Evaluasi Information Gain pada Random Forest, Decision Tree, dan KNN Klasifikasi Tanaman

Dina Andayani[1], Suraya[2],Muhammad Sholeh*[3],Suparyanto[4]
[1]*Program Studi Manajemen Retail, Universitas AKPRIND Indonesia*
[2]*Program Studi Rekayasa Sistem Komputer, Universitas AKPRIND Indonesia*
[3]*Program Studi Informatika, Universitas AKPRIND Indonesia*
[4]*Program Studi Informatika, Sekolah Tinggi Ilmu Komputer EL RAHMA Yogyakarta*
e-mail: [1]dina_asnawi@yahoo.com, [2]suraya@akprind.ac.id, [3]muhash@akprind.ac.id,
[3]anto_suparyanto@stmikelrahma.ac.id
**Correspondence author*

Abstrak

Pemilihan jenis tanaman yang sesuai dengan kondisi tanah dan lingkungan berperan penting dalam meningkatkan produktivitas pertanian. Perkembangan machine learning memungkinkan klasifikasi tanaman berdasarkan karakteristik tanah dan iklim. Penggunaan seluruh fitur dapat meningkatkan kompleksitas model, sementara tidak semua fitur memberikan kontribusi informasi yang sama. Penelitian ini bertujuan menganalisis pengaruh seleksi fitur menggunakan Information Gain terhadap kinerja Random Forest, Decision Tree, dan K-Nearest Neighbors dalam klasifikasi jenis tanaman. Dataset yang digunakan adalah crop recommendation yang terdiri atas 2.200 sampel, tujuh fitur dan 22 kelas tanaman. Seleksi fitur dilakukan menggunakan mutual_info_classif dengan ambang batas >1,0 dan menghasilkan enam fitur terpilih, yaitu humidity, K, rainfall, P, temperature, dan N, sedangkan pH dikeluarkan. Evaluasi model dilakukan menggunakan Stratified 10-Fold Cross-Validation dengan metrik accuracy, precision, recall, dan F1-score. Hasil pengujian menunjukkan bahwa sebelum seleksi fitur, Random Forest, Decision Tree, dan K-Nearest Neighbors memperoleh akurasi masing-masing sebesar 99,32%, 98,64%, dan 97,05%. Setelah seleksi fitur, akurasi masing-masing menjadi 99,54%, 98,64%, dan 98,18%. Pengurangan fitur pH tidak menurunkan kinerja Decision Tree, sementara Random Forest meningkat 0,22 poin persentase dan K-Nearest Neighbors meningkat 1,13 poin persentase. Hasil tersebut menunjukkan bahwa enam fitur terpilih mampu mempertahankan kinerja klasifikasi yang tinggi dengan jumlah fitur yang lebih ringkas.

Kata kunci *:machine learning, Information Gain, seleksi fitur, Random Forest, Decision Tree, K-Nearest Neighbors, klasifikasi tanaman.*

1. PENDAHULUAN

Pertanian merupakan salah satu sektor penting dalam mendukung penyediaan pangan, perekonomian, dan keberlanjutan sumber daya. Salah satu keputusan penting dalam kegiatan budidaya adalah menentukan jenis tanaman yang sesuai dengan karakteristik lahan dan kondisi lingkungan [1],[2]. Ketidak sesuaian antara jenis tanaman dengan kondisi tanah maupun iklim dapat menyebabkan penggunaan sumber daya menjadi kurang efisien dan berpotensi menurunkan produktivitas. Pemilihan tanaman perlu mempertimbangkan berbagai parameter yang menggambarkan kondisi lahan dan lingkungan secara lebih objektif [3], [4], [5].

Pemilihan tanaman secara konvensional pada banyak kasus masih bergantung pada pengalaman petani, pengetahuan lokal, serta pola budidaya yang telah dilakukan sebelumnya. Pendekatan tersebut memiliki nilai penting karena berasal dari pengalaman empiris, tetapi belum selalu mampu mengintegrasikan berbagai parameter lingkungan secara simultan. Perkembangan teknologi digital kemudian membuka peluang untuk mendukung proses pengambilan keputusan pertanian melalui pendekatan berbasis data. Pemanfaatan teknologi digital dalam pertanian dapat membantu meningkatkan efisiensi pengelolaan sumber daya serta mendukung pengambilan keputusan berbasis informasi [6]. Perkembangan tersebut selanjutnya mendorong penerapan *machine learning* dalam berbagai permasalahan pertanian.

Machine learning memungkinkan sistem mempelajari pola hubungan antara kondisi lahan, lingkungan, dan kelas tanaman berdasarkan data historis. Dalam konteks rekomendasi tanaman, beberapa penelitian menunjukkan bahwa parameter unsur hara tanah seperti nitrogen (N), fosfor (P), dan kalium (K), serta parameter lingkungan seperti temperatur, kelembapan, pH, dan curah hujan dapat digunakan sebagai masukan untuk membangun model klasifikasi tanaman. Dey et al. (2024), misalnya, mengevaluasi beberapa algoritma *machine learning* untuk rekomendasi tanaman menggunakan parameter N, P, K, pH, temperatur, rainfall, dan humidity. Penelitian lain juga menunjukkan bahwa kombinasi karakteristik tanah dan kondisi iklim dapat digunakan untuk membangun sistem rekomendasi tanaman berbasis *machine learning*.

Penggunaan *machine learning* untuk rekomendasi tanaman terus berkembang. Hamdani [7], melakukan penelitian dengan model Machine Learning dengan melakukan analisis kelembapan tanah, suhu, dan kelembapan udara untuk memprediksi kebutuhan air tanaman secara real-time. Integrasi IoT dan Machine Learning memungkinkan monitoring lahan berkelanjutan serta penentuan waktu dan jumlah irigasi yang lebih tepat daripada metode konvensional. Sistem ini mendukung pengambilan keputusan irigasi secara akurat dan efisien, sehingga berpotensi meningkatkan produktivitas pertanian dan mengoptimalkan penggunaan air. Penelitian selanjutnya dapat mengembangkan model yang lebih kompleks dengan memanfaatkan data lingkungan yang lebih beragam. Penelitian Adiya [8], mengembangkan sistem prediksi jenis tanaman berdasarkan parameter tanah (N, P, K, suhu, kelembapan, pH) dengan membandingkan GBM dan *LightGBM*. Dataset dari Kaggle diproses melalui normalisasi, label encoding, dan hyperparameter tuning menggunakan *RandomizedSearch*. Hasil menunjukkan akurasi GBM 96,14% dan *LightGBM* 96,82%. Metode *boosting* terbukti efektif untuk rekomendasi tanaman berbasis kondisi tanah dan mendukung pertanian presisi berkelanjutan. Penelitian lain dengan menggunakan dataset 2.200 sampel dengan 22 jenis tanaman dan membandingkan beberapa algoritma, termasuk *Random Forest, Decision Tree*, dan *Logistic Regression*. Hasil penelitian tersebut memperlihatkan bahwa pendekatan *ensemble* dapat memberikan performa yang tinggi dalam rekomendasi tanaman [9].

Hasil [dari beberapa penelitian, penggunaan berbagai parameter dapat meningkatkan informasi yang tersedia bagi model, tidak seluruh fitur selalu memberikan kontribusi yang sama terhadap proses klasifikasi [10]. Fitur yang kurang informatif atau memiliki kontribusi rendah dapat menambah kompleksitas representasi data tanpa memberikan peningkatan performa yang sebanding. Kondisi tersebut menjadikan seleksi fitur sebagai salah satu tahapan penting dalam pengembangan model *machine learning*. Seleksi fitur bertujuan mempertahankan fitur yang paling relevan terhadap variabel target sekaligus mengurangi fitur yang kurang memberikan informasi. Dengan jumlah fitur yang lebih ringkas, model berpotensi memiliki representasi data yang lebih sederhana dan lebih mudah diinterpretasikan [11].

Salah satu metode *filter* yang dapat digunakan untuk seleksi fitur adalah *Information Gain*. Metode ini mengukur seberapa besar suatu fitur memberikan informasi terhadap variabel target melalui pengurangan ketidakpastian atau entropi [12],[13]. Sholeh [12], hasil penelitian melakukan perbandingan dalam penerapan seleksi fitur berbasis *Information Gain* pada *Decision Tree, Logistic Regression,* dan *Random Forest* menggunakan dataset kanker paru (±306 sampel, 16 atribut, dua kelas) dan melaporkan akurasi tertinggi 96,8% pada *Logistic Regression* dengan lima fitur. Namun, studi tersebut menggunakan data berukuran kecil dengan klasifikasi biner, mengevaluasi model melalui satu kali pembagian data tanpa validasi silang maupun uji signifikansi, sehingga selisih akurasi antar skenario hanya setara sekitar satu sampel uji.

Semakin besar nilai *Information Gain,* semakin besar kontribusi informasi fitur tersebut dalam membedakan kelas. Pendekatan berbasis teori informasi telah banyak digunakan untuk mengidentifikasi fitur yang relevan pada dataset berdimensi tinggi . Penelitian [14], mengeksplorasi metode *K-Nearest Neighbors* (K-NN) untuk prediksi diabetes dengan Information Gain sebagai seleksi fitur guna memilih atribut paling informatif, meningkatkan akurasi, dan mengurangi dimensi data yang tidak relevan. Pengujian dilakukan pada empat rasio dataset (90%:10%, 80%:20%, 70%:30%, 60%:40%). Hasil menunjukkan Information Gain secara konsisten meningkatkan akurasi K-NN. Pada rasio 90%:10%, akurasi meningkat dari

69,11% (tanpa *Information Gain*) menjadi 70,96% (dengan *Information Gain*), dan mencapai akurasi tertinggi 72,93% pada K = 17.

Penelitian yang terkait dengan klasifikasi tanaman, penelitian mengenai seleksi fitur menjadi semakin relevan karena parameter tanah dan iklim memiliki karakteristik yang berbeda serta tidak selalu memberikan kontribusi yang sama terhadap setiap jenis tanaman. Parameter N, P, K, temperatur, kelembapan, pH, dan curah hujan dapat memiliki hubungan yang berbeda terhadap kelas tanaman. Penggunaan seluruh fitur belum tentu menghasilkan model yang lebih baik dibandingkan penggunaan subset fitur yang lebih informatif. Penelitian mengenai *feature selection* pada *crop recommendation* juga menunjukkan bahwa pemilihan fitur perlu dievaluasi bersamaan dengan algoritma klasifikasi yang digunakan [15], [16].

Penelitian sebelumnya yang menjadi salah satu dasar penelitian ini telah membandingkan penerapan *Information Gain* pada beberapa algoritma klasifikasi, termasuk *Decision Tree, Logistic Regression* dan *Random Forest* [17], [18]. Penelitian ini diarahkan secara lebih spesifik pada analisis perubahan kinerja *Random Forest, Decision Tree* dan *K-Nearest Neighbors* sebelum dan sesudah pengurangan fitur pada klasifikasi jenis tanaman. Fokus tersebut penting karena setiap algoritma memiliki karakteristik yang berbeda dalam merespons perubahan jumlah dan komposisi fitur. *Decision Tree* membangun aturan keputusan melalui proses pemisahan fitur, *Random Forest* menggabungkan sejumlah pohon keputusan, sedangkan *K-Nearest Neighbors* menentukan kelas berdasarkan kedekatan data dengan sejumlah tetangga terdekat.

Berdasarkan studi terdahulu tersebut, masih terdapat ruang untuk menganalisis secara lebih terarah bagaimana seleksi fitur *Information Gain* memengaruhi performa beberapa algoritma klasifikasi dengan karakteristik yang berbeda pada data rekomendasi tanaman. Penelitian ini tidak hanya berorientasi pada pencapaian akurasi tertinggi, tetapi juga membandingkan perubahan performa sebelum dan sesudah seleksi fitur. Pendekatan tersebut memungkinkan identifikasi apakah fitur yang dikurangi benar-benar memberikan kontribusi penting terhadap model atau justru dapat dihilangkan tanpa menyebabkan penurunan performa yang berarti. Penelitian ini menempatkan seleksi fitur tidak hanya sebagai tahap praprosesing, tetapi sebagai bagian penting dalam evaluasi efisiensi dan efektivitas model klasifikasi.

Berdasarkan permasalahan tersebut, penelitian ini bertujuan untuk menganalisis pengaruh seleksi fitur menggunakan *Information Gain* terhadap kinerja *Random Forest, Decision Tree* dan *K-Nearest Neighbors* dalam klasifikasi jenis tanaman. Dataset yang digunakan terdiri atas 2.200 sampel dengan tujuh parameter kondisi tanah dan lingkungan serta 22 kelas tanaman. Penelitian dilakukan dengan membandingkan performa ketiga model menggunakan seluruh fitur dan fitur hasil seleksi *Information Gain*. Kontribusi penelitian ini adalah sebagai berikut. (1) Membandingkan kinerja *Decision Tree, K-Nearest Neighbors* dan *Random Forest* sebelum dan sesudah seleksi fitur *Information Gain* pada data rekomendasi tanaman dengan 22 kelas. (2) Mengevaluasi kinerja model menggunakan *10-fold cross-validation* sehingga hasil evaluasi tidak bergantung pada satu kali pembagian data dan kinerja model dapat dilaporkan berdasarkan rata-rata hasil pengujian. (3) Menganalisis perubahan akurasi model setelah seleksi fitur untuk mengetahui dampak pengurangan fitur terhadap kinerja masing-masing algoritma.

2. METODE PENELITIAN

Dataset

Dataset yang digunakan adalah farming.csv. Dataset diperoleh dari https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset. Dataset merupakan data publik yang memuat kondisi tanah dan lingkungan untuk klasifikasi jenis tanaman. Dataset terdiri atas 2.200 sampel dengan tujuh fitur input dan satu variabel target. Terdapat 22 jenis tanaman yang masing-masing memiliki 100 sampel sehingga distribusi kelas relatif seimbang. Rincian dari dataset disajikan pada tabel 1.

Berdasar pada Tabel 1, dataset dilakukan proses pengecekan dataset dan proses membersihkan dataset. Dataset tidak memiliki nilai hilang berdasarkan pemeriksaan awal.

Distribusi kelas yang seimbang juga mengurangi potensi permasalahan *class imbalance* dalam proses klasifikasi.

Tabel 1. Informasi dari dataset *crop recommendation*

Fitur	Keterangan	Satuan
N	Kandungan nitrogen tanah	ppm
P	Kandungan fosfor tanah	ppm
K	Kandungan kalium tanah	ppm
Temperature	Suhu udara	°C
Humidity	Kelembapan udara	%
pH	Derajat keasaman tanah	-
Rainfall	Curah hujan	mm
Label	Jenis tanaman	22 kelas

Tahapan Penelitian

Penelitian dilakukan melalui beberapa tahapan yang dimulai dengan pengumpulan dan pemeriksaan dataset, kemudian dilanjutkan dengan eksplorasi data untuk memahami karakteristik data yang digunakan. Selanjutnya, disiapkan skema validasi menggunakan *Stratified 10-Fold Cross-Validation* untuk menjaga proporsi kelas pada setiap *fold*. Pemodelan awal dilakukan menggunakan seluruh fitur sebagai *baseline*, kemudian dilakukan perhitungan *Information Gain* untuk menentukan tingkat relevansi masing-masing fitur terhadap kelas target. Berdasarkan nilai *Information Gain*, fitur-fitur yang memenuhi kriteria seleksi digunakan untuk membangun kembali model klasifikasi. Kinerja model sebelum dan sesudah seleksi fitur selanjutnya dievaluasi dan dibandingkan untuk mengetahui pengaruh seleksi fitur terhadap performa masing-masing algoritma. Tahap akhir penelitian adalah implementasi model terbaik ke dalam aplikasi rekomendasi tanaman.

Seleksi Fitur Information Gain

Information Gain digunakan untuk menentukan tingkat kontribusi masing-masing fitur terhadap kelas tanaman. Secara umum, *Information Gain* dihitung berdasarkan pengurangan entropi setelah suatu fitur digunakan untuk mempartisi data. Information Gain dihitung menggunakan mutual_info_classif pada data fitur dan label. Nilai yang diperoleh digunakan untuk mengurutkan kontribusi informasi setiap fitur. Berdasarkan hasil notebook, skor Information Gain adalah: humidity = 1,729954; K = 1,639248; rainfall = 1,637358; P = 1,304026; temperature = 1,017901; N = 1,000392; dan pH = 0,686067.

Pemodelan

Penelitian ini menggunakan tiga algoritma klasifikasi, yaitu *Decision Tree*, *K-Nearest Neighbors* (KNN), dan *Random Forest*. *Decision Tree* digunakan karena mampu membentuk aturan keputusan berbasis struktur pohon dan relatif mudah diinterpretasikan. Implementasi algoritma ini menggunakan DecisionTreeClassifier dengan random_state=42 dan konfigurasi *default* Scikit-learn, termasuk penggunaan kriteria *Gini impurity*. KNN digunakan sebagai metode klasifikasi berbasis kedekatan antarobservasi dan diimplementasikan menggunakan KNeighborsClassifier dengan n_neighbors=5. Sementara itu, *Random Forest* digunakan sebagai metode *ensemble* yang menggabungkan sejumlah pohon keputusan untuk menghasilkan prediksi dan diimplementasikan menggunakan RandomForestClassifier dengan n_estimators=100 dan random_state=42. Ketiga algoritma tersebut dievaluasi pada dua kondisi, yaitu menggunakan seluruh tujuh fitur dan menggunakan enam fitur yang diperoleh dari hasil seleksi *Information Gain*.

Evaluasi

Kinerja model dievaluasi menggunakan *accuracy*, *precision*, *recall*, dan *F1-score*. Evaluasi dilakukan menggunakan *Stratified K-Fold Cross-Validation* dengan 10 fold untuk

memperoleh estimasi kinerja model yang lebih representatif dengan mempertahankan proporsi kelas pada setiap fold.

Pada algoritma K-Nearest Neighbors (KNN), standardisasi fitur dilakukan menggunakan StandardScaler karena algoritma ini berbasis jarak dan sensitif terhadap perbedaan skala antar fitur. Parameter standardisasi dihitung berdasarkan data latih pada setiap fold, kemudian diterapkan pada data validasi untuk mencegah data leakage. Prosedur evaluasi yang sama diterapkan pada model dengan seluruh fitur dan model dengan fitur hasil seleksi Information Gain. Kinerja kedua model kemudian dibandingkan berdasarkan nilai accuracy, precision, recall, dan F1-score.

3. HASIL DAN PEMBAHASAN

Eksplorasi Dataset

Hasil eksplorasi menunjukkan bahwa dataset memiliki variasi nilai yang cukup besar pada parameter N, P, dan K. Temperatur memiliki nilai rata-rata sekitar 25,62°C, sedangkan kelembapan memiliki rentang yang relatif lebar. Nilai pH mencakup kondisi tanah asam hingga basa, sementara curah hujan memiliki variasi yang cukup tinggi.

Pemeriksaan data menunjukkan tidak terdapat *missing value*. Distribusi kelas juga seimbang karena masing-masing dari 22 kelas tanaman memiliki 100 sampel. Dataset yang seimbang memberikan keuntungan dalam proses evaluasi karena kinerja model tidak terlalu dipengaruhi oleh dominasi kelas tertentu.

Kinerja Model Sebelum Seleksi Fitur

Tahap pengujian awal dilakukan dengan menggunakan seluruh fitur yang tersedia dalam dataset, yaitu nitrogen (N), fosfor (P), kalium (K), temperatur, kelembapan (humidity), pH, dan curah hujan (rainfall). Pengujian ini dilakukan sebagai baseline untuk mengetahui kemampuan awal masing-masing algoritma dalam melakukan klasifikasi jenis tanaman sebelum diterapkan seleksi fitur menggunakan Information Gain. Tiga algoritma yang digunakan, yaitu *Random Forest, Decision Tree* dan *K-Nearest Neighbors* (KNN), diuji menggunakan Stratified 10-Fold Cross-Validation. Hasil pengujian awal tersebut kemudian digunakan sebagai dasar pembanding untuk melihat perubahan kinerja model setelah jumlah fitur dikurangi. Hasil pengujian model menggunakan seluruh fitur disajikan pada tabel 2.

Tabel 2. Kinerja Model Sebelum Seleksi Fitur

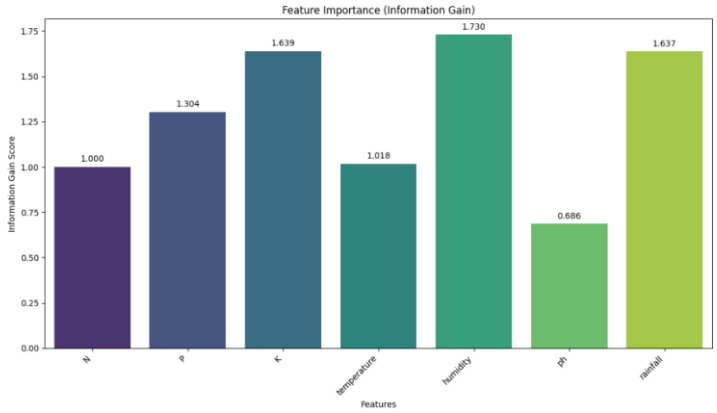
Model	Akurasi (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	99,32	99,37	99,32	99,32
Decision Tree	98,64	98,68	98,64	98,63
K-Nearest Neighbors	97,05	97,40	98,05	98,03

Berdasar pada tabel 2, *Random Forest* menghasilkan akurasi tertinggi sebesar 99,32%. Hasil tersebut menunjukkan kemampuan *Random Forest* dalam menangkap hubungan kompleks dan interaksi antarfitur melalui mekanisme *ensemble*. *Decision Tree* memperoleh akurasi 98,64%. Nilai tersebut menunjukkan bahwa struktur pohon keputusan mampu memisahkan 22 kelas tanaman dengan baik berdasarkan karakteristik tanah dan lingkungan. *K-Nearest Neighbors* menghasilkan akurasi sebesar 97,05%. Nilai tersebut menunjukkan bahwa pendekatan berbasis kedekatan masih mampu mengklasifikasikan sebagian besar kelas tanaman, meskipun kinerjanya berada di bawah dua model berbasis pohon.

Pada KNN proses pada tabel 2 masih menggunakan dataset asli dan sebagai pembanding dilakukan standardisasi fitur dengan menggunakan StandardScaler karena algoritma ini berbasis jarak dan sensitif terhadap perbedaan skala antarfitur. Hasil standardisasi menghasilkan akurasi 95,68%, precision 96,29%, recall 95,68%, dan F1-score 95,67%. Oleh karena itu, hasil KNN yang digunakan dalam perbandingan harus ditetapkan secara konsisten berdasarkan prosedur evaluasi yang sama.

Hasil Seleksi Fitur

Seleksi fitur dilakukan menggunakan *Information Gain*. Hasil nilai *Information Gain* disajikan pada gambar 1.



Gambar 1 Hasil *Information Gain* dari dataset *crop recommendation*

Berdasar pada gambar 1, Hasil *information Gain* menunjukkan bahwa *humidity* (1,730), K (1,639), dan *rainfall* (1,637) merupakan fitur paling berpengaruh, diikuti P (1,304), *temperature* (1,018), dan N (1,000). Fitur pH memiliki skor terendah (0,686) sehingga tidak digunakan dalam proses pembuatan model. Ambang batas 1 dipilih sebagai batas kontribusi minimum, yaitu fitur yang skornya kurang dari 1 dianggap memberikan informasi yang relatif kecil dalam membedakan kelas target.

Kinerja Model Setelah Seleksi Fitur

Berdasarkan hasil seleksi fitur menggunakan *Information Gain*, diperoleh enam fitur terpilih yang selanjutnya digunakan dalam proses pemodelan. Hasil pengujian kinerja model menggunakan enam fitur terpilih ditunjukkan pada Tabel 3.

Tabel 3. Kinerja Model Setelah Seleksi Fitur

Model	Akurasi (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	99,54	99,61	99,55	99,54
Decision Tree	98,64	98,85	98,83	98,64
K-Nearest Neighbors	98,18	98,41	98,18	98,18

Berdasarkan Tabel 2 dan Tabel 3, seleksi fitur dengan menghilangkan fitur pH memberikan perubahan kinerja pada ketiga model. *Random Forest* mengalami peningkatan akurasi dari 99,32% menjadi 99,54% atau sebesar 0,22 poin persentase. *Decision Tree* mempertahankan akurasi sebesar 98,64%, sedangkan *K-Nearest Neighbors* meningkat dari 97,05% menjadi 98,18% atau sebesar 1,13 poin persentase. Perubahan pada *precision, recall* dan *F1-score* juga relatif kecil. Hasil ini menunjukkan bahwa penggunaan enam fitur terpilih tetap mampu mempertahankan kinerja klasifikasi yang tinggi.

Perbandingan Sebelum dan Sesudah Seleksi Fitur

Perbandingan kinerja model sebelum dan sesudah seleksi fitur dilakukan untuk mengetahui pengaruh penerapan *Information Gain* terhadap kemampuan algoritma dalam melakukan klasifikasi jenis tanaman. Perbandingan dilakukan berdasarkan nilai akurasi dari tiga algoritma yang digunakan, yaitu *Random Forest, Decision Tree* dan *K-Nearest Neighbors* (KNN). Hasil perbandingan tersebut digunakan untuk melihat apakah pengurangan jumlah fitur dapat mempertahankan atau meningkatkan kinerja model. Perbandingan akurasi model sebelum dan sesudah seleksi fitur disajikan pada Tabel 4.

Tabel 4. Perbandingan Akurasi Model

Model	Accuracy (Sebelum Seleksi)	Accuracy (Setelah Seleksi)	Perubahan (poin persentase)
Random Forest	99,32%	99,54%	0,22
Decision Tree	98,64%	98,64%	0,00
K-Nearest Neighbors	97,05%	98,18%	1,13

Berdasarkan Tabel 3 dan Tabel 4, seleksi fitur dengan menghilangkan fitur pH memberikan perubahan kinerja pada ketiga model. *Random Forest* mengalami peningkatan akurasi dari 99,32% menjadi 99,54%, *Decision Tree* tetap sebesar 98,64%, sedangkan *K-Nearest Neighbors* meningkat dari 97,05% menjadi 98,18%. Setelah seleksi fitur, *Random Forest* menghasilkan akurasi tertinggi, diikuti *Decision Tree* dan *K-Nearest Neighbors*. Hasil ini menunjukkan bahwa enam fitur terpilih tetap mampu menghasilkan kinerja klasifikasi yang tinggi. Perbedaan perubahan kinerja menunjukkan bahwa setiap algoritma memberikan respons yang berbeda terhadap pengurangan fitur. Pemilihan model selanjutnya dapat mempertimbangkan kinerja, interpretabilitas, dan karakteristik algoritma.

Analisis Fitur

Hasil *Information Gain* menunjukkan bahwa *humidity*, K, dan *rainfall* memiliki skor tertinggi, masing-masing sebesar 1,729954, 1,639248, dan 1,637358. Berdasarkan *threshold* >1,0, enam fitur dipertahankan, yaitu *humidity*, K, *rainfall*, P, *temperature*, dan N, sedangkan pH dengan skor 0,686067 dikeluarkan. Hasil ini menunjukkan bahwa *humidity*, K, dan *rainfall* memiliki nilai informasi relatif lebih tinggi dalam membedakan kelas tanaman pada dataset yang digunakan.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa seleksi fitur menggunakan Information Gain memberikan perubahan kinerja pada *Random Forest, Decision Tree* dan *K-Nearest Neighbors*. Setelah enam fitur terpilih digunakan, *Random Forest* memperoleh akurasi sebesar 99,54%, *Decision Tree* sebesar 98,64%, dan *K-Nearest Neighbors* sebesar 98,18%.

Information Gain dengan threshold >1,0 memilih enam fitur, yaitu *humidity*, K, *rainfall*, P, *temperature* dan N, serta mengeluarkan pH. Setelah seleksi fitur, akurasi *Random Forest* meningkat dari 99,32% menjadi 99,54% atau sebesar 0,22 poin persentase. *Decision Tree* mempertahankan akurasi sebesar 98,64%, sedangkan *K-Nearest Neighbors* meningkat dari 97,05% menjadi 98,18% atau sebesar 1,13 poin persentase. Hasil tersebut menunjukkan bahwa pengurangan fitur pH tidak menurunkan kinerja *Decision Tree* dan justru meningkatkan kinerja *Random Forest* dan *K-Nearest Neighbors* pada dataset yang digunakan.

Secara keseluruhan, enam fitur terpilih mampu mempertahankan kinerja klasifikasi yang tinggi dengan jumlah fitur yang lebih ringkas. Pemilihan model selanjutnya dapat mempertimbangkan nilai kinerja, karakteristik algoritma, dan kebutuhan implementasi, seperti interpretabilitas pada *Decision Tree* atau pendekatan *ensemble* pada *Random Forest*.

5. SARAN

Penelitian ini menggunakan dataset publik dengan 2.200 sampel sehingga hasilnya belum sepenuhnya menggambarkan kemampuan generalisasi model pada kondisi lahan nyata. Penelitian selanjutnya dapat memperluas evaluasi menggunakan dataset dari lokasi pertanian nyata dan membandingkan *Information Gain, Gain Ratio* dan *Chi-Square* dengan evaluasi yang sama. Validasi menggunakan data lapangan dari berbagai kondisi lingkungan juga diperlukan untuk mengetahui konsistensi kemampuan model dalam memberikan rekomendasi tanaman.

6. UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih kepada Direktorat Penelitian, Pengembangan, dan Pengabdian kepada Masyarakat (DP2M) Universitas AKPRIND Indonesia atas dukungan dan fasilitas yang diberikan sehingga penelitian ini dapat dilaksanakan dengan baik. Penulis juga mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan selama proses penelitian, pengolahan data, hingga penyusunan artikel ini.

DAFTAR PUSTAKA

[1] Y. Kamakaula, "Optimasi Pertanian Berkelanjutan : Pengabdian Masyarakat untuk Peningkatan Produktivitas," *Communnity Development Journal*, vol. 4, no. 6, pp. 11463–11471, 2023.

[2] M. Sihite, A. M. Hsb, R. Syahputra, M. R. Amri, R. Alwi, and Sakuntala, "Peran Sektor Pertanian dan Distribusi Pendapatan di Indonesia : Analisis Model Faktor Spesifik," *JURNAL MEDIA AKADEMIK (JMA)*, vol. 3, no. 1, 2025.

[3] N. Fajeriana, "Kesesuaian Lahan dan Kesuburan Tanah pada Lahan Budidaya Kacang Tanah (Arachis hypogaea) di Kampung Kofalit Distrik Salkma Kabupaten Sorong Selatan," *Agroteknika*, vol. 7, no. 1, pp. 51–66, 2024.

[4] R. Manurung, J. Gunawan, R. Hazriani, and J. Suharmoko, "Pemetaan Status Unsur Hara N, P Dan K Tanah pada Perkebunan Kelapa Sawit di Lahan Gambut," *Jurnal Pedon Tropika Edisi*, vol. 3, no. 1, 2020.

[5] A. N. Sandil, M. Montolalu, and R. I. Kawulusan, "Kajian Sifat Kimia Tanah Pada Lahan Berlereng Tanaman Cengkeh (Syzygium Aromaticum L) dI Salurang Kecamatan Tabukan Selatan Tengah," *Soil Environmental*, vol. 21, no. 3, pp. 18–23, 2021.

[6] R. Abiri, N. Rizan, S. K. Balasundram, A. Bayat, and H. Abdul-hamid, "Heliyon Application of digital technologies for ensuring agricultural productivity," *Heliyon*, vol. 9, no. 12, p. e22601, 2023, doi: 10.1016/j.heliyon.2023.e22601.

[7] A. Hamdani, "Implementasi Machine Learning untuk Prediksi Kebutuhan Irigasi Tanaman pada Sistem Smart Agriculture Berbasis Internet of Things," *Karapan Network Journal*, vol. 02, no. 03, 2026.

[8] R. H. Aditya, "Implementasi Metode Boosting Untuk Prediksi Jenis Tanaman Berdasarkan Kondisi Tanah," *135 Jurnal Inovasi Komputer (INOKOM)*, vol. 1, no. 3, pp. 135–145, 2025.

[9] P. N. Sabrina and A. Komarudin, "Prediksi Penyakit Diabetes dengan Metode K-Nearest Neighbor (KNN) dan Seleksi Fitur Information Gain," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 8, no. 6, pp. 11320–11326, 2024.

[10] G. Manikandan and A. Murugappan, "An efficient feature selection framework based on information theory for high dimensional data," *Applied Soft Computing*, vol. 111, p. 107729, Jul. 2021, doi: 10.1016/j.asoc.2021.107729.

[11] M. Chiregi, M. Mazinani, and M. Mirzarezaee, "A New Feature Selection Method Using Deep Learning and Graph Representation in High-Dimensional Datasets," *Knowledge-Based Systems*, vol. 329, p. 114338, Aug. 2025, doi: 10.1016/j.knosys.2025.114338.

[12] M. Sholeh, U. Lestari, and D. Andayati, "Comparison of Feature Selection with Information Gain Method in Decision Tree , Regression Logistic and Random Forest," *Journal of Applied Business and Technology*, vol. 5, no. 3, pp. 146–153, 2024.

[13] A. H. Mohammad, "Comparing Two Feature Selections Methods (Information Gain and Gain Ratio) on Three Different Classification Algorithms using Arabic Dataset," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 6, 2018.

[14] R. Annisa, R. Amelia, M. Putri, and C. Febriyani, "Pengembangan Model Klasifikasi Citra

Tanaman Hutan Melicopelatifolia Berbasis CNN dengan Custom-Built Dataset," *JUKI : Jurnal Komputer dan Informatika*, vol. 6, pp. 174–181, 2024.

[15] M. Abdullah and I. Fahrurrozi, "Model Comparison and Feature Selection for Crop Recommendation," *Jurnal Pepadun*, vol. 7, no. 1, pp. 16–25, 2026, doi: 10.23960/pepadun.v7i3.345.

[16] A. Aryanti and N. Iryani, "A Smart Recommendation System for Crop Seed Selection Using Gradient Boosting Based on Environmental and Geospatial Data," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 6, pp. 2965–2973, 2025.

[17] E. Sari, H. Nurdiniyah, A. Nur, A. Yusuf, R. L. Prasetyo, and R. Shabihah, "Seleksi Fitur Berbasis Mutual Information untuk Optimalisasi Model Prediksi Tingkat Kematian Penderita Gagal Jantung Menggunakan Machine Learning," *Jurnal, Juritek Teknik, Ilmiah Komputer, Elektro*, vol. 5, no. 2, 2025, doi: 10.51903/Juritek.V5i2.5084.

[18] Z. J. Zumantara *et al.*, "Perbandingan algoritma random forest dan logistic regression dalam prediksi penyakit diabetes," *Swadharma*, vol. 6, no. 1, 2026.