Penerapan Machine Learning terhadap Klasifikasi Kanker Payudara pada Pasien

Fayza Apriliza[1], Dzakiyyah Al Kaazhim[1], and Naufal Mufid F[2]
[1]*Program Studi Sistem Informasi, Fakultas Informatika Institut Teknologi Telkom Purwokerto*
[2]*Program Studi Informatika, Fakultas Teknologi Industri Institut Teknologi Nasional Bandung*

***Abstract*— Machine learning adalah bagian dari Artificial Intelligence (AI) yang membuat sistem memiliki kemampuan belajar secara otomatis dan meningkatkan kemampuannya berdasarkan pengalaman tanpa diprogram secara eksplisit. Kanker payudara identik dengan sebuah keganasan yang dapat berakibat pada kematian. Penyakit ini merupakan masalah kesehatan baik di negara maju maupun negara berkembang. Keterlambatan dalam mendeteksi gejala kanker payudara menyebabkan risiko kematian menjadi jauh lebih tinggi. Oleh Karena itu, pada paper dan percobaan ini dilakukan proses klasifikasi jenis sel kanker pada pasien pengidap kanker payudara. Dewasa ini perkembangan teknologi semakin pesat dan hal tersebut mempengaruhi banyak bidang tidak terkecuali bidang kesehatan atau bidang medis. Dengan perkembangan teknologi tadi proses pengambilan keputusan dapat dipermudah dengan melakukan automasi. Salah satu contohnya adalah pengimplementasian program untuk menentukan atau mengklasifikasi jenis sel kanker payudara menggunakan beberapa algoritma klasifikasi. Ada banyak algoritma pada metode klasifikasi, namun algoritma yang digunakan untuk proses klasifikasi jenis sel kanker payudara pada penelitian ini adalah algoritma terbaik berdasarkan perbandingan beberapa algoritma, seperti Support Vector Classifier (SVC), K-Nearest Nerighbor (KNN), Logistic Regression, Random Forest, Decision Tree, dan Gradient Boosting Classifier. Hasil yang diperoleh algoritma terbaik adalah *Random Forest* dengan *accuracy score* sebesar 0.953216 dan F1-*Score* sebesar 0.936508.**

I. PENDAHULUAN

Machine learning adalah bagian dari Artificial Intelligence (AI) yang membuat sistem memiliki kemampuan belajar secara otomatis dan meningkatkan kemampuannya berdasarkan pengalaman tanpa diprogram secara eksplisit. Fokus machine learning terdapat pada pengembangan program komputer yang dapat mengakses data dan belajar dari data tersebut. Hingga saat ini terdapat beberapa algoritma Machine Learning yang dapat digunakan dan dikembangkan untuk berbagai tujuan.

Kanker payudara disebut juga carcinoma mamae merupakan suatu jenis kanker yang dapat menyerang siapa saja baik kaum wanita maupun pria. Kanker payudara ini tumbuh dalam kelenjar susu, jaringan lemak, maupun pada jaringan ikat payudara. Kanker payudara ini diidentikkan dengan sebuah keganasan yang dapat berakibat pada kematian. Oleh karena itu, kanker payudara ini masih menjadi hal yang menakutkan khususnya pada kaum wanita.

Berdasarkan WHO, kanker payudara adalah kanker yang paling sering terjadi pada wanita, berdampak pada 2,1 juta wanita setiap tahunnya, dan menyebabkan jumlah terbesar kematian akibat kanker payudara [1]. Penyakit ini juga merupakan masalah kesehatan baik di negara maju maupun negara berkembang. Menurut Departemen Kesehatan, angka penderita di Indonesia sendiri ada sebesar 876.665 orang [2]. Penyakit ini ada di peringkat kedua setelah kanker rahim dikarenakan rata-rata penderitanya adalah 10 dari 100 ribu perempuan [3]. Keterlambatan dalam mendeteksi gejala kanker payudara menyebabkan banyak penderita baru mengetahui kondisinya setelah memasuki stadium yang tinggi (rata-rata pada stadium III dan IV). Pada kondisi ini, risiko kematian menjadi jauh lebih tinggi.

Oleh Karena itu, pada paper dan percobaan ini dilakukan proses klasifikasi jenis sel kanker menggunakan beberapa algoritma klasifikasi. Ada banyak algoritma pada metode klasifikasi, namun yang digunakan untuk proses klasifikasi jenis sel kanker payudara pada penelitian ini adalah Support Vector Classifier (SVC), K-Nearest Nerighbor (KNN), Logistic Regression, Random Forest, Decision Tree, dan Gradient Boosting Classifier.

Penulisan ini bertujuan untuk mengetahui apa saja yang mempengaruhi diagnosis jenis sel kanker, khususnya pada kanker payudara dan bagaimana hasil akurasi dari model yang telah dibuat. Hasil diagnosis dapat menghasilkan suatu klasifikasi penentuan jenis sel kanker payudara bersifat ganas atau jinak.

II. LANDASAN TEORI

A. Kanker Payudara

Kanker payudara adalah pertumbuhan sel yang abnormal pada jaringan payudara seseorang. Payudara wanita terdiri dari lobulus (kelenjar susu), duktus (saluran susu), lemak dan jaringan ikat, pembuluh darah dan *limfe*. Sebagian besar kanker payudara bermula pada sel-sel yang melapisi duktus (kanker duktal), beberapa bermula di lobulus (kanker lobular), serta sebagian kecil bermula di jaringan lain [4]. Ada dua jenis kanker payudara yaitu kanker payudara benign (jinak) dan kanker payudara malignant (ganas).

B. Machine Learning

Salah satu bagian dari Artificial Intelligent (AI) yang berfokus pada pengembangan sistem yang mampu belajar sendiri tanpa harus diprogram berulang kali adalah Machine Learning. Machine Learning merupakan pemrograman komputer untuk mencapai kriteria/performa tertentu dengan menggunakan sekumpulan data training atau pengalaman di masa lalu [5]. Penelitian terkini mengungkapkan bahwa machine learning terbagi menjadi tiga kategori: Supervised Learning, Unsupervised Learning, Reinforcement Learning (Somvanshi & Chavan, 2016) [6].

C. *Klasifikasi*

Klasifikasi merupakan salah satu ilmu yang terdapat pada machine learning dan merupakan algoritma supervised learning karena memiliki label/target. Terdapat banyak metode yang ada dalam klasifikasi supervised learning, diantaranya adalah Regresi Logistik, K-nearest Neighboor, Super Vector Machine, Naive Bayes, Decision Tree, dan Random Forest[7].

D. *Random Forest*

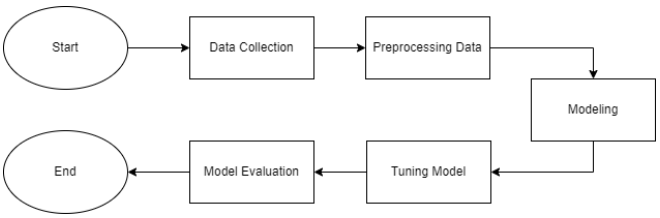
Random Forest adalah algoritma turunan dari pohon keputusan yang membagi dataset menjadi beberapa bagian dan menggunakan algoritma pohon keputusan pada setiap subset. Random Forest dapat mengatasi overfitting data [8]. Random Forest merupakan salah satu algoritma supervised learning yang dalam prosesnya dilakukan pemecahan data secara random ke dalam decision tree dan menggabungkan tree dengan training pada data yang dimiliki [9]. Algoritma ini berupa kombinasi dari beberapa tree predictors atau bisa disebut decision trees dimana setiap tree bergantung pada nilai random vector yang dijadikan sampel secara bebas dan merata pada semua tree dalam forest tersebut. Hasil prediksi dari Random Forest didapatkan melalui hasil terbanyak dari setiap individual decision tree (voting untuk klasifikasi dan rata-rata untuk regresi). Untuk RF yang terdiri dari N trees dirumuskan sebagai berikut:

$$l(y) = argmax\left(\sum_{n=1}^{N} I_{hn}(y) = c\right) \qquad (1)$$

Dimana I adalah fungsi indikator dan *hn* adalah tree ke-n dari RF.

III. METODOLOGI

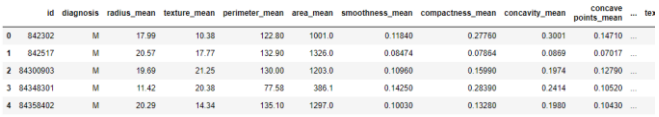
Penelitian ini memiliki beberapa tahapan yang ditunjukan pada Gambar 1.



Gambar 1 Alur penelitian

Penjelasan detail dari masing-masing tahapan sebagai berikut:

1. *Data Collection*
   Pada tahapan ini dilakukan pencarian *dataset* yang akan digunakan. *Dataset* yang digunakan pada *project* ini adalah Breast Cancer Wisconsin (Diagnostic) Dataset yang diunduh melalui link: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. Project ini akan memprediksi apakah kanker payudara yang diderita seorang pasien termasuk ganas atau tidak. Gambar 3 menunjukan gambaran *dataset* yang digunakan.



	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	tex
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

Gambar 2 Dataset

Dataset yang digunakan memiliki jumlah data sebanyak 569 data dengan jumlah pasien kanker payudara *benign* sebanyak 357 data dan jumlah data pasien kanker payudara *malignant* sebanyak 212. Variabel respon atau target data (Y) dalam percobaan adalah kolom diagnosis yang terdiri dari dua kategori yaitu benign dan malignant. Sedangkan variabel predictor (X) pada percobaan adalah semua kolom dari dataset kecuali diagnosis, yaitu *radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst,* dan *fractal_dimension_worst*. Berikut informasi terkait dataset:

a. Dataset terdiri atas *numerical data* dan *categorical data*. *Categorical data* hanya terdapat di kolom diagnosis yang mana kolom tersebut akan menjadi data *target project* ini. Nantinya, perlu mengubah *categorical data* tersebut menjadi *numerical* dengan *label encoding.*
b. Adanya kolom yang tidak diperlukan untuk proses *modeling*, yakni kolom id yang berisi nomor identitas pasien dan kolom Unnamed: 32
c. Tidak terdapat *missing value* selain dikolom Unnamed: 32 sehingga kita tidak memerlukan proses *handling missing value*
d. *Range data* yang berbeda-beda sehingga memerlukan proses *rescale* data dengan *normalization* atau *standardization*
e. Jumlah *feature* yang sangat banyak, yaitu sebanyak 33 *features*.

2. *Pre-processing Data*
   Pada tahap *pre-processing data* dilakukan beberapa tahapan sebagai berikut.
   a. Melakukan *drop column* yang tidak diperlukan untuk proses *modeling* (kolom yang tidak digunakan adalah kolom id dan kolom Unnamed: 32).
   b. *Label encoding*.
   c. Melakukan pengecekan *missing value* untuk memastikan.
   d. Melakukan pemisahan *data feature* dan *data target*.

e. Melakukan *rescale data* dengan menggunakan *standardization*.
f. Melakukan *feature selection*.

3. *Modeling*
   a. Melakukan pemisahan *data test* dan *data train* dengan perbandingan 3:7.
   b. Melakukan persiapan model algoritma, seperti *Support Vector Machine Classifier, K-Neighbors Classifier, Logistic Regression Random Forest Classifier, Decision Tree Classifier*, dan *Gradient Boosting Classifier*.
   c. Melakukan evaluasi semua model

4. *Tuning Model*
   Dari beberapa algoritma yang telah dibandingkan, dipilih model dengan *accuracy* dan *F1-Score* terbaik untuk dilakukan *tuning hyperparameter*. Pada proses ini dilakukan dua uji coba, yakni *tuning* menggunakan *RandomizedSearchCV* dan *GridSearchCV*.

5. *Evaluation*
   Pada tahapan dilakukan evaluasi terhadap model yang dibangun dengan menggunakan *accuracy* dan *F1-Score*. Akurasi adalah perbandingan antara true prediksi baik positif maupun negatif terhadap keseluruhan data. Sedangkan, *F1-Score* adalah perbandingan presisi dan perolehan rata-rata tertimbang.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN} \times 100\%$$

Gambar *3 F1 score* dan *Accuracy*

IV. ANALISIS DAN PEMBAHASAN

A. *Data Collection*

Dataset yang digunakan yaitu terkait dengan data diagnosis Kanker Payudara Wisconsin. Dataset tersebut bersumber dari kaggle, dimana dataset tersebut memiliki 569 data di dalam 33 kolom.

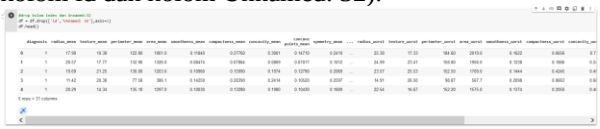
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 33 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
 16  smoothness_se            569 non-null    float64
 17  compactness_se           569 non-null    float64
 18  concavity_se             569 non-null    float64
 19  concave points_se        569 non-null    float64
 20  symmetry_se              569 non-null    float64
 21  fractal_dimension_se     569 non-null    float64
 22  radius_worst             569 non-null    float64
 23  texture_worst            569 non-null    float64
 24  perimeter_worst          569 non-null    float64
 25  area_worst               569 non-null    float64
 26  smoothness_worst         569 non-null    float64
 27  compactness_worst        569 non-null    float64
 28  concavity_worst          569 non-null    float64
 29  concave points_worst     569 non-null    float64
 30  symmetry_worst           569 non-null    float64
 31  fractal_dimension_worst  569 non-null    float64
 32  Unnamed: 32              0 non-null      float64
dtypes: float64(31), int64(1), object(1)
memory usage: 146.8+ KB
```

Gambar 4 Kolom pada dataset

B. *Pre-processing Data*

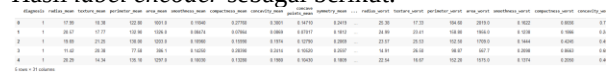
Pre-processing adalah tahapan dimana data atau dataset dipersiapkan untuk diolah lebih lanjut. Pada tahapan ini data yang tidak diinginkan akan dieliminasi dan data yang akan digunakan akan ditingkatkan kualitasnya dengan beberapa proses sebagai berikut.

a. Melakukan *drop column* yang tidak diperlukan untuk proses *modeling* (kolom yang tidak digunakan adalah kolom id dan kolom Unnamed: 32).



Gambar 5 *Drop column*

b. *Label encoding*.
   Ketika *Label encoding* digunakan untuk mengubah nilai kategorikal menjadi nilai numerik pada beberapa kolom. Hasil *label encoder* sebagai berikut.



Gambar 6 Hasil *label encoder*

c. Melakukan pengecekan *missing value* untuk memastikan.
   Ketika dataset tidak memiliki missing value, maka tidak perlu me-replace dengan nilai median atau mean. Hasil pengecekan missing value sebagai berikut.

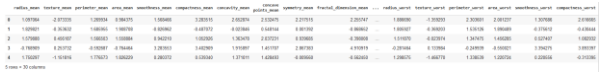
diagnosis	0
radius_mean	0
texture_mean	0
perimeter_mean	0
area_mean	0
smoothness_mean	0
compactness_mean	0
concavity_mean	0
concave points_mean	0
symmetry_mean	0
fractal_dimension_mean	0
radius_se	0
texture_se	0
perimeter_se	0
area_se	0
smoothness_se	0
compactness_se	0
concavity_se	0
concave points_se	0
symmetry_se	0
fractal_dimension_se	0
radius_worst	0
texture_worst	0
perimeter_worst	0
area_worst	0
smoothness_worst	0
compactness_worst	0
concavity_worst	0
concave points_worst	0
symmetry_worst	0
fractal_dimension_worst	0

dtype: int64

Gambar 7 Pengecekan *missing value*

d. Melakukan pemisahan *data feature* dan *data target*.
Pada kasus ini yang menjadi *data target* ialah diagnosis sedangkan yang menjadi *data features* adalah kolom lain selain diagnosis.

e. Melakukan *rescale data* dengan menggunakan *standardization*.
Selanjutnya, dilakukan *rescale data* dengan menggunakan standarisasi. Hasil standarisasi sebagai berikut.



Gambar 8 Hasil *standarisasi*

f. Melakukan *feature selection*.
Berdasarkan feature selection tersebut didapatkan bahwa concave points_worst, perimeter_worst, concave points_mean, radius_worst, perimeter_mean, area_worst, radius_mean, area_mean, concavity_mean, concavity_worst, compactness_mean, compactness_worst, radius_se, perimeter_se, dan area_se merupakan lima belas feature atau atribut yang sangat mempengaruhi diagnosis.

C. *Modeling*

Kasus Diagnosis Kanker Payudara Wisconsin ini termasuk ke kategori klasifikasi sehingga model yang digunakan adalah model algoritma klasifikasi. Beberapa model algoritma klasifikasi digunakan dan model dilatih dengan parameter default. Hasil *accuracy score* dan F1-Score dari setiap model ditunjukan pada Gambar 9.

	Model	Accuracy Score	F1 Score
0	Random Forest	0.953216	0.936508
1	Logistic Regression	0.941520	0.918033
2	Decision Tree	0.912281	0.883721
3	K-Neighbors	0.953216	0.934426
4	Support Vector Machine	0.941520	0.918033
5	Gradient Boosting	0.941520	0.921875

Gambar 9 Hasil *accuracy score* dan *f1 score*

Berdasarkan hasil tersebut, *random forest* merupakan model dengan F1-Score paling tinggi sehingga model tersebut yang digunakan. Untuk mendapatkan tingkat error yang lebih minimum dilakukan *tuning hyperparameter* dengan menggunakan *randomized search* dan *GridSearch*.

D. *Tuning Model*

Dari beberapa algoritma yang telah dibandingkan, dipilih model dengan accuracy dan F1-Score terbaik untuk dilakukan tuning hyperparameter. Pada proses ini dilakukan dua uji coba, yakni tuning menggunakan RandomizedSearchCV dan GridSearchCV. Berikut merupakan hasil tuning hyperparameter menggunakan RandomizedSearchCV.

	precision	recall	f1-score	support
0	0.95	0.97	0.96	107
1	0.95	0.92	0.94	64
accuracy			0.95	171
macro avg	0.95	0.95	0.95	171
weighted avg	0.95	0.95	0.95	171

Gambar 10 Hasil *tuning parameter* menggunakan *RandomizeSearchCv*

Berikut merupakan hasil tuning hyperparameter menggunakan GridSearch.

	precision	recall	f1-score	support
0	0.95	0.97	0.96	107
1	0.95	0.92	0.94	64
accuracy			0.95	171
macro avg	0.95	0.95	0.95	171
weighted avg	0.95	0.95	0.95	171

Gambar 11 Hasil *tuning parameter* menggunakan *GridSearch*

E. *Evaluation*

Pada tahapan dilakukan evaluasi terhadap model yang dibangun dengan menggunakan accuracy dan F1-Score. Akurasi adalah perbandingan antara true prediksi baik positif maupun negatif terhadap keseluruhan data. Sedangkan, F1-

Score adalah perbandingan presisi dan perolehan rata-rata tertimbang.

Setelah model random forest dilakukan tuning hyperparameter menggunakan randomized search dan GridSearch, model dilatih dan diuji kembali yang mana didapatkan accuracy score sebesar 0.95 dan F1-Score 0,96.

V. KESIMPULAN

Hal-hal yang mempengaruhi diagnosis kanker bisa berbeda-beda antar jenis kanker. Pada diagnosis kanker payudara sangat dipengaruhi oleh concave points_worst, perimeter_worst, concave points_mean, radius_worst, perimeter_mean, area_worst, radius_mean, area_mean, concavity_mean, concavity_worst, compactness_mean, compactness_worst, radius_se, perimeter_se, dan area_se.

Pada penelitian dilakukan percobaan perbandingan beberapa algoritma pada dataset. Berdasarkan percobaan tersebut didapatkan bahwa metode atau model Random Forest memiliki accuracy score dan F1-Score tertinggi dibandingkan metode algoritma lainnya.

REFERENCES

[1] Kusumawaty, J. *et al.* (2021) 'Efektivitas Edukasi SADARI (Pemeriksaan Payudara Sendiri) Untuk Deteksi Dini Kanker Payudara', *ABDIMAS: Jurnal Pengabdian Masyarakat*, 4(1), pp. 496–501. doi: 10.35568/abdimas.v4i1.1177.

[2] Kusminarto. Deteksi Sangat Dini Kanker Payudara, Jawaban untuk Menghindar. Departemen Kesehatan. 2006, [internet], Tersedia dalam:<https://www.litbang.depkes.go.id/aktual/kliping/payudara190906.htm>

[3] Anon. (2005) Kanker Payudara Stadium Dini dapat Diobati. 2005. [internet]Tersedia dalam:<https://situs.kesrepro.info/ aging/ agu/ 2005/ag01.htm>

[4] Ellis, E.O., Schnitt, S.J., S.-Garau, X., Bussolati, G., Tavassaoli, F.A., Eusebi, V. Pathology and Genetic of Tumours of The Breast and Female Genital Organs / WHO Classification of Tumours. Washington: IARC Press; 2003. P.10, 34-6.

[5] Chazar, C. and Erawan, B. (2020) 'Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine', *INFORMASI (Jurnal Informatika dan Sistem Informasi)*, 12(1), pp. 67–80. doi: 10.37424/informasi.v12i1.48.J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," *IEEE J. Quantum Electron.*, submitted for publication.

[6] Roihan, A., Sunarya, P. A. and Rafika, A. S. (2020) 'Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper', *IJCIT (Indonesian Journal on Computer and Information Technology)*, 5(1), pp. 75–82. doi: 10.31294/ijcit.v5i1.7951.

[7] Pamungkas, F. S., Prasetya, B. D. and Kharisudin, I. (2020) 'Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python', *PRISMA, Prosiding Seminar Nasional Matematika*, 3, pp. 692–697. Available at: https://journal.unnes.ac.id/sju/index.php/prisma/article/view/37875.

[8] Ferdyandi, M., Setiawan, N. Y. and Bachtiar, F. A. (2022) 'Prediksi Potensi Penjualan Makanan Beku berdasarkan Ulasan Pengguna Shopee menggunakan Metode Decision Tree Algoritma C4 . 5 dan Random Forest (Studi Kasus Dapur Lilis)', 6(2), pp. 588–596.

[9] Wandani, A. (2021) 'Sentimen Analisis Pengguna Twitter pada Event Flash Sale Menggunakan Algoritma K-NN, Random Forest, dan Naive Bayes', *Jurnal Sains Komputer & Informatika (J-SAKTI*, 5(2), pp. 651–665.