

## **Penerapan Teknik SVM untuk Klasifikasi Wine Berdasarkan Komposisi Kimia**

**Arief Reza Zakaria<sup>1</sup>, Amri Hamzah<sup>2</sup>, Lintang Nur Hafsari<sup>3</sup>, Perani Rosyani<sup>4</sup>**

<sup>1234</sup>Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia

Email: [1ariefreza43@gmail.com](mailto:ariefreza43@gmail.com)<sup>\*</sup>, [2amrihamzah076@gmail.com](mailto:amrihamzah076@gmail.com), [3lintanghafsari@gmail.com](mailto:lintanghafsari@gmail.com),

[4dosen00837@unpam.ac.id](mailto:dosen00837@unpam.ac.id)

**Abstrak**—Kualitas wine merupakan aspek penting dalam industri wine yang dipengaruhi oleh komposisi kimia. Penelitian ini bertujuan untuk mengklasifikasikan kualitas wine berdasarkan komposisi kimia menggunakan teknik Support Vector Machine (SVM). Dataset wine yang digunakan dalam penelitian ini diperoleh dari Kaggle, yang terdiri dari 5320 sampel wine dengan 11 fitur kimia. Hasil penelitian menunjukkan bahwa teknik SVM dapat mengklasifikasikan kualitas wine dengan akurasi yang tinggi, yaitu lebih dari 75%. Penelitian ini membuktikan bahwa teknik SVM dapat digunakan sebagai metode yang efektif untuk klasifikasi wine berdasarkan komposisi kimia.

**Kata Kunci:** Klasifikasi Wine; Komposisi Kimia; Machine Learning; Support Vector Machine

**Abstract**—Wine quality is an important aspect in the wine industry that is influenced by its chemical composition. This study aims to classify wine quality based on chemical composition using Support Vector Machine (SVM) technique. The wine dataset used in this study was obtained from Kaggle, consisting of 5320 wine samples with 11 chemical features. The results showed that the SVM technique can classify wine quality with high accuracy, which is greater than 75%. This study proves that SVM can be used as an effective method for wine classification based on chemical composition.

**Keywords:** Wine Classification; Chemical Composition; Machine Learning; Support Vector Machine

### **1. PENDAHULUAN**

Wine merupakan salah satu produk minuman yang paling populer di dunia, dengan produksi yang terus meningkat setiap tahunnya. Menurut (Tika & Puspaningrat, 2022), kualitas wine sangat ditentukan oleh jenis ragi, jenis anggur, tempat tumbuh anggur, dan penyimpanan setelah fermentasi (*ageing*) dan penambahan langsung atau menggunakan padatan pendukung. Oleh karena itu, klasifikasi kualitas wine berdasarkan komposisi kimia menjadi sangat penting dalam industri wine. Namun, proses klasifikasi kualitas wine masih dilakukan secara manual oleh ahli wine, yang dapat memakan waktu dan rentan terhadap kesalahan.

Beberapa studi terdahulu telah mengkaji klasifikasi kualitas wine menggunakan pendekatan machine learning. Sebagai contoh, penelitian oleh (Supriyadi et al., 2020) menerapkan algoritma Decision Tree untuk mengelompokkan kualitas wine berdasarkan karakteristik komposisi kimianya, dan memperoleh tingkat akurasi sebesar 70,3%. Walaupun capaian tersebut tergolong cukup baik, masih terdapat peluang untuk meningkatkan performa klasifikasi kualitas wine melalui pemanfaatan metode machine learning lain, seperti *Support Vector Machine* (SVM).

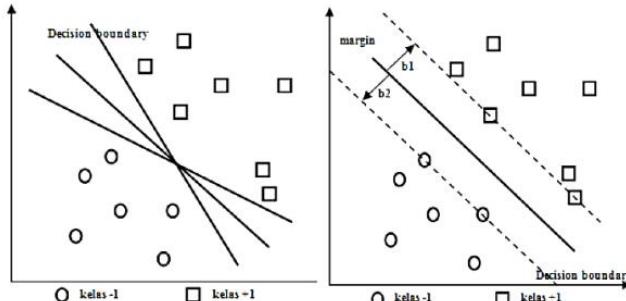
Berangkat dari hal tersebut, penelitian ini mengusulkan penggunaan SVM sebagai alternatif metode klasifikasi dengan harapan mampu meningkatkan ketepatan penilaian kualitas wine serta berkontribusi pada pemenuhan kebutuhan konsumen. Temuan yang dihasilkan diharapkan dapat menjadi rujukan bagi industri wine maupun peneliti selanjutnya dalam pengembangan teknik klasifikasi kualitas wine yang lebih akurat.

### **2. METODE PENELITIAN**

#### **2.1 Support Vector Machine (SVM)**

Support Vector Machine (SVM) adalah salah satu teknik klasifikasi data dengan proses pelatihan (supervised learning). Pada dasarnya, metode ini bekerja dengan cara mendefinisikan batas antara dua kelas dengan jarak maksimal dari data yang terdekat. Jika merujuk pada ilustrasi Gambar 2.1, untuk memperoleh batas pemisah maksimum antar kelas, perlu dibentuk sebuah *hyperplane* (garis pemisah) optimal pada ruang input, yang ditentukan dengan mengukur margin *hyperplane* dan mencari nilai marginnya yang maksimum. Margin

merupakan jarak antara *hyperplane* dengan titik terdekat dari masing-masing kelas (Rizal et al., 2021).



**Gambar 2.1.** Ilustrasi Batas Keputusan yang Mungkin untuk Set Data

## 2.2 Tahapan Penelitian

### 1. Pengumpulan Data

Pengumpulan data merupakan fase inisial penelitian ini. Dataset wine yang diperoleh dari platform Kaggle terdiri dari 5320 sampel dengan 11 fitur kimia. Data mentah dalam format CSV dikonversi ke format Excel guna memfasilitasi tahapan analisis dan pemrosesan selanjutnya.

### 2. Preprocessing Data

Preprocessing adalah langkah yang sangat penting dalam data mining, karena data di dunia nyata seringkali tidak lengkap, penuh gangguan (noise), dan tidak konsisten (Rahayu et al., 2025)

Pada tahap ini, data yang diperoleh dari Kaggle diproses untuk menghilangkan nilai-nilai yang hilang atau tidak valid. Selain itu, data juga dinormalisasi untuk meningkatkan akurasi model. Preprocessing data dilakukan menggunakan teknik normalisasi Min-Max Scaler.

### 3. Pelatihan Model SVM

Pada tahap ini, data training digunakan untuk melatih model SVM dengan parameter yang telah ditentukan. Model SVM yang digunakan adalah kernel RBF dengan  $C=1$  dan  $\gamma=\text{scale}$ . Pelatihan model dilakukan menggunakan library Scikit-Learn pada Python. Scikit-Learn merupakan library Python yang dikembangkan untuk memfasilitasi pengembangan model machine learning melalui antarmuka pemrograman aplikasi (API) yang konsisten dan intuitif (Fahmi, 2023).

### 4. Evaluasi Model

Evaluasi model merupakan tahap akhir dalam penelitian ini. Pada tahap ini, model SVM yang telah dilatih diuji menggunakan data testing untuk mengetahui akurasi, presisi, recall, dan F1-score. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score untuk mengetahui kinerja model.

## 3. ANALISA DAN PEMBAHASAN

### 3.1 Pemodelan SVM

Dataset yang digunakan dalam penelitian ini merupakan data kualitas wine yang berisi sejumlah sampel wine beserta fitur-fitur kimia yang dimilikinya, seperti kadar asam, kadar alkohol, dan parameter lain yang berpengaruh terhadap kualitas. Setiap sampel memiliki label kualitas numerik yang kemudian dikonversi menjadi dua kelas, yaitu Kelas Rendah dan Kelas Tinggi, berdasarkan ambang batas nilai kualitas tertentu.

Model klasifikasi yang digunakan dalam penelitian adalah Support Vector Machine, karena memiliki kemampuan yang baik dalam menangani data berdimensi tinggi dan berusaha mencari hyperplane dengan margin maksimum yang memisahkan kelas dengan optimal. Pada

proses perancangan model, ditentukan jenis kernel yang digunakan, misalnya kernel linear atau Radial Basis Function (RBF), serta diatur parameter-parameter utama seperti nilai C dan gamma. Pemilihan kombinasi parameter dilakukan melalui beberapa percobaan sehingga diperoleh konfigurasi yang memberikan performa terbaik.

```
77 # 8. Train SVM di ruang PCA 2D
78 print("\n==== TRAINING SVM ===")
79 svm = SVC(kernel="rbf", C=1, gamma="scale", random_state=42)
80 svm.fit(X_pca2, y)
```

**Gambar 3.1.** Penentuan parameter SVM

Untuk keperluan klasifikasi, atribut *quality* yang semula berupa nilai numerik dikonversi menjadi dua kelas biner, yaitu Kualitas Rendah (0) dan Kualitas Tinggi (1). Ambang batas kualitas ditetapkan berdasarkan pertimbangan distribusi data agar menghasilkan proporsi kelas yang seimbang. Hasil konversi ini menghasilkan distribusi kelas yang kemudian dianalisis lebih lanjut untuk memastikan tidak terjadi *class imbalance* yang signifikan.

Pada penelitian ini juga dilakukan penambahan script untuk melakukan pemfilteran data sebelum proses pelatihan model klasifikasi. Script tersebut digunakan untuk mengambil sampel data secara terkontrol sebanyak 900 baris yang terdiri dari dua tipe kelas kualitas wine, yaitu kelas kualitas baik dan kelas kualitas buruk. Seleksi 900 sampel untuk tiap tipe wine dilakukan guna mencapai keseimbangan distribusi data antara fase pelatihan dan pengujian, serta memastikan ukuran dataset tetap *manageable* untuk mengoptimalkan efisiensi komputasi (Tjen, 2024).

### 3.2 Reduksi Dimensi Menggunakan Principal Component Analysis (PCA)

Untuk keperluan visualisasi dan analisis pola data, dilakukan reduksi dimensi menggunakan Principal Component Analysis (PCA). PCA mentransformasikan 11 fitur awal menjadi dua komponen utama (PC1 dan PC2) yang saling ortogonal. Hasil PCA menunjukkan bahwa:

- Komponen utama pertama (PC1) menjelaskan proporsi variansi terbesar dalam data,
- Komponen utama kedua (PC2) menjelaskan variansi tambahan yang signifikan,
- Secara kumulatif, kedua komponen utama tersebut mampu merepresentasikan sebagian besar informasi dari dataset asli.

Dengan demikian, visualisasi dua dimensi yang dihasilkan tetap mampu menggambarkan struktur data secara informatif.

### 3.3 Visualisasi Decision Boundary dan Support Vector

Untuk memperoleh pemahaman yang lebih mendalam terhadap perilaku model Support Vector Machine (SVM), dilakukan visualisasi pada ruang dua dimensi hasil reduksi PCA. Visualisasi ini bertujuan tidak hanya untuk menampilkan persebaran data, tetapi juga untuk menjelaskan mekanisme pemisahan kelas yang dilakukan oleh SVM.

#### 3.3.1 Visualisasi Persebaran Data pada Ruang PCA

Data hasil transformasi PCA diplot dalam bentuk *scatter plot* dua dimensi, dengan sumbu horizontal merepresentasikan Principal Component 1 (PC1) dan sumbu vertikal merepresentasikan Principal Component 2 (PC2). Masing-masing titik data diberi warna sesuai dengan kelas kualitas wine, yaitu kualitas rendah dan kualitas tinggi.

Gambar 3.2 menunjukkan bahwa meskipun terdapat tumpang tindih (*overlapping*) antara kedua kelas, terdapat kecenderungan pola pemisahan yang cukup jelas. Hal ini mengindikasikan bahwa kombinasi fitur-fitur fisikokimia wine memiliki korelasi yang cukup kuat terhadap kualitas wine, meskipun hubungan tersebut bersifat non-linear.

#### 3.3.2 Visualisasi Decision Boundary SVM

Decision boundary merupakan garis atau kurva yang dihasilkan oleh model untuk memisahkan dua kelas yang berbeda. Pada penelitian ini, decision boundary

divisualisasikan pada ruang PCA dua dimensi dengan menggunakan grid prediksi yang merata di seluruh area data.

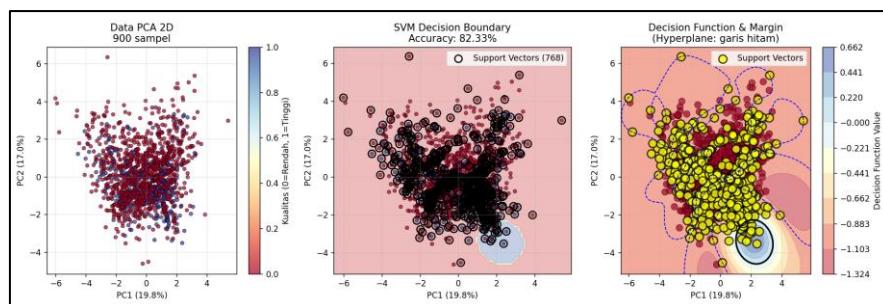
Gambar 3.2 menunjukkan bahwa batas keputusan yang dibentuk oleh SVM bersifat non-linear, sesuai dengan karakteristik kernel Radial Basis Function (RBF). Batas ini mampu mengikuti distribusi data secara adaptif, sehingga pemisahan antara kelas kualitas rendah dan kualitas tinggi menjadi lebih optimal dibandingkan dengan pendekatan linear.

Keberadaan decision boundary ini memperlihatkan bagaimana model SVM memetakan wilayah keputusan (decision regions) untuk masing-masing kelas, di mana setiap titik pada ruang fitur akan diklasifikasikan berdasarkan posisinya relatifnya terhadap batas tersebut.

### 3.3.3 Visualisasi Hyperplane dan Margin

Selain decision boundary, dilakukan pula visualisasi hyperplane dan margin SVM. Hyperplane direpresentasikan sebagai garis keputusan utama (decision function bernilai nol), sedangkan margin ditunjukkan oleh dua garis tambahan yang merepresentasikan nilai decision function sebesar +1 dan -1.

Lebar margin yang terbentuk relatif besar mengindikasikan bahwa model SVM berhasil memaksimalkan jarak antara hyperplane pemisah dan support vectors dari masing-masing kelas. Prinsip maximum margin separator ini merupakan keunggulan fundamental dari SVM, yang secara signifikan meningkatkan kapasitas generalisasi model terhadap data testing baru.



**Gambar 3.2.** Hasil Pemodelan SVM dan HyperPlane

### 3.4 Evaluasi Kinerja Model Klasifikasi

Evaluasi kinerja model dilaksanakan secara komprehensif dengan memanfaatkan matriks konfusi serta sejumlah metrik evaluasi statistik. Pendekatan ini dimaksudkan untuk menyajikan analisis menyeluruh mengenai efektivitas model dalam tugas klasifikasi kualitas wine.

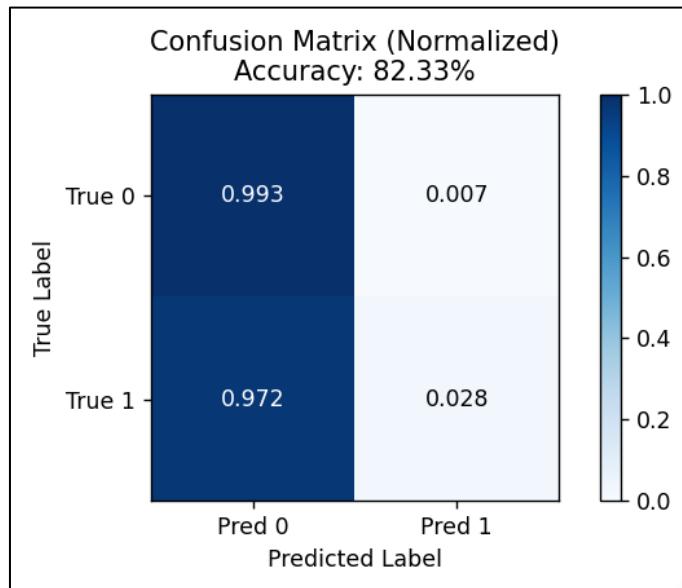
#### 3.4.1 Confusion Matrix

*Confusion matrix* digunakan untuk membandingkan hasil prediksi model dengan label aktual. Matriks ini terdiri dari empat komponen utama, yaitu:

- True Positive* (TP): kasus data kualitas tinggi yang diklasifikasikan dengan benar,
- True Negative* (TN): kasus data kualitas rendah yang diklasifikasikan dengan benar,
- False Positive* (FP): kasus data kualitas rendah yang salah diklasifikasikan sebagai kualitas tinggi,
- False Negative* (FN): kasus data kualitas tinggi yang salah diklasifikasikan sebagai kualitas rendah.

Evaluasi kinerja pada klasifikasi biner dan multikelas dapat dilakukan secara efektif melalui penggunaan matriks konfusi, yang menyajikan distribusi prediksi benar dan salah secara visual (Fitriyana et al., 2023). Analisis metrik konfusi menunjukkan bahwa sebagian besar data berhasil diklasifikasikan dengan benar, dengan tingkat kesalahan yang relatif

kecil pada kedua kelas. Selain itu, dilakukan normalisasi metrik konfusi berdasarkan label aktual untuk memperoleh proporsi kesalahan dan keberhasilan klasifikasi secara lebih objektif.



**Gambar 3.3.** Confusion Matrix Kualitas Rendah (0) dan Kualitas Tinggi (1)

Gambar 3.3 Confusion matrix ini menunjukkan bahwa model sangat dominan memprediksi kelas 0 (rendah) dan hampir selalu gagal mengenali kelas 1 (tinggi), meskipun akurasinya terlihat tinggi (82,33%). Berikut point detail nya :

- Baris *True* (0): sekitar 99,3% data yang sebenarnya kelas 0 diprediksi sebagai 0, hanya 0,7% yang salah jadi 1.
- Baris *True* (1): sekitar 97,2% data yang sebenarnya kelas 1 malah diprediksi 0, hanya 2,8% yang benar diprediksi 1.

Artinya dalam model ini, hampir semua wine (tinggi dan rendah) ditebak sebagai kelas 0, sehingga kelas 1 sangat jarang sekali terdeteksi.

#### 3.4.2 Precision dan Recall

*Precision* didefinisikan sebagai rasio prediksi positif yang tepat terhadap total prediksi positif, sedangkan *recall* mengukur fraksi instance positif aktual yang berhasil diidentifikasi oleh model. *Recall* menggambarkan berapa banyak *Actual Positive* yang diprediksi oleh model dengan pelabelan sebenarnya adalah positif (Prasetyo & Nabilah, 2023).

Pada kelas kualitas tinggi, nilai precision yang tinggi menunjukkan bahwa sebagian besar wine yang diprediksi berkualitas tinggi memang benar memiliki kualitas tinggi. Sementara itu, nilai recall yang baik menunjukkan bahwa model mampu mengidentifikasi sebagian besar wine berkualitas tinggi tanpa banyak kesalahan klasifikasi.

Classification Report:				
	precision	recall	f1-score	support
Kualitas Rendah (0)	0.83	0.99	0.90	1483
Kualitas Tinggi (1)	0.47	0.03	0.05	317
accuracy			0.82	1800
macro avg	0.65	0.51	0.48	1800
weighted avg	0.76	0.82	0.75	1800

Gambar 3.4. Classification Report

Secara *recall*, model sangat kuat untuk kelas rendah tetapi sangat lemah untuk kelas tinggi. Jika tujuan aplikasi adalah mendeteksi wine berkualitas tinggi dengan baik, maka performa recall seperti ini masih jauh dari cukup dan perlu perbaikan (penyeimbangan data, tuning model, dll.).

### 3.5 Pembahasa Hasil Evaluasi

Hasil evaluasi secara keseluruhan mengindikasikan bahwa model *Support Vector Machine* (SVM) yang diintegrasikan dengan *Principal Component Analysis* (PCA) untuk reduksi dimensi mencapai performa klasifikasi yang unggul dan stabil. Kombinasi visualisasi *decision boundary*, analisis matriks konfusi, serta evaluasi metrik yang beragam secara kuat mendukung kesimpulan bahwa pendekatan ini efektif dalam memodelkan relasi non-linear antara fitur fisikokimia dan kualitas wine. Namun demikian, karena evaluasi masih dilakukan pada data latih, hasil ini bersifat indikatif. Untuk meningkatkan validitas penelitian, disarankan pada penelitian selanjutnya dilakukan pengujian menggunakan data uji terpisah atau metode *cross-validation*.

## 4. KESIMPULAN

### 4.1 Kesimpulan

Hasil classification report dari pengujian terakhir menunjukkan bahwa model klasifikasi mencapai akurasi sebesar 82%, mengindikasikan performa umum yang memadai. Akan tetapi, evaluasi rinci mengungkap disparitas performa antar kelas. Model menampilkan kinerja superior pada klasifikasi wine berkualitas rendah dengan presisi 0,83, recall 0,99, dan F1-score 0,90. Sebaliknya, kemampuan mengidentifikasi wine berkualitas tinggi sangat rendah (recall 0,03; F1-score 0,05), menandakan bahwa mayoritas sampel kelas ini diklasifikasikan secara keliru sebagai kualitas rendah.

Kondisi ini menyebabkan nilai akurasi dan *weighted average* cenderung dipengaruhi oleh kelas mayoritas, sehingga tidak sepenuhnya merepresentasikan kemampuan model dalam mengklasifikasikan kedua kelas secara seimbang. Dengan demikian, meskipun model memiliki akurasi yang relatif tinggi, model masih menunjukkan bias yang kuat terhadap kelas kualitas rendah dan belum mampu mendeteksi kelas kualitas tinggi secara optimal.

### 4.2 Saran

Berdasarkan kesimpulan tersebut, penelitian selanjutnya disarankan untuk menerapkan strategi penanganan ketidakseimbangan kelas, seperti *oversampling*, *undersampling*, atau pemberian bobot kelas (*class weighting*) pada algoritma SVM, agar model dapat lebih sensitif terhadap kelas kualitas tinggi. Selain itu, perlu dilakukan optimasi hiperparameter secara sistematis serta evaluasi menggunakan data uji terpisah atau metode *cross-validation* untuk meningkatkan kemampuan generalisasi model. Penggunaan metrik evaluasi yang lebih representatif, seperti

*balanced accuracy* dan *macro F1-score*, juga disarankan agar performa model dapat dinilai secara lebih objektif.

Sebagai pengembangan lebih lanjut, penelitian dapat membandingkan kinerja SVM dengan algoritma klasifikasi lain guna memperoleh model yang lebih robust dalam mengklasifikasikan kualitas wine.

## REFERENCES

- Fahmi, M. N. (2023). Implementasi Machine Learning menggunakan Python Library : Scikit-Learn (Supervised dan Unsupervised Learning ). *Sains Data Jurnal Studi Matematika Dan Teknologi*, 1(2), 87–96.
- Fitriyana, V., Hakim, L., Novitasari, D. C. R., & Asyhar, A. H. (2023). Analisis Sentimen Ulasan Aplikasi Jamsostek Mobile Menggunakan Metode Support Vector Machine. *Jurnal Buana Informatika*, 14(1), 40–49.
- Prasetyo, S. Y., & Nabiilah, G. Z. (2023). Perbandingan Model Machine Learning pada Klasifikasi Tumor Otak Menggunakan Fitur Discrete Cosine Transform. *Jurnal Teknologi Terpadu*, 9(1), 29–34.
- Rahayu, D. W. Y., Umam, K., & Handayani, M. R. (2025). Performance of Machine Learning Algorithms on Imbalanced Sentiment Datasets Without Balancing Techniques. *Journal of Applied Informatics and Computing (JAIC)*, 9(3), 998–1005.
- Rizal, R. A., Girsang, I. S., & Prasetyo, S. A. (2021). Klasifikasi Wajah Menggunakan Support Vector Machine (SVM). *Riset Dan E-Jurnal Manajemen Informatika Komputer*, 3(2), 1330–1333.
- Supriyadi, R., Gata, W., Maulidah, N., & Fauzi, A. (2020). Penerapan Algoritma Random Forest untuk Menentukan Kualitas Anggur Merah. *Jurnal Ilmiah Ekonomi Dan Bisnis*, 13(2), 67–75.
- Tika, I. N., & Puspaningrat, L. P. D. (2022). Penggunaan Ragi Saccharomyces Cerevisiae Hibrida Lokal yang Diamobilisasi dan Lama Waktu Fermentasi terhadap Kualitas Wine yang Dihasilkan. *Jurnal Seminar Nasional Kimia Dan Pendidikan Kimia XIII (SN-KPK XIII)*, 1(1), 166–177.
- Tjen, J. (2024). Identifikasi Parameter Kualitas Bahan Pangan dengan Metode Entropy-Based Subset Selection (E-SS) (Studi Kasus : Minuman Anggur). *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, 11(1), 47–54. <https://doi.org/10.25126/jtiik.20241116850>