Jurnal Jendela Matematika

Jurnal Jendela Matematika
Volume 2 No 01 Januari 2024
ISSN: 2986-7134 (Print) / ISSN: 2985-9875 (Online)
The article is published with Open Access at: https://www.ejournal.jendelaedukasi.id/index.php/JJM

Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube

Siti Mahmuda ✉, Universitas Mulawarman

✉ *sitimahmuda@fmipa.unmul.ac.id*

Abstract: Random Forest is an ensemble classifier in a machine-learning algorithm. The ensemble classifier aimed to improve model accuracy and classification performance. Based on accuracy measures, Random Forest shows the best performance with existing ensemble classifiers like Support Vector Machine (SVM) and AdaBoost. Hence, this research implement Random Forest on categorical of Youtube channel content. The predictor variables of classification are amount of subscribers, amount of videos, amount of video.s view, and age of Youtube channel. The Random Forest showed that the number of trees selected was 100 and the m being tried was 1. Amount of subscribers were the most influential variable in categorical of Youtube channel content with an importance of 19,04%. The Accuracy of classification was 77,27%.

Keywords: Random forest, ensemble classifier, youtube, accuracy.

Abstrak: *Random Forest* adalah metode klasifikasi *ensemble* dalam algoritma pembelajaran mesin. Metode klasifikasi *ensemble* bertujuan untuk meningkatkan akurasi model dan kinerja klasifikasi. Berdasarkan ukuran akurasi, *Random Forest* menunjukkan performa terbaik diantara metode klasifikasi yang ada, seperti *Support Vector Machine* (SVM) dan *AdaBoost*. Oleh karena itu, penelitian ini menerapkan metode klasifikasi *Random Forest* pada kategori konten kanal *Youtube*. Variabel prediktor klasifikasi adalah jumlah*subscribers*, jumlah video, jumlah penayangan video, dan lama kanal *Youtube*. Metode *Random Forest* menunjukkan jumlah pohon yang dipilih adalah 100 dan mtry adalah 1. Jumlah *subscribers* merupakan variabel yang paling berpengaruh dalam pengkategorian konten kanal *Youtube* dengan tingkat kepentingan sebesar 19,04%. Akurasi klasifikasi yang dihasilkan sebesar 77,27%.

Kata kunci: *Random forest*, klasifikasi *ensemble*, *youtube*, akurasi.

Received 29 November 2023; **Accepted** 11 Januari 2024; **Published** 25 Januari 2024

Citation: Mahmuda, S. (2024). Implementasi Metode Random Forest pada Kategori Konten Kanal Youtube. *Jurnal Jendela Matematika*, 2 (01), 21-31.

PENDAHULUAN

Youtube adalah salah satu situs yang dapat digunakan untuk menyebarkan konten berupa video pada internet. Sebuah kanal *Youtube* dapat mengunggah dan menyebarkan video dengan berbagai macam kategori konten. Berdasarkan data klasifikasi yang terdapat pada situs www.kaggle.com, kategori konten *Youtube* dapat berupa pendidikan, warta berita, hiburan dan lain sebagainya. Kategori konten *Youtube* dapat diklasifikasikan dengan menggunakan metode klasifikasi. Salah satu metode klasifikasi yang menggunakan teknik pohon keputusan adalah *Classification And Regression Trees* (CART). Proses perhitungan CART lebih cepat dan dapat menyelesaikan kasus dengan jumlah data yang besar. Namun, metode ini masih mempunyai kelemahan yaitu menghasilkan pohon yang kurang stabil, dimana perubahan kecil pada data *training* dapat menyebabkan perubahan yang signifikan pada pohon yang terbentuk (Dian, 2014) dan cenderung *overfitting* (Dhawangkara & Riksakomora, 2014). Sehingga untuk meningkatkan kestabilan dan menghindari *overfitting* maka diterapkan metode *ensemble* (Dian, 2014).

Metode klasifikasi *ensemble* merupakan metode klasifikasi yang menggabungkan beberapa algoritma klasifikasi untuk meningkatkan kekuatan model dan kinerja klasifikasi (Syarif dkk, 2012). Metode klasifikasi *ensemble* dianggap lebih efektif daripada pengklasifikasi tunggal karena alasan berikut : (1) dataset *training* tidak selalu memberikan informasi yang cukup untuk melakukan prediksi yang akurat, (2) proses *training* dalam klasifikasi lemah di kondisi tertentu, (3) ruang prediksi yang dicari mungkin tidak berisi fungsi target sebenarnya, sementara metode klasifikasi *ensemble* dapat memberikan prediksi yang baik (Diettarich, 2000). Metode klasifikasi *ensemble* dianggap lebih resisten terhadap *noise* dan mampu meminimalkan bias dan keragaman dibandingkan metode klasifikasi tunggal.

Beberapa metode yang menerapakan klasifikasi *ensemble* yaitu *Random Forest Support Vector Machine* (SVM), dan *AdaBoost*. Rata-rata akurasi yang dihasilkan ketiga algoritma tersebut cukup baik dan memiliki kinerja yang hampir sama dalam klasifikasi, namun metode *Random Forest* memiliki nilai metrik yang paling tinggi dibandingkan metode lainnya (Arumnisa & Wijayanto, 2023). Penelitian lain menunjukkan bahwa metode *Random Forest* memiliki hasil terbaik untuk semua kriteria akurasi, dengan membandingkan kinerja beberapa metode klasifikasi *ensamble*, antara lain *Bagging*, *Random Forest*, *AdaBoost*, *XGBoost*, dan *LightGBM* (Li & Weidong, 2020).

Penelitian ini bertujuan untuk implementasi metode klasifikasi *ensemble* yaitu metode *Random Forest* pada data kategori konten kanal *Youtube*. Hasil penelitian memberikan informasi mengenai klasifikasi kategori konten kanal *Youtube*, tingkat keakuratan klasifikasi, dan variabel yang paling berpengaruh terhadap kategori konten kanal *Youtube*.

METODE

Random Forest

Breiman pertama kali memperkenalkan metode *Random Forest* pada tahun 2001. Metode *Random Forest* memiliki dua fungsi untuk pemecahan suatu kasus, yaitu klasifikasi dan prediksi. Teknik dasar yang digunakan metode *Random Forest* adalah pohon keputusan. Dengan kata lain metode *Random Forest* merupakan kumpulan pohon keputusan untuk klasifikasi dan prediksi data dengan memberikan masukan ke dalam akar di bagian atas kemudian turun ke daun di bagian bawah (Wulansari, 2018). Hasil analisis metode *Random Forest* untuk klasifikasi adalah bentuk setiap pohon dari pohon-pohon yang terbangun, sedangkan hasil prediksi diperoleh dari nilai rata-rata setiap pohon (Lingga, 2011).

Metode *Random Forest* merupakan hasil pengembangan metode *Classification and Regression Tree* (CART) yang menerapkan metode agregasi *bagging* atau *bootstrap* dan pemilihan fitur secara acak. *Bagging* merupakan salah satu metode yang dapat meningkatkan hasil algoritma klasifikasi. Dasar dari metode *bagging* ini adalah metode *ensemble* (Samudera, 2019). Menurut (Rahmawati, 2015) , algoritma metode *Random Forest* yaitu :

1. Mengambil *n* data sampel dari dataset awal menggunakan metode *bootstrap resampling* dengan pengembalian.
2. Menyusun pohon klasifikasi dari setiap dataset hasil *bootstrap resampling*, menentukan pengklasifikasi terbaik berdasarkan variabel atribut yang diambil secara acak. Hitungan variabel yang diambil secara acak dapat ditentukan dengan menghitung $\frac{1}{2}\sqrt{m}$, $\sqrt{m}$ atau $2\sqrt{m}$, dimana *m* adalah banyaknya variabel prediktor.
3. Memprediksi klasifikasi data sampel berdasarkan pohon klasifikasi yang terbentuk.
4. Mengulangi langkah 1-3 hingga diperoleh jumlah pohon klasifikasi yang diinginkan, kemudian melakukan pengulangan sebanyak *k* kali.
5. Memprediksi klasifikasi data sampel akhir dengan mengkombinasikan hasil prediksi pohon klasifikasi yang diperoleh.

Algoritma *Random Forest* mempunyai nilai *m* yang dapat berbeda-beda. Nilai *m* merupakan banyaknya variabel prediktor yang digunakan sebagai pemisah dalam pembentukan pohon klasifikasi. Nilai *m* yang semakin besar akan menyebabkan korelasi yang semakin tinggi.

Informasi Gain

Perhitungan yang digunakan pada saat membangun pohon keputusan dengan metode CART adalah informasi *gain*. Informasi *gain* menggambarkan ukuran pemilihan variabel yang digunakan untuk klasifikasi oleh setiap simpul dalam sebuah pohon klasifikasi. Misalkan N adalah simpul untuk memisahkan setiap kelas berdasarkan variabel dari dataset yang dinotasikan dengan D. Pemisahan simpul dilakukan berdasarkan informasi *gain* tertinggi dari variabel tersebut. Rumus untuk mendapatkan informasi *gain* adalah sebagai berikut :

$$Informasi\ Gain(IG) = entropi(D) - entropi_A(D) \quad (1)$$

Nilai entropi(D) dapat diperoleh dengan menggunakan rumus (2) dan nilai $\text{entropi}_A(D)$ dapat diperoleh melalui rumus (3).

$$Entropi(D) = -\sum_{i=1}^{c} p_i^2 log(p_i) \quad (2)$$

dimana :

c : jumlah kelas target (variabel prediktor)

p_i : proporsi kelas ke-i partisi *D*

$$Entropi_A(D) = \sum_{j=1}^{v} \frac{n_j}{n} \times entropi(D_j) \quad (3)$$

dimana :

v : jumlah partisi

n : jumlah observasi

n_j : jumlah observasi *j*

D : partisi ke-j

Proses partisi setiap iterasi pada algoritma pohon klasifikasi pada dasarnya mencari metode partisi yang memberikan informasi *gain* tertinggi (Sartono & Dharmawan, 2023).

Akurasi

Jika model klasifikasi telah dihasilkan berdasarkan dataset, maka perlu diketahui seberapa baik kinerja prediksi model klasifikasi tersebut. Cara paling intuitif untuk melihat baik tidaknya hasil prediksi dari suatu model adalah dengan melihat akurasi prediksinya, yaitu

membandingkan kelas hasil prediksi dengan yang sebenarnya (Sartono & Dharmawan, 2023).

Menurut Carlo Vercellis dalam penelitian (Lareno, 2014), ada dua alasan untuk mengukur nilai akurasi suatu model prediksi. Pertama, nilai akurasi diperlukan untuk menentukan model prediksi dengan hasil paling akurat. Kedua, nilai akurasi dapat mendeteksi kekurangan pada model yang dibentuk. Model prediksi menghasilkan kinerja berdasarkan analisis menggunakan beberapa metode yang dapat diukur berdasarkan kesalahan dari hasil prediksi. Nilai *error* dapat diperoleh dengan menggunakan perhitungan *Mean Absolute Percentage Error* (MAPE). MAPE adalah persentase rata-rata dari seluruh perbedaan antara data aktual dan data hasil prediksi (Lingga, 2011).

$$MAPE = \frac{1}{n}\sum_{I=1}^{n}\frac{|X_i - F_i|}{X_i} \times 100 \quad (4)$$

dimana :

X_i : nilai aktual ke-i

F_i : nilai prediksi ke-i

n : jumlah observasi

Berdasarkan rumus MAPE pada persamaan (4) yang menggambarkan nilai *error* model prediksi, maka akurasi model dapat diperoleh dengan menggunakan perhitungan sebagai berikut :

$$Accuracy = 100\% - MAPE \quad (5)$$

Variabel Penting (Variable Importance)

Analisis tingkat kepentingan variabel dalam pemodelan dapat ditemukan dalam berbagai istilah, seperti analisis kepentingan variabel dan dominasi varibael. Dalam pemodelan prediktif, variabel prediktor yang penting adalah variabel yang apabila tidak dimasukkan ke dalam model maka kualitas hasil prediksi akan menurun secara signifikan.

Berdasarkan definisi di atas, secara intuitif untuk mengukur pentingnya suatu variabel prediktor digunakan perbandingan antara kualitas prediksi model yang melibatkan variabel tersebut dengan model prediksi yang tidak melibatkan variabel tersebut. Diasumsikan terdapat p variabel prediktor dan l merupakan salah satu variabel diantara variabel prediktor tersebut, maka dapat dirumuskan :

$$VI_l = akurasi(M_w) - akurasi(M_s) \quad (6)$$

dimana :

VI_l : penurunan tingkat akurasi model tanpa variabel l

M_w : model tanpa variabel l

M_s : model dengan variabel l

Variabel l akan dianggap penting ketika variabel tersebut dikeluarkan dari model, terjadi penurunan tingkat akurasi yang tinggi dan begitu pula sebaliknya.

Proses perhitungan VI_l umumnya bergantung pada akurasi data uji. Karena data uji diambil secara acak dari data lengkap, jika proses VI_l dihitung berkali-kali, kemungkinan akan diperoleh nilai akurasi yang berbeda-beda karena proses pengacakan. Perhitungan VI_l bergantung pada perbedaan rata-rata akurasi dari beberapa pengulangan. Statistik yang diperoleh dikenal sebagai penurunan rata-rata akurasi.

Selain menggunakan perubahan nilai akurasi, pada model klasifikasi berbasis pohon, ukuran kepentingan variabel juga dapat didasarkan pada perubahan nilai *gini*. Nilai *gini* dari simpul model M_w dibandingkan dengan nilai *gini* model M_s. Ukuran ini dikenal sebagai *mean decrease gini* (Sartono & Dharmawan, 2023).

Variabel dan Struktur Data

Dataset dalam penelitian ini adalah data kategori konten kanal *Youtube* dengan variabel-variabel atributnya. Jumlah kanal *Youtube* yaitu sebanyak 108. Dataset ini merupakan data

sekunder yang diperoleh dari situs *www.kaggle.com* tahun 2023. Terdapat 5 variabel yang digunakan dalam peneltian ini, disajikan pada Table 1.

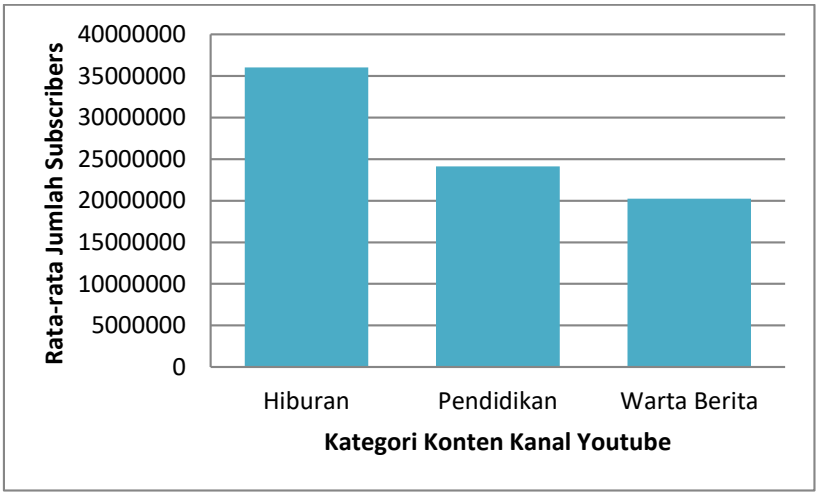
TABEL 1. *Variabel-variabel dataset*

No	Variabel	Deskripsi
1	Kategori konten kanal *Youtube*	Jenis kategori yang dilihat dari konten-konten video yang diunggah oleh sebuah kanal *Youtube*.
2	Jumlah *subscribers*	Jumlah akun *Youtube* yang berlangganan konten video atau informasi dari sebuah kanal *Youtube*.
3	Jumlah video	Jumlah video yang diunggah dan disebarkan oleh sebuah kanal *Youtube*.
4	Jumlah penayangan video	Jumlah berapa kali konten dari sebuah kanal *Youtube* dilihat atau tayang.
5	Lama akun *Youtube*	Usia akun *Youtube* sejak terdaftar di situs *Youtube*.

Varibel respon adalah variabel pada data yang berbentuk kategori, yaitu kategori konten kanal *Youtube*. Kategori konten kanal *Youtube* dalam penelitan ini adalah pendidikan, warta berita, dan hiburan. Variabel prediktornya antara lain jumlah *subscriber*, jumlah video, jumlah penayangan video, dan lama akun *Youtube*.

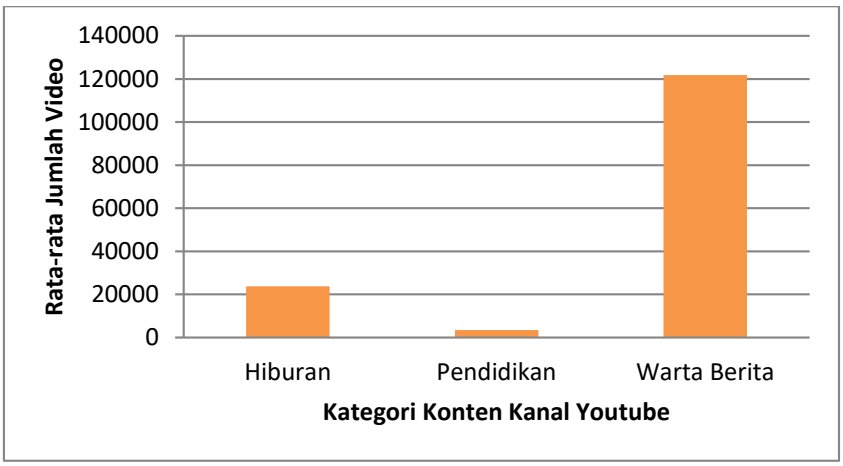
HASIL DAN PEMBAHASAN PENELITIAN

Statistika Deskriptif



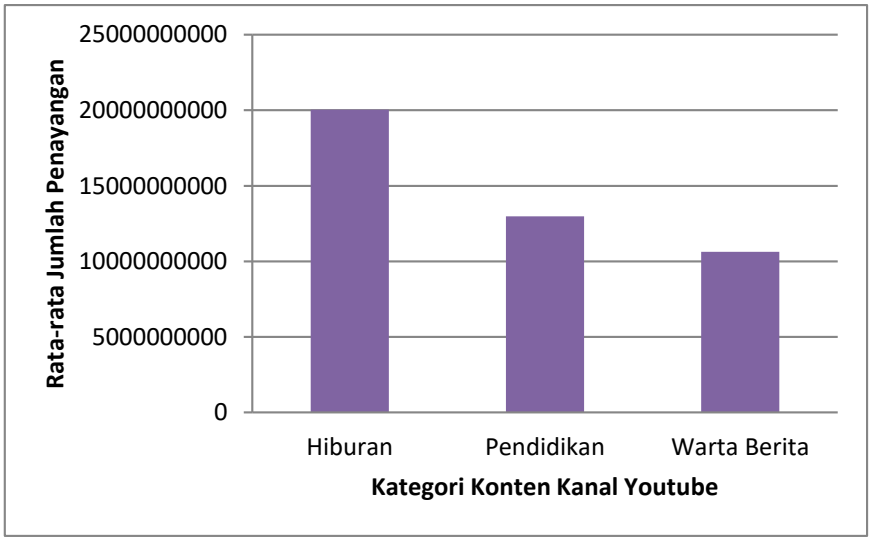
GAMBAR 1. Perbandingan rata-rata jumlah *subscribers* kanal *Youtube*

Perbandingan rata-rata jumlah *subscribers* di antara 3 kategori konten kanal *Youtube* dapat dilihat pada diagram batang Gambar 1. Diagram batang tersebut menunjukkan bahwa rata-rata jumlah *subscribers* terbanyak dimiliki oleh kanal *Youtube* dengan kategori konten hiburan. Kemudian disusul oleh kategori konten kanal *Youtube* bertema pendidikan. Rata-rata jumlah *subscribers* paling sedikit pada kategori konten kanal *Youtube* berupa warta berita.



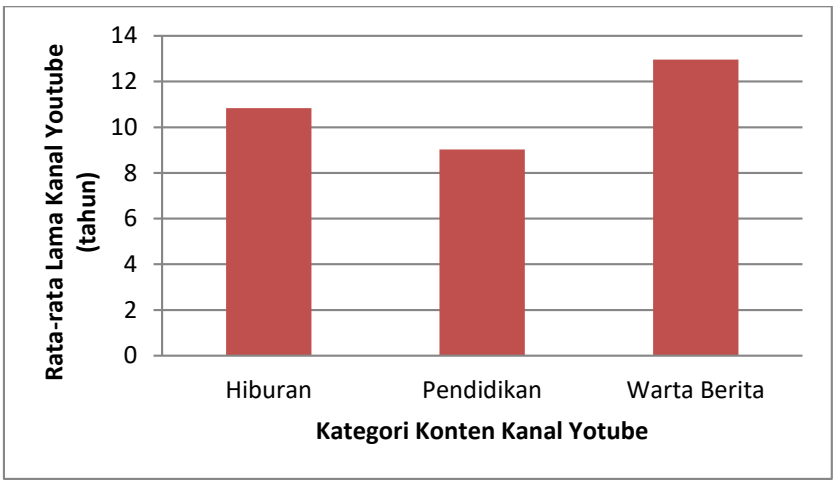
GAMBAR 2. Perbandingan rata-rata jumlah video kanal *Youtube*

Diagram batang pada Gambar 2 menunjukkan perbandingan rata-rata jumlah video dari 3 kategori konten kanal *Youtube*. Kategori pendidikan memiliki rata-rata jumlah video yang diunggah dan disebarkan melalui kanal *Youtube* paling sedikit. Konten kanal *Youtube* kategori pendidikan menunjukkan rata-rata jumlah video paling banyak, sedangkan kategori hiburan menyusul setelahnya.



GAMBAR 3. Perbandingan rata-rata jumlah penayangan video kanal *Youtube*

Diagram batang pada Gambar 3 tentang perbandingan rata-rata jumlah penayangan video kanal *Youtube* pada 3 kategori konten menunjukkan kecederungan yang sama dengan perbandingan rata-rata jumlah *subscribers*-nya pada Gambar 1. Hal ini memberikan indikasi bahwa jumlah penayangan video suatu kanal *Youtube* berbanding lurus dengan jumlah *subscriber*-nya. Dengan kata lain, jika jumlah *subscriber* semakin banyak, maka dapat meningkatkan jumlah penayangan suatu video di kanal *Youtube* dan sebaliknya.



GAMBAR 4. Perbandingan rata-rata jumlah lama kanal *Youtube*

Perbandingan rata-rata lama kanal *Youtube* pada Gambar 4 menunjukkan bahwa rata-rata usia kanal *Youtube* kategori warta berita yang paling lama, yaitu 13 tahun. Kategori hiburan memiliki rata-rata usia semenjak kanal dibuat adalah 11 tahun dan kategori konten kanal *Youtube* berupa pendidikan berusia rata-rata 9 tahun. Perbandingan ini juga selaras dengan perbandingan rata-rata jumlah video kanal *Youtube* pada Gambar 2. Sehingga dapat dikatakan bahwa semakin lama usia suatu kanal *Youtube*, maka jumlah video yang diunggah atau disebarkan terindikasi akan semakin banyak pula.

Data Latih dan Data Uji

Sebelum mengimplementasikan metode *Random Forest*, dataset awal dibagi menjadi 2, yaitu data latih dan data uji. Data latih adalah data yang digunakan untuk melatih algoritma membentuk sebuah model, sedangkan data uji adalah data yang digunakan untuk mengevaluasi performa atau akurasi dari model.

Penelitian ini membentuk proporsi data latih sebanyak 80% dan data uji sebanyak 20% dari dataset keseluruhan. Total observasi seluruh dataset adalah 108, sehingga data latih berjumlah 86 observasi dan data uji berjumlah 22 observasi. Algoritma *Random Forest* mengambil data latih dan data uji secara acak dari keseluruah dataset.

Algoritma Metode Random Forest

Algoritma *Random Forest* menggunakan nilai *mtry* untuk menentukan jumlah variabel prediktor yang dipilih secara acak. Penggunaan *mtry* yaitu sebagai kandidat pemisah dalam sebuah pohon klasifikasi. Penentuan nilai *mtry* pada penelitian ini disajikan sebagai berikut (Rahmawati, 2015) :

$$mtry_1 = \frac{\sqrt{jumlah\ variabel\ prediktor}}{2} = \frac{\sqrt{4}}{2} = 1$$

$$mtry_2 = \sqrt{jumlah\ variabel\ prediktor} = \sqrt{4} = 2$$

$$mtry_3 = \sqrt{jumlah\ varibel\ prediktor} \times 2 = \sqrt{4} \times 2 = 4$$

Perhitungan di atas menunjukkan jumlah variabel prediktor yang akan dievaluasi untuk menentukan jumlah variabel yang digunakan dalam partisi simpul pohon klasifikasi.

TABEL 2. *Misklasifikasi setiap mtry*

Mtry	Misklasifikasi
1	23,26%
2	26,74%
4	26,74%

Pemilihan nilai *mtry* didasarkan pada rata-rata tingkat misklasifikasi dari model *Random Forest*. Semakin kecil tingkat misklasifikasi, maka semakin optimal jumlah *mtry* atau variabel prediktor digunakan untuk membangun model. Tabel 2 menunjukkan bahwa *mtry* yang paling optimum berjumlah 1, karena *mtry* tersebut menghasilkan misklasifikasi yang paling kecil. Sehingga penelitian ini menggunakan *mtry* bernilai 1.

TABEL 3. *Misklasifikasi setiap jumlah pohon*

Jumlah pohon	Misklasifikasi
100	20,93%
500	23,26%
1000	22,09%

Akurasi klasifikasi *Random Forest* juga ditentukan oleh banyaknya jumlah pohon yang dibentuk (*number of tree*). Breiman (2001) menyatakan bahwa tingkat misklasifikasi *Random Forets* akan menuju ke nilai tertentu saat ukuran pohon *Random Forest* semakin besar (Breiman, 2001). Dalam penelitian ini dicoba tiga jumlah pohon yaitu 100, 500, dan 1000, karena analisis klasifikasi menggunakan jumlah pohon tersebut diperkirakan dapat mengurangi tingkat misklasifikasi (Dewi dkk, 2011).

Berdasarkan Tabel 3 klasifikasi *Random Forest* pada penelitian ini menetapkan jumlah pohon sebanyak 100, karena mempunyai tingkat misklasifikasi terkecil yaitu 20,93%. Artinya rata-rata kesalahan klasifikasi dengan jumlah pohon yang dibangun 100 merupakan tingkat kesalahan minimum.

Model klasifikasi *Random Forest* dengan jumlah *mtry* 1 dan pohon yang dibangun sebanyak 100 memperoleh hasil prediksi dari data uji yang ditampilkan pada Tabel 4.

TABEL 4. *Prediksi data uji*

Data Uji	Prediksi		
	Pendidikan	Warta Berita	Hiburan
Pendidikan	7	0	3
Warta Berita	1	4	0
Hiburan	1	0	6

Tabel 4 menunjukkan prediksi kategori konten kanal *Youtube* data uji yang berjumlah 22. Hasil prediksi yang benar untuk kategori pendidikan sebanyak 7 observasi dan 3 observasi dengan hasil prediksi yang salah. Pada kategori warta berita terdapat 1 prediksi yang salah dan 4 prediksi yang benar. Sedangkan prediksi yang salah pada kategori hiburan sebanyak 1 observasi dan prediksi yang benar sebaanyak 6 observasi. Kesimpulan dari hasil prediksi 22 observasi data uji adalah 5 observasi salah klasifikasi dan 17 observasi diprediksi dengan benar.

Akurasi

Akurasi model klasifikasi *Random Forest* dapat diperoleh dengan menggunakan rumus (4) dan (5) :

$$MAPE = \frac{1}{n}\sum_{I=1}^{n}\frac{|X_i - F_i|}{X_i} \times 100$$

$$MAPE = \frac{3+1+1}{22} \times 100 = 22{,}73$$

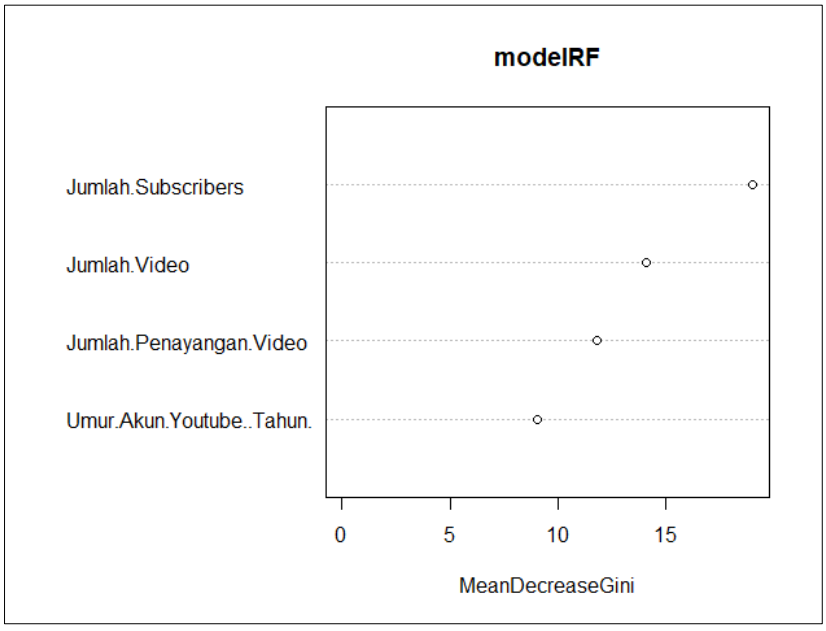
$$Accuracy = \ 100\% - MAPE$$

$$Accuracy = \ 100\% - 22{,}73 = 77{,}27\%$$

Berdasarkan perhitungan di atas diperoleh nilai akurasi klasifikasi data uji sebesar 77,27%. Nilai akurasi tersebut menunjukkan model klasifikasi *Random Forest* mampu mengklasifikasikan kategori konten kanal *Youtube* dengan benar sebesar 77,27%.

Variabel Penting

Gambar 5 dan Tabel 5 menunjukkan nilai *mean decrease gini* yang berguna untuk mendeskripsikan data dan memahami variabel yang penting dalam membangun model dan menentukan prediksi. Dengan demikian dalam penelitian ini diketahui bahwa variabel jumlah *subscribers* merupakan variabel dengan tingkat kepentingan tertinggi yang dapat mempengaruhi penentuan prediksi kategori konten kanal *Youtube* dengan nilai *mean decrease gini* sebesar 19,04. Kemudian tingkat kepentingan variabel selanjutnya adalah variabel jumlah video dengan nilai *mean decrease gini* 14,10, lalu variabel jumlah penayangan video dengan nilai *mean decrease gini* sebesar 11,78, dan variabel lama akun *Youtube* dengan nilai *mean decrease gini* sebesar 9,06 yang juga merupakan variabel dengan tingkat kepentingan terendah. Variabel dengan tingkat kepentingan yang tinggi menjadi pendorong keakuratan model yang dibentuk karena hasil prediksi akan mendekati nilai sebenarnya.



GAMBAR 5. *Mean decrease gini* variabel prediktor

TABEL 5. *Mean decrease gini* variabel prediktor

Variabel prediktor	Mean decrease gini
Jumlah *subscribers*	19,04
Jumlah video	14,10
Jumlah penayangan video	11,78
Lama akun *Youtube*	9,06

SIMPULAN

Hasil peneltian ini memberikan beberapa kesimpulan, yaitu :

1. Statistik deskriptif menunjukkan jumlah penayangan video suatu kanal *Youtube* berbanding lurus dengan jumlah *subscribers*-nya dan dapat diketahui bahwa semakin lama usia suatu kanal *Youtube*, maka jumlah video yang diunggah atau disebarkan terindikasi akan semakin banyak pula.
2. Klasifikasi menggunakan metode *Random Forest* dengan nilai *mtry*=1 dan jumlah pohon=100 menghasilkan tingkat akurasi sebesar 77,27%, yang berarti akurasi metode klasifikasi *Random Forest* untuk data kategori konten kanal *Youtube* tinggi.
3. Variabel penting yang paling besar pengaruhnya dalam pengkategorian konten kanal *Youtube* berdasarkan nilai *mean decrease gini* adalah variabel jumlah *subscribers*.

DAFTAR PUSTAKA

1. Arumnisa R.I. dan Wijayanto A.W. (2023). Perbandingan Metode Ensemble Learning: Random Forest, SVM, AdaBoost pada Klasifikasi Indeks Pembangunan Manusia (IPM). *Jurnal Sistem Informasi*, vol.12(1).
2. Breiman. L. (2001). Random Forest. *Machine Learning Journal*, vol. 45(5).
3. Dewi N.K., Syafitri U.D., dan Mulyadi S.Y. (2011) . Penerapan Metode Random Forest dalam Driver Analysis. *Statistics and Its Application*, vol.16(1).
4. Dhawangkara M. dan Riksakomora E. (2014). Prediksi Intensitas Hujan Kota Surabaya dengan Matlab Menggunakan Teknik Random Forest dan CART (Studi Kasus Kota Surabaya). *Jurnal Teknik ITS*, vol.6(1).
5. Dian S. (2014). Klasifikasi Rumah tangga Sangat Miskin di Kabupaten Jombang Menurut Paket Bantuan Rumah Tangga yang Diharapkan dengan Pendekatan Random Forest Classification and Regression Trees (RF-CART). *Jurnal Sains dan Seni*, pp 1-7.
6. Diettarich T.G. (2000). Ensemble Method in Machine Learning. *Springer-Verlag Berlin*, vol.1857, pp.1-15.
7. *Global Youtube Statistitics 2023 Dataset*. https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023. Diakses pada tanggal 29 November 2023.
8. Lareno B. (2014). Analisa dan Perbandingan Akurasi Model Prediksi Rentet Waktu Arus Lintas Jangka Pendek. *CSRID Journal*, vol. 6(3).
9. Li Y dan Weidong C. (2020). A Comparative Perfomance Assesment of Ensemble Learning for Credit Scoring. *Mathematics*, vol.8(10).
10. Lingga P.R.D. (2011). Deteksi Gempa Berdasarkan Data Twitter menggunakan Decision Tree, Random Forest, dan SVM. *Jurnal Teknik ITS*, vol.160.

11. Rahmawati I. (2015). Klasifikasi Faktor-faktor yang Mempengaruhi Korban Kecelakaan Lalu Lintas di Surabaya dengan Pendekatan Regresi Logistik Multinomial dan Random Forest. *Tesis ITS*. Surabaya.
12. Samudra A.Y. (2019). Pendekatan Random Forest untuk Model Peramalan Harga Tembakau Rajangan di Kabupaten Temanggung. *Tesis*. Yogyakarta.
13. Sartono B dan Dharmawan H. (2023). Pemodelan Prediksi Berbasis Pohon Klasifikasi. *IPB Press Bogor*.
14. Syarif I dkk. (2012). Application of bagging, Boosting, and Stacking to Intrusion Detection. *Springer-Verlag Berlin*, vol.7376, pp.593-602.
15. Wulansari M.J. (2018). Analisis Faktor-faktor yang Mempengaruhi Seseorang Terkena Penyakit Diabetes Melitus Menggunakan Regresi Random Forest. *Tesis.* Yogyakarta.

PROFIL SINGKAT

Siti Mahmuda adalah penulis yang berasal dari Universitas Mulawarman, Samarinda, Kalimantan Timur.