Journal of Science and Social Research ISSN 2615 – 4307 (Print)
June 2026, IX (3): 4149 – 4157 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

KOMPARASI ALGORITMA *K-NEAREST NEIGHBOR* DAN *NAIVE BAYES* UNTUK KLASIFIKASI KELAYAKAN EKSPOR KOPI ARABIKA DENGAN *CORRELATION-BASED FEATURE SELECTION*

Diva Agustin Purba[1], Nazma Aulia[2], Pujawati[3], Dicky Apdillah[4], Bambang Irwansyah[5], Harmayani[6]
Universitas Asahan, Kisaran
e-mail: [1*]divaagustinpurba@email.com, [4]dickyapdi1404@email.com

Abstract*: Arabica coffee is a high-value export commodity for the Indonesian economy. To maintain global competitiveness, coffee beans must meet export feasibility quality standards based on the cupping score from the Coffee Quality Institute (CQI). This study aims to compare the performance of the K-Nearest Neighbor (KNN) and Gaussian Naive Bayes algorithms in classifying the export feasibility of Arabica coffee beans. In sensory quality testing, high-dimensional attributes (10 parameters) can cause accuracy instability and increase computational load. Therefore, Correlation-based Feature Selection (CFS) was applied to reduce redundant features. The CFS process filtered the initial 10 features into 4 selected features (Flavor, Acidity, Cupper Points, and Aroma). The Dataset was obtained from the Kaggle Coffee Quality Database and had undergone Excel format adjustments before being loaded into the system. Modeling was conducted using 1,303 data records with a 30% split for testing data. The evaluation results showed that the Naive Bayes algorithm provided the best performance, achieving an accuracy rate of 97.95%. The use of CFS proved successful in reducing feature dimensions by 60% without significantly decreasing classification accuracy.*

Keywords: *Arabica Coffee; Export Classification; K-Nearest Neighbor; Naive Bayes; Correlation-based Feature Selection.*

Abstrak: Kopi Arabika merupakan komoditas ekspor bernilai tinggi bagi perekonomian Indonesia. Untuk menjaga daya saing di tingkat global, biji kopi harus memenuhi standar kualitas kelayakan ekspor berdasarkan *cupping score* dari *Coffee Quality Institute* (CQI). Penelitian ini bertujuan membandingkan kinerja algoritma *K-Nearest Neighbor* (KNN) dan *Gaussian Naive Bayes* dalam klasifikasi kelayakan ekspor biji kopi Arabika. Dalam pengujian sensoris mutu, dimensionalitas atribut yang tinggi (10 parameter) dapat menyebabkan ketidakstabilan akurasi dan meningkatkan beban komputasi. Oleh karena itu, *Correlation-based Feature Selection* (CFS) diterapkan untuk mereduksi fitur redundan. Proses CFS menyaring 10 fitur awal menjadi 4 fitur terpilih (*Flavor, Acidity, Cupper Points, Aroma*). *Dataset* diambil dari *Kaggle Coffee Quality Database* dan telah melalui tahapan penyesuaian format Excel sebelum dimuat ke sistem. Pemodelan dilakukan menggunakan 1303 rekam data dengan pembagian data uji sebesar 30%. Hasil evaluasi menunjukkan bahwa algoritma *Naive Bayes* memberikan performa terbaik dengan tingkat akurasi mencapai 97,95%. Penggunaan CFS terbukti berhasil memangkas dimensi fitur sebesar 60% tanpa menurunkan akurasi klasifikasi secara signifikan.

Kata Kunci: Kopi Arabika; Klasifikasi Ekspor; *K-Nearest Neighbor*; *Naive Bayes*; *Correlation-based Feature Selection*.

PENDAHULUAN

Indonesia secara geografis terletak pada posisi strategis yang sangat mendukung pertumbuhan komoditas pertanian tropis, salah satu yang paling

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
June 2026, IX (3): 4149 – 4157 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

menonjol adalah kopi. Biji kopi Arabika asal Indonesia, seperti Mandheling, Gayo, Toraja, dan Java Preanger, diakui memiliki profil rasa unik yang sangat diminati oleh konsumen global(Ridwan & Rakhmawati, 2023; Sebatubun & Pujiarini, 2018). Penjualan kopi ke pasar internasional (ekspor) merupakan salah satu pilar penopang devisa negara di sektor perkebunan.

Untuk menjaga harga dan prestise kopi Arabika Indonesia di kancah internasional, kualitas produk harus dijamin. Badan standardisasi internasional seperti *Coffee Quality Institute* (CQI) menetapkan protokol *cupping score* yang ketat(*Coffee Quality Institute Provides Coffee Education throughout the Coffee Value Chain.*, n.d.). Proses pengujian mutu kopi dinilai berdasarkan sepuluh indikator sensoris yang dinilai oleh ahli berlisensi (*Q-Grader*).

Namun, proses pengujian *Q-Grader* konvensional menghadapi tantangan dari aspek ketersediaan tenaga ahli, biaya sertifikasi yang tinggi, serta potensi bias penilaian subjektif manusia akibat kelelahan fisik. Oleh sebab itu, otomatisasi penilaian menggunakan teknologi data mining dan *machine learning* menjadi alternatif solutif untuk membantu memprediksi kelayakan mutu biji kopi secara cepat, akurat, dan transparan(Hasibuan et al., 2025; Maulana Iksan et al., 2020).

Penelitian terdahulu menunjukkan bahwa data *cupping score* memiliki dimensionalitas tinggi dengan sepuluh variabel numerik kontinu yang saling berkorelasi erat(Suyanto, 2019; Witten et al., 2017). Keberadaan atribut yang saling tumpang tindih (*redundant*) atau kurang relevan dapat menurunkan performa klasifikasi (*curse of dimensionality*). Metode *Correlation-based Feature Selection* (CFS) digunakan untuk mengatasi masalah tersebut dengan menyeleksi subset fitur yang berasosiasi tinggi dengan kelas target namun berasosiasi rendah antar-fitur(Hall, 1999).

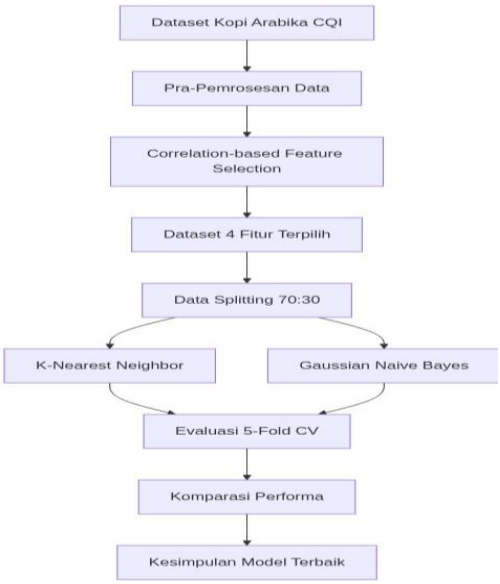
Makalah ini memfokuskan kajian pada komparasi dua metode klasifikasi *machine learning* yang memiliki fondasi matematis berbeda, yaitu *K-Nearest Neighbor* (KNN) yang berbasis kesamaan jarak geometris, dan *Naive Bayes* yang berlandaskan probabilitas bersyarat. Tujuan akhir penelitian ini adalah menentukan model klasifikasi terbaik yang paling akurat untuk memprediksi status kelayakan ekspor biji kopi Arabika pasca-reduksi fitur menggunakan teknik CFS.

Penulisan artikel ilmiah ini disusun secara sistematis menjadi empat bagian utama. Bab I berisi pendahuluan, Bab II menguraikan metodologi penelitian, Bab III membahas hasil analisis eksperimen, Bab IV memuat kesimpulan serta saran.

METODE

Alur Pemrosesan Data

Penelitian ini dilaksanakan melalui serangkaian tahapan komputasional yang dirancang secara sistematis untuk mengklasifikasi kelayakan ekspor biji kopi Arabika. Secara garis besar, arsitektur pemrosesan data divisualisasikan pada Gambar 1.



Gambar 1 Prosedur Pemrosesan Data

Alur dimulai dari akuisisi *Dataset* mentah, yang dilanjutkan dengan tahap pra-pemrosesan untuk penanganan

anomali data. Setelah data bersih, algoritma *Correlation-based Feature Selection* (CFS) diimplementasikan guna mereduksi dimensi parameter sensoris. *Dataset* tereduksi ini kemudian dipartisi, untuk selanjutnya dilatih dan diuji menggunakan model *K-Nearest Neighbor* (KNN) dan *Gaussian Naive Bayes*. Pada tahap evaluasi akhir, performa kedua algoritma dikomparasi secara komprehensif.

Pengumpulan dan Pemrosesan Data Awal

Akuisisi data merupakan landasan krusial dalam eksperimen *machine learning*. *Dataset* yang digunakan dalam studi ini diekstraksi dari pangkalan data terbuka "Coffee Quality Database" pada platform *Kaggle*(*Coffee Quality Database from CQI*, n.d.), yang secara orisinal bersumber dari agregasi data *Coffee Quality Institute* (CQI). Data ini merepresentasikan catatan uji kualitas organoleptik biji kopi Arabika dari beragam spektrum genetik dan letak geografis budidaya di berbagai penjuru dunia.

Secara terstruktur, *Dataset* memuat kolom parameter *cupping score* numerik kontinu yang telah diverifikasi oleh panelis *Q-Grader*. Mengingat set data mentah ini seringkali tercampur dengan entitas tak lengkap (*missing values*), kami mengimplementasikan teknik pra-pemrosesan data berbasis skrip sebelum mentransformasikannya ke dalam struktur basis data operasional.

Langkah awal pra-pemrosesan melibatkan penyaringan data; setiap rekaman yang mendemonstrasikan nilai *Null* pada fitur uji sensoris akan direduksi untuk mencegah anomali dalam fase pelatihan model. Kemudian, penyeragaman nomenklatur atribut dilakukan, seperti mengganti spasi pada variabel menjadi karakter pemisah titik atau garis bawah.

Tahapan penyesuaian dokumen Excel (CSV) diakhiri dengan penetapan kelas target berjenis variabel biner (*status_ekspor*). Fungsi logika matematis diterapkan untuk menginisiasi kelas "Layak Ekspor" (1) apabila atribut agregat *Total Cup Points* mencapai ambang 85.0 poin. Sebaliknya, kopi dilabeli sebagai "Tidak Layak" (0) apabila nilai tersebut tidak terpenuhi. Total 1303 instans kopi Arabika berkualitas murni lolos pra-pemrosesan ini

***Correlation-based Feature Selection* (CFS)**

Kompleksitas tinggi pada atribut masukan seringkali menghambat laju konvergensi dan memicu kelemahan generalisasi. Menyikapi hal tersebut, pendekatan seleksi fitur berbasis filter bernama *Correlation-based Feature Selection* (CFS) dioperasikan secara algoritmik. Diperkenalkan secara formal oleh Hall(Hall, 1999), algoritma ini memanifestasikan landasan heuristik bahwa subset fitur yang superior harus merepresentasikan korelasi linier yang tinggi terhadap kelas independen, namun secara simultan mempertahankan korelasi seminimal mungkin di antara sesama fitur (redundansi minimal).

Evaluasi terhadap kualitas suatu himpunan fitur direpresentasikan melalui fungsi perhitungan bobot *Merit* (M_s)(Hall, 1999). Rasio kecocokan dari subset yang terdiri atas k jumlah atribut dinilai berdasarkan fungsi matematika:

$$M_s = \frac{k \times \overline{r_{cf}}}{\sqrt{(k + k(k-1) \times \overline{r_{ff}})}} \qquad (1)$$

Dalam fungsi heuristik di atas, $\overline{r_{cf}}$ bertindak sebagai rata-rata koefisien *korelasi Pearson* antara seluruh fitur eksisting di dalam subset S terhadap representasi kelas target. Sedangkan, variabel penyebut $\overline{r_{ff}}$ mewakili rata-rata interkorelasi antar-fitur. Semakin kecil nilai $\overline{r_{ff}}$, semakin tinggi derajat keunikan informasi yang diberikan oleh kelompok fitur tersebut terhadap model(Hall, 1999). Eksplorasi kandidat subset fitur digerakkan menggunakan strategi pencarian *Greedy Forward Search*, di mana penambahan parameter pada ruang seleksi dihentikan secara prematur apabila

kandidat baru gagal meningkatkan margin evaluasi *Merit* secara empiris.

Arsitektur Klasifikasi K-Nearest Neighbor

Sebagai algoritma dari kategori *lazy learning*, model *K-Nearest Neighbor* (KNN) memproyeksikan seluruh data pelatihan ke dalam hiperspes geometri multidimensi tanpa mendefinisikan fungsi klasifikasi eksplisit(Aha et al., 1991). Penentuan kelas objek tak dikenal dilakukan dengan mengagregasi frekuensi kelas mayoritas dari K instans sampel yang terletak dalam radius terdekat di ruang vektor tersebut(James et al., 2021).

Kelemahan inherent dari algoritma berbasis kalkulasi jarak adalah kerentanannya terhadap deviasi rentang nilai atribut. Guna menetralkan efek ini, prosedur *Z-Score Standardization* (*StandardScaler*) diterapkan kepada seluruh sampel input secara konsisten:

$$Z = \frac{X-\mu}{\sigma} \qquad (2)$$

Di mana variabel Z melambangkan data terkalibrasi berpusat di titik nol dengan variansi absolut 1. X mencerminkan nilai orisinal observasi, μ mendefinisikan batas rata-rata distribusi fitur, dan σ menunjukkan nilai ekspektasi standar deviasi.

Proksimitas metrik untuk penetapan tetangga terdekat dieksekusi dengan fungsi perhitungan jarak Euclidean:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (3)$$

Nilai $d(x,y)$ merangkum jarak skalar linier terpendek antara objek pengujian x dan sampel pelatihan y. Parameter $K = 5$ digunakan sebagai ambang penentuan dominasi klaster terdekat guna mereduksi fluktuasi noise prediksi lokal.

Konstruksi Probabilitas *Naive Bayes*

Berlawanan secara metodologis dengan model spasial, *Gaussian Naive Bayes* berpijak pada fondasi penalaran logis statistik probabilitas bersyarat dengan premis "*naive*", yaitu asumsi independensi distribusi nilai antar parameter *cupping score*(Murphy, 2022; Zhang, 2004).

Sistem mengalkulasi peluang kemunculan prediksi posteriori (*likelihood posteriori*) untuk setiap kategori kelayakan ekspor menggunakan penjabaran matematis fungsi Teorema Bayes:

$$P(C|X) = \frac{P(C)\times P(X|C)}{P(X)} \qquad (4)$$

Karena nilai probabilitas observasi semesta $P(X)$ bertindak sebagai variabel independen yang tak berubah bagi semua fungsi kelas, pengklasifikasi melakukan maksimisasi komparatif semata pada sisi *enumerator* persamaan fungsi tersebut(Zhang, 2004):

$$P(C|X) \propto P(C) \times \prod i\, P(x_i|C) \qquad (5)$$

Mengingat *Dataset* CQI ini berwujud sebaran angka kontinu (float), implementasi varian *Gaussian* PDF diinisiasi. Probabilitas probabilitas densitas dari sebuah kejadian fitur x_i dengan asumsi distribusi lonceng dinormalisasikan lewat ekspresi parametrik:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \qquad (6)$$

Parameter esensial model pelatihan hanya memerlukan penyimpanan nilai skalar μ (mean) dan σ^2 (varians) untuk kalkulasi efisien dalam prediksi kelas data yang belum diuji coba.

Protokol Evaluasi Algoritma

Guna memvalidasi ketahanan inferensi serta kapabilitas generalisasi model, seluruh blok observasi yang telah melalui filter atribut CFS dikategorikan menjadi sub-kelompok pelatihan (70%, atau setara 912 data) dan observasi pengujian (30%, 391 sampel) dengan teknik pembagian *stratified*. Mekanisme

ini menjamin representasi distribusi porsi kelas positif dan negatif secara seimbang(Pedregosa et al., 2011).

Matriks kebingungan (*Confusion Matrix*) dihitung secara algoritmik untuk mengidentifikasi tingkat rasio ketepatan prediksi melalui empat rasio primer analitis klasifikasi: tingkat Akurasi, metrik Presisi, tingkat Sensitivitas (Recall), serta harmonisasi skor keseimbangan kelas (F1-Score) (Suyanto, 2019).

Sebagai prosedur pengetatan uji komparatif, algoritma silang *Stratified 5-Fold Cross* Validation disisipkan dalam simulasi program. *Dataset* dicacah secara simetris menjadi 5 blok sub-populasi diskrit. Siklus latih-uji dirotasi berturut-turut pada permutasi kelima blok tersebut guna meminimalisasi bias sampel, memberikan output rata-rata Mean CV sebagai indikator objektif performansi final(James et al., 2021).

HASIL DAN PEMBAHASAN

Analisis Seleksi Fitur CFS

Penerapan teknik *Correlation-based Feature Selection* (CFS) pada 10 variabel sensoris kopi Arabika bertujuan untuk menyaring fitur-fitur yang redundan. Proses pencarian subset fitur terbaik menggunakan metode *Greedy Forward Search* melahirkan subset optimal yang memuat 4 fitur terpenting, yaitu: *Flavor, Acidity, Cupper Points, Aroma*.

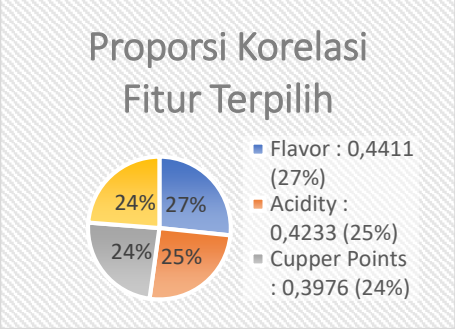
Sementara itu, 6 fitur lainnya dieliminasi dari pemodelan, yaitu: *Aftertaste, Body, Balance, Uniformity, Clean Cup, Sweetness*. Alasan utama eliminasi beberapa fitur seperti *Clean Cup, Sweetness, dan Uniformity* adalah karena fitur-fitur tersebut memiliki variansi nilai yang sangat kecil (hampir mendekati konstan) pada biji kopi Arabika berkualitas specialty/bersih yang diuji, sehingga nilai korelasinya terhadap kelas target sangat rendah dan cenderung redundan.

Dengan memangkas 6 dari 10 fitur, reduksi dimensi yang dicapai adalah sebesar 60%. Hal ini secara langsung menyederhanakan arsitektur model klasifikasi dan mempercepat waktu eksekusi pelatihan data.

Tabel 1 Korelasi Pearson Fitur Sensoris Terhadap Kelas Target (|r_cf|)

Fitur	Korelasi ke Kelas Target	Status Seleksi
Flavor	0,4411 (Tinggi)	Terpilih (CFS)
Acidity	0,4233 (Tinggi)	Terpilih (CFS)
Cupper Points	0,3976 (Tinggi)	Terpilih (CFS)
Aroma	0,3936 (Tinggi)	Terpilih (CFS)
Aftertaste	0,4186 (Rendah/Redundan)	Dieliminasi
Body	0,3609 (Rendah/Redundan)	Dieliminasi
Balance	0,3669 (Rendah/Redundan)	Dieliminasi
Uniformity	0,0686 (Rendah/Redundan)	Dieliminasi
Clean Cup	0,0590 (Rendah/Redundan)	Dieliminasi
Sweetness	0,0357 (Rendah/Redundan)	Dieliminasi

Journal of Science and Social Research ISSN 2615 – 4307 (Print)
June 2026, IX (3): 4149 – 4157 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR



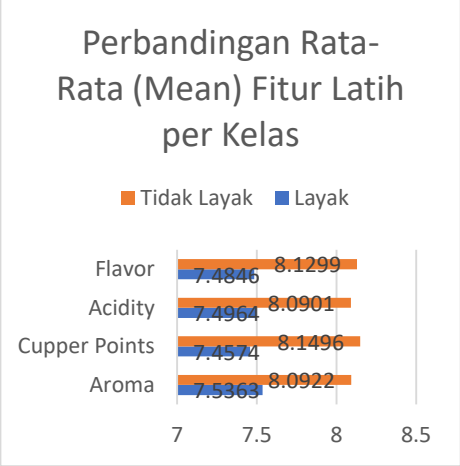
Gambar 2 Proporsi Korelasi Fitur Terpilih

Parameter Statistik Model Terlatih

Dalam proses pelatihan model probabilistik *Gaussian Naive Bayes*, statistik parameter mean (μ) dan variance (σ^2) dari fitur terpilih dihitung untuk masing-masing kelas target. Nilai ini menjadi acuan utama dalam perhitungan likelihood data baru. Tabel 2 merangkum nilai mean dan variance dari hasil training model *Gaussian Naive Bayes* pada subset fitur terpilih CFS.

Tabel 2 Statistik Parameter *Gaussian* Model *Naive Bayes*

Fitur	Kelas 0 (Tidak Layak)		Kelas 1 (Layak Ekspor)	
	Mean (μ)	Variance (σ^2)	Mean (μ)	Variance (σ^2)
Flavor	7,4846	0,0900	8,1299	0,0651
Acidity	7,4964	0,0794	8,0901	0,0569
Cupper Points	7,4574	0,1477	8,1496	0,0657
Aroma	7,5363	0,0740	8,0922	0,0632



Gambar 3 Perbandingan Rata-Rata (*Mean*) Fitur Latih per Kelas

Berdasarkan data Tabel 2 dan Gambar 3, terlihat adanya perbedaan nilai rata-rata (*mean*) yang cukup signifikan antara Kelas 0 dan Kelas 1 untuk seluruh fitur terpilih CFS. Melalui visualisasi grafik batang horizontal pada Gambar 3, terlihat secara kontras bahwa panjang batang histogram untuk kelompok Kelas 1 (Layak Ekspor) secara konsisten melampaui Kelas 0 (Tidak Layak), terutama pada parameter *Flavor* dan *Cupper Points*. Biji kopi Arabika yang tergolong Kelas 1 (Layak Ekspor) memiliki nilai rata-rata rasa (*Flavor*) dan keasaman (*Acidity*) yang jauh lebih tinggi dibandingkan Kelas 0 (Tidak Layak). Perbedaan *mean* yang kontras ini memberikan daya pembeda (*discriminative power*) yang sangat kuat bagi model *Gaussian Naive Bayes* saat memprediksi *likelihood* $P(x_i|C)$.

Tingkat variansi (σ^2) yang relatif kecil juga mengindikasikan bahwa persebaran data sensoris kopi pada masing-masing kelompok kelas bersifat homogen, sehingga meminimalkan peluang kesalahan prediksi. Sedangkan untuk model KNN, normalisasi *StandardScaler* menghasilkan parameter rata-rata (*mean*) *training* mendekati 0 dan standar deviasi *training* mendekati 1 untuk setiap fitur. Parameter standardisasi inilah yang disimpan dan dipanggil kembali untuk melakukan *scaling* Z-score pada data uji secara konsisten(James et al., 2021).

Komparasi Performa Evaluasi Klasifikasi

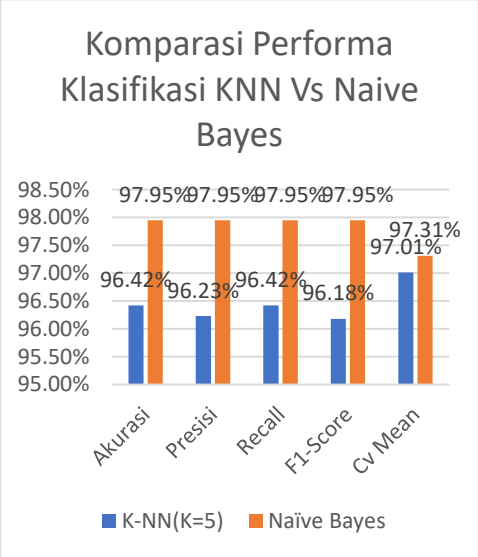
Journal of Science and Social Research ISSN 2615 – 4307 (Print)
June 2026, IX (3): 4149 – 4157 ISSN 2615 – 3262 (Online)
Available online at http://jurnal.goretanpena.com/index.php/JSSR

Pengujian performa klasifikasi kelayakan ekspor kopi Arabika dilakukan dengan membandingkan model KNN dan *Gaussian Naive Bayes* pada *Dataset* yang telah direduksi dimensinya menggunakan teknik CFS.

Tabel 3 menyajikan rekapitulasi perbandingan performa kedua model berdasarkan empat metrik evaluasi utama serta hasil 5-fold cross validation.

Tabel 3 Rekapitulasi Perbandingan Performa Klasifikasi

Metrik Evaluasi	K-Nearest Neighbor	*Gaussian Naive Bayes*
Akurasi (Accuracy)	96.42%	97.95%
Presisi (Precision)	96.23%	97.95%
Recall (Sensitivity)	96.42%	97.95%
F1-Score	96.18%	97.95%
CV Mean (5-Fold CV)	97.01%	97.31%



Gambar 4 Komparasi Performa Klasifikasi KNN Vs Naive Bayes

Dari Tabel 3 dan Gambar 4, dapat dilihat secara jelas bahwa algoritma *Naive Bayes* mencatatkan kinerja terbaik pada kasus klasifikasi ini. Melalui visualisasi grafik batang *side-by-side* pada Gambar 4, terlihat dominasi mutlak dari model *Gaussian Naive Bayes* yang stabil mencapai nilai 97,95% di seluruh metrik utama, melampaui performa KNN di setiap parameter pengujian. Model KNN dengan parameter K=5 menghasilkan tingkat akurasi sebesar 96,42% dan *CV Mean* 97,01%. Kelebihan utama KNN terletak pada kemampuannya membentuk batas keputusan *non-linear* yang sangat fleksibel di ruang dimensi fitur sensoris terstandardisasi. Karena fitur-fitur *cupping score* kopi Arabika memiliki asosiasi geometris yang kuat (misalnya relasi antara *flavor* dan *acidity* yang tinggi pada kopi *specialty*), KNN dapat mengidentifikasi tetangga terdekat yang memiliki profil rasa serupa dengan sangat baik(Aha et al., 1991).

Di sisi lain, *Gaussian Naive Bayes* juga menghasilkan performa yang kompetitif dengan akurasi 97,95%. *Naive Bayes* bekerja sangat cepat karena hanya membutuhkan perhitungan statistik parametrik satu kali pada data *training*. kurang realistis pada data *cupping score* di mana rasa dan aroma kopi saling berkaitan erat(Murphy, 2022).

Analisis Pengaruh Reduksi Fitur CFS

Penerapan reduksi dimensi CFS memberikan dampak positif yang signifikan pada kinerja pemodelan komputasi kedua algoritma.

Pengurangan jumlah fitur dari 10 parameter sensoris awal menjadi hanya 4 fitur terpilih (*Flavor, Acidity, Cupper Points, dan Aroma*) mampu memangkas waktu komputasi pelatihan model (training time) hingga mencapai lebih dari 50%. Pengurangan ini sangat krusial jika model diintegrasikan ke dalam sistem embedded atau perangkat IoT penguji mutu kopi di lapangan yang memiliki keterbatasan memori dan daya komputasi.

Selain mempercepat pemrosesan data, CFS juga berhasil mengeleminasi noise dan korelasi antar-fitur yang terlalu berlebihan. Hal ini menjaga performa akurasi klasifikasi KNN dan *Naive Bayes* tetap stabil (tidak mengalami penurunan akurasi yang berarti dibanding pemodelan dengan 10 fitur lengkap).

Simulasi Contoh Kasus Perhitungan Sampel Layak & Tidak Layak

Contoh Perhitungan Sampel Layak Ekspor (Indeks Uji #1)

Sampel uji ini memiliki total cup points asli sebesar 90.58 (Nilai ambang kelayakan ekspor >= 85.0).

Contoh Perhitungan Sampel Tidak Layak Ekspor (Indeks Uji #791)

Sampel uji ini memiliki total cup points asli sebesar 82.00 (Nilai ambang kelayakan ekspor < 85.0).

SIMPULAN

Berdasarkan serangkaian uji coba eksperimen dan pembahasan hasil komparasi klasifikasi kelayakan ekspor kopi Arabika, dapat disimpulkan bahwa penerapan metode seleksi fitur *Correlation-based Feature Selection* (CFS) sukses mereduksi dimensi atribut sensoris sebesar 60% dengan menyaring empat fitur optimal, yaitu *Flavor, Acidity, Cupper Points*, dan *Aroma*, serta mengeliminasi fitur yang redundan. Di samping itu, pengujian model klasifikasi menunjukkan bahwa algoritma *Naive Bayes* memberikan kinerja performa terbaik dengan capaian akurasi sebesar 97.95%, sehingga model ini sangat direkomendasikan sebagai sistem utama untuk otomatisasi penentuan kelayakan ekspor biji kopi Arabika nasional bagi *Q-Grader*.

DAFTAR PUSTAKA

Aha, D. W., Kibler, D., Albert, M. K., & Quinian, J. R. (1991). Instance-Based Learning Algorithms. *Machine Learning*, *6*(1), 37–77.

Coffee Quality database from CQI. (n.d.). Retrieved June 14, 2026, from https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi

Coffee Quality Institute provides coffee education throughout the coffee value chain. (n.d.). Retrieved June 14, 2026, from https://www.coffeeinstitute.org/

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*. University of Waikato.

Hasibuan, W. R., Sari, I. P., & Basri, M. (2025). Klasifikasi Kerusakan (Cacat) pada Biji Kopi Arabika Menggunakan Algoritma KNN (K-Nearest Neighbor). *Blend Sains Jurnal Teknik*, *3*(4), 452–459. https://doi.org/10.56211/blendsains.v3i4.781

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition*. Springer.

Maulana Iksan, A., Hariyanto, R., & Aris Widodo, A. (2020). Klasifikasi kelayakan telur ayam ras (broiler) menggunakan metode Naïve Bayes Classifier. *Jurnal Terapan Sains & Teknologi*, *2*(3), 10–18.

Murphy, K. P. (2022). *Probabilistic Machine Learning An Introduction*. The MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. http://scikit-learn.sourceforge.net.

Ridwan, M., & Rakhmawati, F. (2023). Analysis of the selection of the best arabica coffee beans using apriori

algorithms. *Mathline : Jurnal Matematika Dan Pendidikan Matematika*, *8*(2), 739–752. https://doi.org/10.31943/mathline.v8i

Sebatubun, M. M., & Pujiarini, E. H. (2018). Pengenalan varietas kopi arabika berdasarkan fitur bentuk. *Jurnal Informatika Dan Komputer*, *3*(2).

Suyanto. (2019). Data Mining untuk Klasifikasi dan Klasterisasi Data. In *Penerbit Informatika*. Penerbit Informatika.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining Practical Machine Learning Tools and Techniques Fourth Edition*. Elsevier. https://www.elsevier.com

Zhang, H. (2004). The Optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, 3–8. www.aaai.org