehatan dari pemakaian rokok elektrik dirasa lebih sedikit/ terbatas. Sementara rokok elektrik aerosol mungkin mengandung lebih sedikit racun daripada rokok tradisional biasa, penelitian yang meneliti apakah rokok elektrik lebih tidak berbahaya daripada rokok tradisional tidak meyakinkan. Beberapa bukti menunjukkan bahwa penggunaan rokok elektrik dapat memfasilitasi penghentian merokok, tetapi data definitif masih kurang [5]. Inilah sebabnya mengapa pemakaian rokok elektrik telah memicu perdebatan diantara para profesional kesehatan yang memiliki tujuan yang sama yaitu mengurangi penyakit dan kematian terkait tembakau. Tetapi mereka tidak setuju tentang rokok elektrik yang dianggap membuat masalah penyakit dan kematian menjadi lebih menurun atau lebih meningkat [2].

Dengan permasalahan tersebut, menggunakan pengetahuan datamining yang ada. Penelitian ini mencoba membandingkan penggunaan algoritma yang ada untuk mengklasifikasikan opini tentang berbahaya atau tidaknya penggunaan rokok elektrik bagi masyarakat umum saat ini. Algoritma yang dibandingkan adalah *Decision Tree*, *Naïve Bayes* dan *Logistic Regression*. Dimana pada penelitian sebelumnya algoritma *Decision Tree* merupakan salah satu algoritma yang paling akurat dalam mengklasifikasikan beberapa masalah seperti memprediksi status pembayaran non finansial terburuk pada kredit komersial dimana *Decision Tree*, *Naïve Bayes* dan *Logistic Regression* untuk mendapatkan algoritma yang paling akurat dan mengetahui hasil klasifikasi bahaya atau tidaknya penggunaan rokok elektrik berdasarkan pandangan masyarakat itu sendiri menggunakan 4 faktor pembanding yaitu seberapa sering. melihat bagaimana orang menggunakan/ menawarkan/ mengiklankan/ mempromosikan rokok elektrik, seberapa sering mereka melihat tayangan tentang bahaya rokok elektrik, memiliki teman dekat yang usan memiliki akurasi yang lebih baik daripada *Logistic Regression* [6]. Pada kasus perbandingan *Naïve Bayes* dan *Decision Tree* pada seleksi fitur menggunakan algoritma genetika *untuk* masalah klasifikasi, juga menunjukkan bahwa Decision *Tree* memiliki akurasi yang sedikit lebih baik daripada *Naïve Bayes*. Namun pada penelitian yang membandingkan penggunaan *Naïve Bayes*, *Decision Tree* dan *k-Nearest Neighbor* dalam menemukan desain alternatif pada alat simulasi energi, menunjukkan bahwa pada penelitian ini, *Naïve Bayes* mengungguli *Decision Tree* dan *kNearest Neighbor* [7].

Perbandingan penggunaan algoritma *naïve Bayes* dan *Logistic Regression* pada Teori, Implementasi, dan Validasi Eksperimental dengan hasil yang menunjukkan bahwa *Logistic Regression* dengan teknik *gradient climbing* dapat mengungguli *classifier* dari *Naïve Bayes* umum. Namun, dengan asumsi pengklasifikasi Gaussian *Naive Bayes*, baik pengklasifikasi *Naïve Bayes* maupun *Logistic Regression* memiliki kinerja yang sama [8]. Dengan adanya perbedaan hasil perbandingan tersebut, penelitian ini juga perlu membandingkan algoritma klasifikasi mana yang paling baik diterapkan dalam meningkatkan akurasi pengklasifikasian opini tentang berbahaya atau tidaknya penggunaan rokok elektrik.

2. Metode Penelitian

Metode penelitian yang dilakukan pada penelitian ini yaitu penelitian eksperimen, dan memiliki beberapa tahapan yaitu:

2.1. Metode Pengumpulan Data

Langkah awal adalah pengumpulan data kuisioner terkait rokok elektrik. Dataset berfokus pada opini yang membahas hal-hal terkait aspek pendukung utama dalam pengenalan rokok elektrik hingga penggunaan rokok elektrik. Tahap selanjutnya adalah proses pemberian label pada data melalui penyematan status tweet yang dilihat dari sentimen negatif dan sentimen prositif. Melalui cara manual (memberikan informasi nilai untuk setiap kuesioner yang diisi) proses pelabelan dilakukan.

Pada tahap awal, data mentah dikumpulkan dengan menggunakan pertanyaan dan pernyataan dalam kuesioner sebagai variabel yang diolah, yaitu Apakah Anda pernah melihat kesan orang menggunakan/ menawarkan/ mengiklankan/ mempromosikan rokok elektrik/ vape? Pernahkah Anda melihat tayangan tentang bahaya rokok elektrik? Saya punya teman dekat perokok elektrik, saya punya keluarga (ayah/ ibu/ kakak) perokok elektrik. Pertanyaan dan pernyataan dalam kuesioner diambil sebagai data dari beberapa responden yang tersebar di beberapa kota besar di Indonesia seperti Jakarta, Medan, Semarang, Surabaya, dan lain-lain. Data yang terkumpul disimpan dalam format file excel. Dataset yang telah dikumpulkan dari kuesioner ini adalah data *unsupervised*. Untuk diolah menggunakan teknik *supervised learning*, data angket yang telah dikumpulkan sebelumnya perlu dibuatkan atau diberi label, proses pelabelan dilakukan secara manual dengan memberikan status pada setiap pertanyaan. Sukai label A untuk pertanyaannya Pernahkah Anda melihat kesan orang menggunakan/ menawarkan/ mengiklankan/ mempromosikan rokok elektrik/ vape? Dengan jawaban Ya dan Tidak, beri label B untuk

pertanyaan ?, Pernahkah Anda melihat kesan tentang bahaya rokok elektrik? Dengan tidak pernah, sangat jarang, jarang dan sering, label C untuk pernyataan saya memiliki teman dekat yang merokok vape dengan Ya dan Tidak, dan label D dengan pernyataan saya memiliki keluarga (ayah/ ibu/ saudara) yang merokok elektrik dengan Ya dan Tidak.

2.2. *Cross Validation*

Dalam memvalidasi atau menilai keakuratan dalam suatu model yang ada pada suatu dataset dapat menggunakan suatu teknik yang disebut dengan dengan cross validation. Validasi juga dilakukan sebagai tes standar untuk memprediksi tingkat kesalahan. Ada 3 langkah utama dalam validasi *cross* yaitu:

1. Cadangan sebagian dari kumpulan data sampel.
2. Menggunakan sisa data-set melatih model.
3. Uji model menggunakan bagian cadangan dari kumpulan data.

Pengolahan data awal dilakukan pada penggunaan RapidMiner, terdapat operator *cross validation* yang merupakan suatu operator yang ada pada rapidmaner dan memiliki 2 subproses yaitu yang pertama adalah subproses *training* dimana subproses ini dapat dipakai dalam mencoba atau melatih model yang sudah dibuat dan yang kedua subproses *testing* yaitu subproses yang digunakan sebagai pengujian dan pengukuran dari kinerja model tersebut, untuk memberikan nilai k (jumlah iterasi) maka pada operator *cross validation* terdapat parameter yang dapat digunakan yaitu *Number of fold*, dengan penggunaan operator ini terdapat s*ampling type* digunakan untuk memilih teknik *sampling* yang dapat membagi dataset [9]. Standar deviasi yang merupakan hasil perolehan dari pengukuran dalam penyebaran data yang mengambarkan atau menunjukan jarak rata-rata dari nilai tengah menuju suatu titik nilai tertentu akan dihasilkan berdasarkan kinerja model yang menggunakan cross validation dalam penentuan akurasinya. Hal ini dapat diartikan semakin besar standar deviasi yang dihasilkan, maka penyebaran dari nilai tengahnya juga akan semakin besar, dan begitu juga sebaliknya.

2.3. *Confusion Matrix*

Terdapat suatu konsep dalam datamining dimana penghitungan akurasi dapat menggunakan salah satu metode yaitu *confusion matrix*, dimana pada confusion matrix ini perhitungan dapat menghasilkan 4 keluaran yaitu akurasi, *recall*, *precision*, dan *error rate*. Hasil evaluasi model klasifikasi didasarkan pada pengujian untuk memperkirakan objek benar dan salah [10]. Pada tahap *confusion matrix* kinerja dapat diukur dengan TP, TN, FP dan FN, seperti gambaran dan rumusan berikut [11]:

Tabel 1. Komposisi Tabel Hasil *Confusion Matrix*

Kelas	Terklarifiaksi Positif	Terklarifikasi Negatif
Positif	True Positif (TP)	False Negative (FN)
Negatif	False Positif (FP)	True Negative (TN)

True Positive (TP) merupakan jumlah dari data bernilai benar dan memiliki nilai kebenaran datanya adalah benar. False Negative (FN) adalah jumlah dari data salah dan memiliki nilai kebenaran datanya adalah salah. False Negative (FN) adalah jumlah dari data benar dan dianggap oleh sistem memiliki nilai kebenaran datanya adalah salah. True Negative (TN) adalah jumlah dari data salah dan dianggap oleh sistem memiliki nilai kebenarannya benar.

Adapun rumusan dalam menghitung keakuratan dalam mengklasifikasikan data adalah:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+Fp+FN}} \quad (1)$$

2.4. ROC Curve

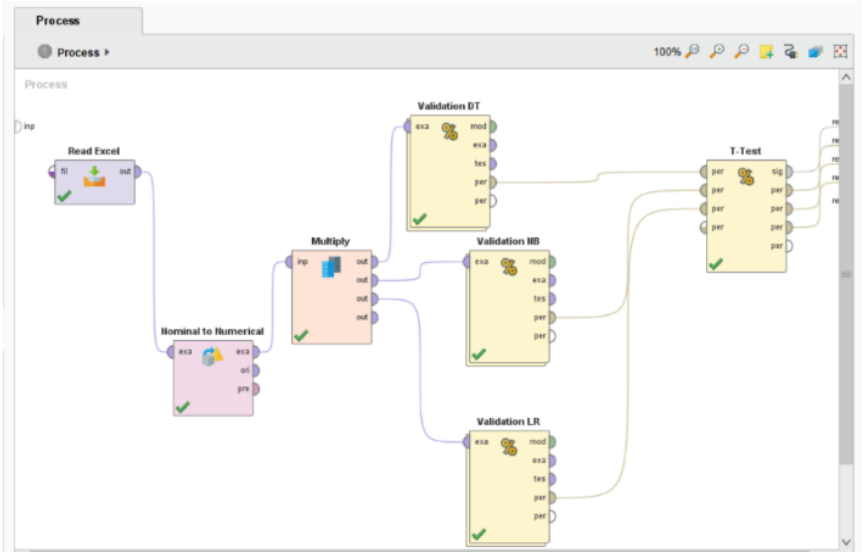
ROC Curve merupakan salah satu prosedur analisis model klasifikasi yang sudah ditentukan sebelumnya. Salah satu manfaat dari ROC curve adalah parameter model yang diinginkan berdasarkan sifat kategoris modelnya. Tingkat akurasi, skalabilitas, kecepatan dan interpretabilitas merupakan beberapa kriteria dari penggunaan metode [12].

2.5. Validasi

Kinerja dari suatu metode perlu diuji atau dievaluasi secara sistematis [13] dan evaluasi klasifikasi dilakukan berdasarkan pada pengujian diobjek benar dan salah sehingga menghasilkan suatu validasi data yang dapat digunakan untuk menentukan jenis skema pembelajaran atau hasil yang terbaik, berdasarkan data pelatihan untuk melatih rencana pembelajaran untuk memaksimalkan penggunaan data.

3. Hasil dan Pembahasan

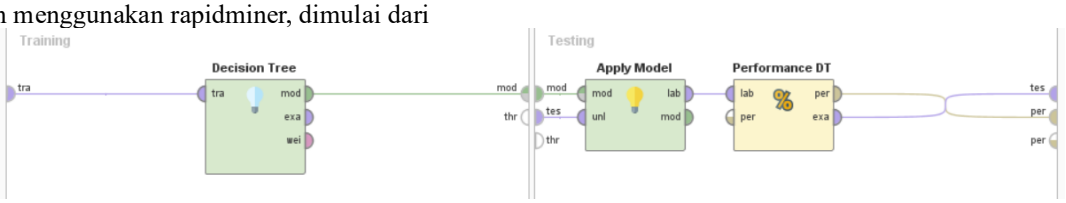
Dalam melakukan perbandingan 3 algoritma klasifikasi ini, terdapat beberapa tahapan, yang di awali dengan tahap modeling. Pada tahap ini, pemilihan teknik *datamining* dilakukan dengan menentukan algoritma yang akan digunakan dalam penelitian dengan menggunakan sebuah tools yaitu *Rapid Miner* versi 9.5. Adapun hasil pengujian model yang telah dilakukan dengan menggunakan tools tersebut adalah mengklasifikasikan rokok elektrik aman/tidak berbahaya bagi kesehatan dan rokok elektrik memang berbahaya bagi kesehatan menggunakan algoritma *Decision Tree*, *Naïve Bayes* dan *Logistic Regression* untuk mendapatkan hasil nilai akurasi tertinggi atau terbaik. Berikut ini adalah desain model *Rapidminer* yang digunakan:



Gambar 1. Perbandingan Desain Model Algoritma *Decision Tree*, *Naïve Bayes* dan *Logistic Regression*

Hasil akurasi yang diperoleh akan dipengaruhi oleh setting dan penggunaan operator maupun parameter yang dipakai pada *framework rapid miner*, selain itu model yang terbentuk pun juga akan terpengaruh oleh hal tersebut, sebagai gambaran adalah penggunaan bantuanmulti operator dalam melakukan perbandingan.

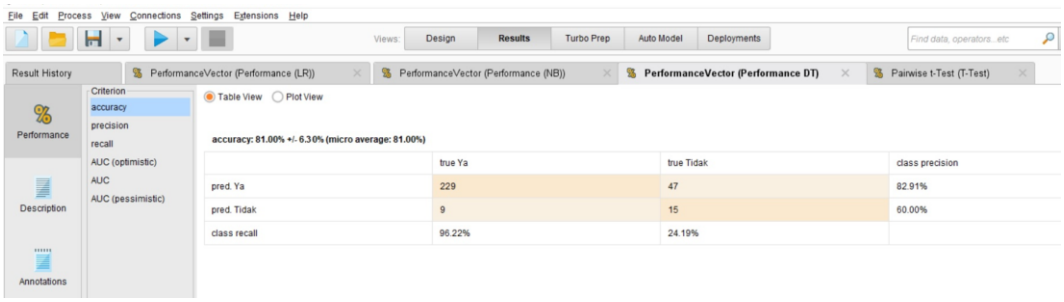
Gambar diatas juga menggambarkan model pengujian algoritma Decision Tree, Naïve Bayes dan Logistic Regression menggunakan rapidminer, dimulai dari memasukkan data kemudian setting role set yang kemudian menentukan label yang ada kemudian memproses dokumen yang berisi seperti pada Gambar 1. masing-masing algoritma adalah seperti contoh gambar dibawah untuk salah satu algoritma yaitu Decision Tree :



Gambar 2. Contoh Desain Proses *10-Fold Cross Validation* untuk salah satu algoritma yaitu *Decision Tree*

Gambar 2 menjelaskan desain proses pada operator *Decision Tree* validasi silang pada Gambar 1. Pengujian ini dilakukan dengan menggambil data dari operator *read* excel dan data ini merupakan data bersih yang sudah melalui *preprocessing*, Mengolah dokumen dari file untuk mengubah file menjadi dokumen. Proses validasi terdiri dari data latih dan data uji. Kemudian masuk ke model algoritma Decision Tree yang didalamnya terdapat perhitungan algoritma, kemudian model diterapkan, setelah itu masuk evaluasi kinerja, kemudian muncul nilai akurasi dan AUC.

Tahap selanjutnya yaitu evaluasi model. Pada tahap evaluasi ini akan diketahui model yang sudah berhasil dibuat pada langkah sebelumnya akan memiliki nilai utilitas sebanyak apa. Validasi silang 10 kali lipat digunakan untuk evaluasi. Dari hasil pengujian yang telah dilakukan pada model algoritma *decision tree* yang digunakan menghasilkan nilai *Accuracy* (*Confusion Matrix*). Setelahnya adalah menghitung akurasi dari masing-masing algoritma. Dari hasil evaluasi model dengan menggunakan algoritma Decision Tree yang sebelumnya telah dilakukan tergambar hasil nilai *Accuracy* (*Confusion Matrix*) nya adalah sebagai berikut:



Gambar 3. Hasil Pengujian menggunakan *rapid miner* pada algoritma *Decision Tree*

Berdasarkan hasil pengujian pada gambar 3, akurasi model algoritma dapat dihitung dengan menggunakan persamaan berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+Fp+FN}$$

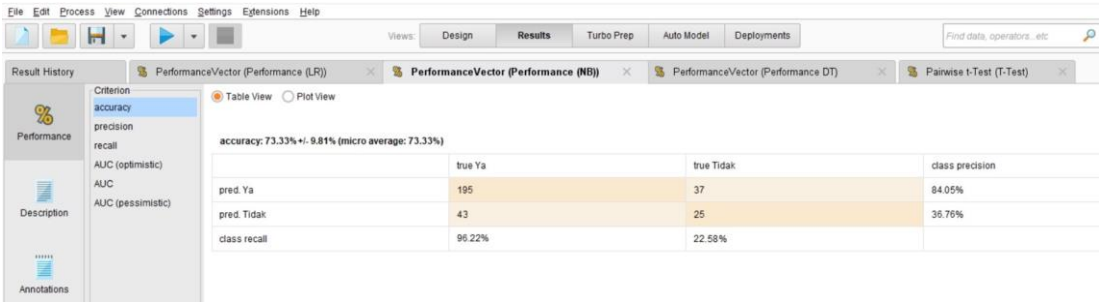
$$= \frac{229+14}{229+48+9+14}$$

$$= \frac{243}{300}$$

$$= 0{,}81$$

Jumlah *True Yes* (TP) sebanyak 229 record yang tergolong aman dan *False Not yes* (FN) sebanyak 9 record yang tergolong *False unsafe*. 48 catatan Benar Tidak berikutnya diklasifikasikan sebagai tidak aman dan 14 catatan Tidak Benar diklasifikasikan sebagai Tidak Aman. Berdasarkan gambar 3 diatas menunjukkan bahwa tingkat akurasi menggunakan algoritma *Decision Tree* adalah 81,00%.

Perhitungan untuk akurasi algoritma *Naïve Bayes*, didapatkan bahwa dengan algoritma *Naïve Bayes* nilai *Accuracy* (*Confusion Matrix*) yang diperoleh adalah:



Gambar 4. Hasil Pengujian menggunakan *rapid miner* pada algoritma *Naïve Bayes*

Berdasarkan hasil pengujian pada gambar 4, akurasi model algoritma dapat dihitung dengan menggunakan persamaan berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+Fp+FN}$$

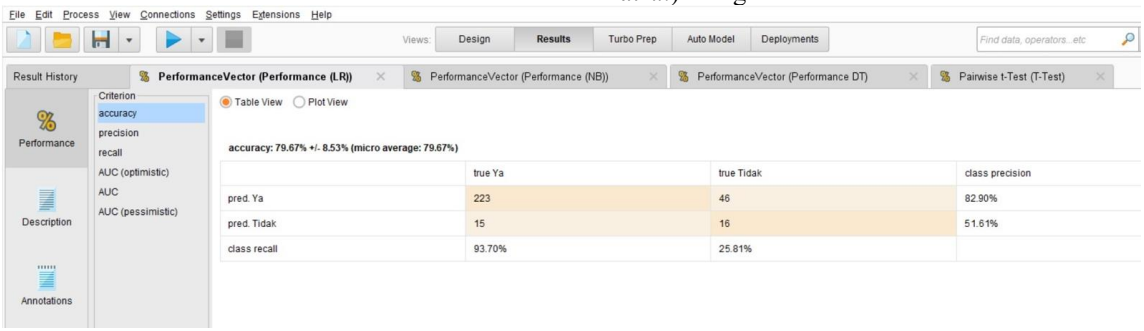
$$= \frac{195+25}{195+37+43+25}$$

$$= \frac{243}{300}$$

$$= 0{,}7333$$

Jumlah *True Yes* (TP) sebanyak 195 *record* yang tergolong aman dan *False Not yes* (FN) sebanyak 43 *record* yang tergolong *False unsafe*. 37 catatan Benar Tidak berikutnya diklasifikasikan sebagai tidak aman dan 25 catatan Benar Tidak diklasifikasikan sebagai Tidak Aman. Pada tabel 3 di atas memperlihatkan jumlah akurasi dari penggunaan algoritma *Naïve Bayes* adalah 73,33%.

Yang terakhir adalah penilaian akurasi dari algoritma *Logistic Regression*. Berdasarkan hasil pengujian model sebelumnya yang menggunakan algoritma *Logistic Regression* dapat terlihat hasil nilai *Accuracy* (*Confusion Matrix*) sebagai berikut:



Gambar 5. Hasil Pengujian menggunakan *rapid miner* pada algoritma *Logistic Regression*

Berdasarkan hasil pengujian pada gambar 5, akurasi model algoritma Naïve Bayes dapat dihitung dengan menggunakan persamaan berikut:

$$Accuracy = \frac{TP+TN}{TP+TN+Fp+FN}$$

$$= \frac{223+16}{223+46+15+16}$$

$$= \frac{239}{300}$$

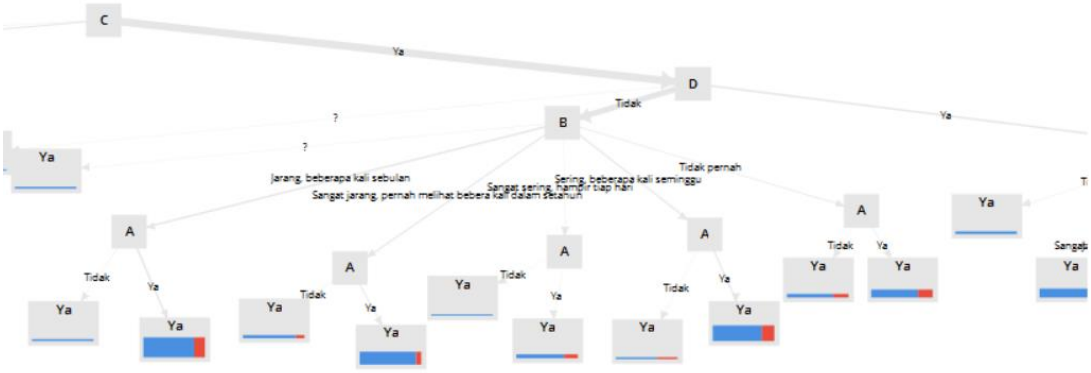
$$= 0{,}7966$$

Jumlah *True Yes* (TP) sebanyak 223 *record* yang tergolong aman dan *False Not yes* (FN) sebanyak 15 *record* yang tergolong *False unsafe*. Selanjutnya 46 Benar Tidak diklasifikasikan sebagai tidak aman dan 16 catatan Benar Tidak diklasifikasikan sebagai Tidak Aman. Berdasarkan gambar 5 di atas menunjukkan bahwa tingkat akurasi menggunakan algoritma *Logistic Regression* adalah 79,67%. Tahapan terakhir yaitu membandingkan hasil akurasi dari masing-masing algoritma. Berdasarkan hasil olah analisis dari 3 algoritma yang sudah dilakukan sebelumnya, maka dapat dirangkum hasilnya seperti pada tabel berikut ini:

Tabel 5. Hasil dari Perbandingan Performa ke-3 Algoritma

	Accuracy	AUC
Decision Tree	81,00%	0.628
Naïve Bayes	73.33%	0.679
Logistic Regression	79.67%	0.663

Dapat dilihat pada tabel bahwa akurasi tertinggi terdapat pada algoritma *Decision Tree* yang menghasilkan grafik seperti berikut:



Gambar 6. Grafik Hasil *Decision Tree*

4. Kesimpulan

Algoritma decision tree menghasilkan tingkat nilai akurasi yang tertinggi dibandingkan algoritma *Naïve Bayes* dan *Logistic Regression* dalam mengklasifikasi opini bahaya penggunaan rokok elektrik dengan akurasi 81,00% dan nilai AUC 0,628. Grafik hasil pengolahan data menggunakan algoritma *Decision Tree* menunjukkan bahwa pendapat bahwa rokok elektrik memang berbahaya. Dan faktor terbesar yang mempengaruhi pendapat tersebut adalah dari variabel C yaitu seseorang yang memiliki teman dekat yang merokok vape/rokok elektrik. Dengan hasil akurasi yang tinggi tersebut, penerapan algoritma *Decision Tree* pada penelitian ini memiliki akurasi yang lebih tinggi sehingga dapat digunakan untuk memberikan solusi atas masalah pro dan kontra apakah penggunaan rokok elektrik berbahaya atau tidak.

Daftar Pustaka

[1] World Heald Organization, *Who Global Report on Trends in Prevalence of Tobacco Smoking 2000-2025, Second Edition*. 2018.

[2] A. Sameera, R. K. Patel, K. Bharadwaj, M. M. A. Mir, and S. T. Ahmed, "Pros and Cons of e-cigarettes- A brief note," *J. Chem. , Biol. Phys. Sci.*, vol. 5, no. 4, pp. 4083–4088, 2015.

[3] G. Heydari, A. E. Ahmady, F. Chamyani, M. Masjedi, and L. Fadaizadeh, "Electronic Cigarette, Effective or Harmful for Quitting Smoking and Respiratory Health: A Quantitative Review Papers," *Lung India*, vol. 34, no. 1, pp. 25–28, 2017, doi: 10.4103/0970-2113.197119.

[4] D. Leader, "The Pros and Cons of Vaping Are They a Safer Alternative For People Living with COPD?," 2020. https://www.verywellhealth.com/the-pros-and-cons-of-e-cigarettes-915015.

[5] P. Callahan-Lyon, "Electronic Cigarettes: Human Health Effects," *Tob. Control*, vol. 23, no. SUPPL. 2, 2014, doi: 10.1136/tobaccocontrol-2013-051470.

[6] J. M. Rudd and J. Lewis Priestley, "A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit," *Grey Lit. from PhD Candidates*, vol. 5, no. January, 2017, [Online]. Available:

http://digitalcommons.kennesaw.edu/dataphdgreylit/5.

[7] A. Ashari, I. Paryudi, and A. M. Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 11, pp. 33–39, 2013.

[8] T. K. Bhowmik, "Naive bayes vs logistic regression: Theory, implementation and experimental validation," *Intel. Artif.*, vol. 18, no. 56, pp. 14–30, 2015, doi: 10.4114/ia.v18i56.1113.

[9] K. S. Nugroho, "Validasi Model Klasifikasi Machine Learning pada RapidMiner," *https://ksnugroho.medium.com/*, 2020. https://ksnugroho.medium.com/validasi-model-machine-learning-pada-rapidminer-50be0080df14.

[10] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery*, vol. 53, no. 9. Boca Raton: CRC Press, 2009.

[11] Ainurrohmah, "Akurasi Algoritma Klasifikasi pada Software Rapidminer dan Weka," in *Prisma*, 2021, vol. 4, pp. 493–499, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/.

[12] C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*. West Sussex: John Wiley & Sons Ltd Registered, 2009.

[13] J. Han, M. Kamber, and J. Pei, *Data mining Concepts and Techniques (Third Edition)*. Waltham: Morga Kaufmann Publishers, 2015.