KLASIFIKASI VARIETAS KACANG MENGGUNAKAN XGBOOST DENGAN PENYESUAIAN *CLASS WEIGHTING*

Thesa Adi Saputra Yusri [1], Deddy Rudhistiar [2], Andika Putri Ratnasari [3]
[1] Program Studi Matematika, Universitas Negeri Yogyakarta, Indonesia
[2] Program Studi Teknik Informatika, Institut Teknologi Nasional Malang, Indonesia
[3] Program Studi Statistika, Universitas Negeri Yogyakarta, Indonesia
thesaadisaputrayusri@uny.ac.id

ABSTRAK

Kondisi data tidak imbang (*imbalance*) merupakan salah satu tantangan utama dalam masalah klasifikasi terkait kualitas atau penyakit pada bidang agrikultur. Penelitian ini bertujuan untuk mengimplementasikan algoritma XGBoost dalam klasifikasi varietas kacang kering dengan fokus pada penanganan ketidakseimbangan kelas melalui pembobotan kelas. Dataset yang digunakan terdiri dari tujuh jenis kacang kering dengan berbagai karakteristik fisik yang diukur dalam piksel, yang meliputi fitur dimensi dan bentuk. Proses normalisasi dilakukan menggunakan teknik min-max normalization untuk memastikan skala data konsisten. Untuk menangani ketidakseimbangan kelas, teknik pembobotan kelas diterapkan dalam XGBoost, yang memberikan bobot lebih pada kelas minoritas. *Grid Search* dengan 5-*fold cross-validation* digunakan untuk menemukan kombinasi hyperparameter terbaik, yang menghasilkan akurasi *cross-validation* sebesar 92.5% dan skor terbaik pada 92.8%. Evaluasi model pada data uji menunjukkan akurasi 93%, dengan hasil precision, recall, dan f1-score yang seimbang pada setiap kelas. Hasil ini menunjukkan bahwa XGBoost dengan pembobotan kelas dapat mengatasi ketidakseimbangan kelas dan memberikan akurasi yang tinggi pada klasifikasi kacang kering.

Keyword : *Kacang kering, XGBoost, Class weighting, imbalance dataset*

1. PENDAHULUAN

Kacang kering (*dry bean*) memegang peran penting dalam industri agrikultur sebagai sumber nutrisi esensial dan komoditas ekonomi signifikan [1]. Dalam konteks ini, klasifikasi yang akurat dari varietas kacang kering menjadi esensial untuk optimalisasi rantai pasokan dan penetapan harga yang lebih adil [2]. Teknologi machine learning, terutama algoritma XGBoost, telah diidentifikasi sebagai salah satu metode unggul dalam klasifikasi berbagai jenis tanaman pangan [3].

Dataset UCI *Machine Learning Repository* menyediakan basis untuk pengujian dan evaluasi algoritma ini, yang mencakup tujuh varietas kacang dengan karakteristik morfologis yang berbeda [1]. Namun, seperti dataset lain yang sering digunakan dalam penelitian, ketidakseimbangan kelas tetap menjadi tantangan utama yang mempengaruhi kinerja algoritma klasifikasi [4], [5].

Dalam situasi ketidakseimbangan kelas, model cenderung memberikan prediksi yang lebih akurat untuk kelas mayoritas, sementara performa menurun drastis untuk kelas minoritas [6]. Ini menjadi isu krusial ketika varietas kacang tertentu, yang mungkin kurang terwakili, justru memiliki nilai ekonomi atau gizi yang tinggi [7], [8].

Penerapan *class weighting* dalam XGBoost dapat mengatasi ketidakseimbangan kelas dengan cara memberikan bobot lebih pada observasi kelas minoritas dalam proses pelatihan [9]. Ini menandai peningkatan signifikan dibandingkan pendekatan *resampling* yang dapat mengubah distribusi alami data dan meningkatkan biaya komputasi [10].

Sejumlah penelitian telah menunjukkan efektivitas XGBoost dalam berbagai aplikasi agrikultur. Temuan pada [11] menyoroti kemampuan XGBoost dalam meningkatkan akurasi prediksi hasil panen di Malaysia, dengan nilai R^2 mencapai 0,98. Melalui analisis SHAP, faktor-faktor seperti suhu, curah hujan, dan penggunaan pestisida berhasil diidentifikasi sebagai penentu utama hasil panen, mendukung pengambilan keputusan bagi petani untuk meningkatkan ketahanan pangan. Penelitian [12] juga mendemonstrasikan keberhasilan model RF-XGBoost dalam memprediksi penjualan produk pertanian siklus pendek di Tiongkok, dengan peningkatan akurasi sebesar 24% dan penurunan MAPE hingga 12%. Model ini mengoptimalkan rantai pasok sekaligus mengurangi limbah makanan.

Penelitian [13] mengonfirmasi keunggulan XGBoost dibandingkan *Artificial Neural Network* (ANN) dalam prediksi kualitas tomat, dengan hasil akurasi tinggi untuk °Brix (R^2 = 0,98) dan kandungan likopen (R^2 = 0,87). Penelitian ini menekankan pengaruh signifikan kultivar dan faktor lingkungan terhadap kualitas hasil agrikultur. Secara keseluruhan, studi-studi ini menegaskan bahwa XGBoost adalah alat yang sangat efektif untuk mengatasi tantangan di sektor agrikultur modern, baik dalam meningkatkan akurasi prediksi hasil panen, meminimalkan limbah, maupun mendukung pertanian presisi berbasis data.

Penyesuaian berupa pembobotan pada kelas atau label dalam XGBoost menjadi salah satu faktor penentu keberhasilan model dalam menghadapi dataset yang tidak seimbang [14]. Penelitian [15]

menemukan bahwa penyesuaian bobot ini dapat meningkatkan F1-score hingga sebanyak 5% hingga 8% pada skenario klasifikasi yang tidak seimbang. Penelitian ini juga menunjukan nilai antara presisi dan recall yang lebih seimbang melalui penyesuaian bobot ini.

Dalam konteks industri pengolahan kacang, penerapan XGBoost dengan pembobotan kelas membuka peluang otomatisasi yang lebih efisien dan akurat, yang dapat mengurangi ketergantungan pada inspeksi manual serta menghemat sumber daya [16], [17]. Berdasarkan permasalahan di atas, penelitian ini bertujuan untuk mengimplementasikan algoritma XGBoost dalam klasifikasi kacang kering, dengan fokus pada penanganan ketidakseimbangan kelas melalui pembobotan pada setiap kelas. .

2. TINJAUAN PUSTAKA

Penerapan model XGBoost dalam klasifikasi data *imbalance* telah mendapatkan perhatian yang signifikan dalam beberapa tahun terakhir, terutama karena kemampuannya dalam mengatasi masalah ketidakseimbangan data dan meningkatkan akurasi prediksi melalui pendekatan *ensemble learning*. XGBoost (*eXtreme Gradient Boosting*) merupakan implementasi efisien dari *gradient boosting decision trees* (GBDT) yang mengoptimalkan fungsi objektif dengan memanfaatkan informasi deret pertama (*gradien*) dan deret kedua (*Hessian*) dari fungsi loss. Secara teoretis, model XGBoost dibangun secara iteratif dengan menambahkan pohon keputusan lemah untuk mengoreksi kesalahan prediksi model sebelumnya, sehingga fungsi objektif pada iterasi ke-t dapat diekspresikan pada Persamaan 1.

$$\mathcal{L}^t = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \quad (1)$$

dimana ℓ adalah fungsi kerugian, $\hat{y}_i^{t-1}$ merupakan prediksi hingga iterasi sebelumnya, f_t adalah fungsi (pohon) baru yang ditambahkan, dan Ω merupakan regularisasi yang mengendalikan kompleksitas model [3].

Dalam konteks pertanian, Penelitian [18] mengaplikasikan XGBoost bersama dengan transformasi Box-Cox dan evaluasi menggunakan repeated k-fold cross-validation untuk klasifikasi kacang kering. Hasil penelitian tersebut menunjukkan bahwa meskipun akurasi pada fase training mencapai mendekati 99%, akurasi validasi hanya sebesar 91%, yang mengindikasikan adanya potensi overfitting dan perlunya penyesuaian lebih lanjut pada parameter model.

Seiring dengan perkembangan tersebut, Penelitian [19] melakukan optimasi *hyperparameter* XGBoost untuk prediksi beban dengan menggunakan metode *grid search*. Penelitian ini menemukan bahwa penyesuaian parameter seperti *n_estimators*, *max_depth*, dan *learning_rate* secara signifikan meningkatkan performa model, sehingga model dapat lebih responsif terhadap karakteristik spesifik dataset pertanian.

Penelitian perbandingan juga telah dilakukan untuk mengevaluasi kinerja XGBoost relatif terhadap algoritma lain. Penelitian [20] membandingkan XGBoost dan Support Vector Machine (SVM) pada klasifikasi opini publik. Hasil studi tersebut menunjukkan bahwa XGBoost tidak hanya unggul dari segi kecepatan *training*, tetapi juga menghasilkan akurasi yang lebih tinggi, terutama ketika parameter *scale_pos_weight* digunakan untuk menangani ketidakseimbangan kelas.

Lebih lanjut, Penelitian [21] meneliti performa model klasifikasi berbasis XGBoost pada data intrusi pada jaringan IoT. Penelitian ini mengungkapkan bahwa dengan penerapan teknik pembobotan kelas, model XGBoost dapat mengurangi bias terhadap kelas mayoritas dan meningkatkan sensitivitas terhadap kelas minoritas, yang esensial untuk aplikasi dalam seleksi produk pertanian.

Secara keseluruhan, literatur terkini menunjukkan bahwa XGBoost merupakan algoritma yang fleksibel dan efektif untuk mengatasi permasalahan klasifikasi, khususnya dalam aplikasi pertanian. Pendekatan yang menggabungkan optimasi hyperparameter, teknik transformasi data (misalnya Box-Cox), dan penyesuaian bobot kelas telah terbukti mampu meningkatkan akurasi dan sensitivitas model terhadap kelas minoritas. Kontribusi penelitian-penelitian ini tidak hanya memperkaya literatur mengenai metode ensemble dalam machine learning, tetapi juga membuka peluang aplikasi praktis untuk meningkatkan efisiensi seleksi dan pengolahan produk pertanian melalui sistem klasifikasi otomatis.

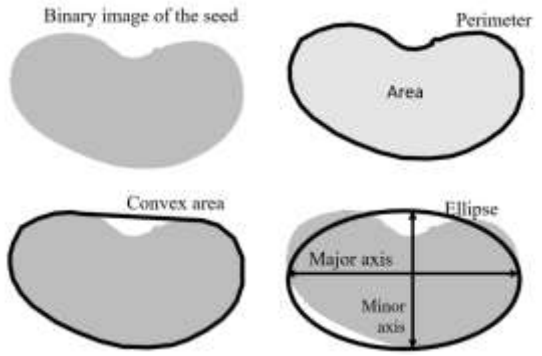
3. METODOLOGI

3.1. Deskripsi Data

Dataset yang digunakan dalam penelitian ini adalah data kacang kering (dry bean) yang tersedia di UCI *Machine Learning Repository*. Dataset ini terdiri dari tujuh jenis kacang kering yang diklasifikasikan berdasarkan berbagai karakteristik fisik, seperti bentuk, ukuran, warna, dan keragaman. Untuk mendukung klasifikasi otomatis, dikembangkan sebuah sistem berbasis visi komputer. Dalam sistem ini, kamera beresolusi tinggi digunakan untuk menangkap gambar sebanyak 13.611 butir kacang dari ketujuh jenis tersebut. Setelah gambar diperoleh, data melalui proses segmentasi dan ekstraksi fitur. Proses ini menghasilkan 16 fitur utama yang terbagi menjadi 12 fitur dimensi dan empat fitur bentuk, sebagaimana dijelaskan oleh [22].

Fitur-fitur tersebut meliputi *Area* (A), yaitu jumlah piksel dalam batas area kacang; *Perimeter* (P), yaitu panjang tepi kacang; Major Axis Length (L) dan Minor Axis Length (l), yang merupakan panjang sumbu utama dan sumbu minor dari kacang; serta Aspect Ratio (K), yang dihitung dari rasio L

terhadap l. Eksentrisitas (Eccentricity, Ec) mengukur keelipsan kacang berdasarkan momen wilayahnya, sementara *Convex Area* (C) adalah jumlah piksel dalam poligon cembung terkecil yang mencakup area kacang. Selain itu, fitur seperti *Equivalent Diameter* (Ed), *Extent* (Ex), *Solidity* (S), *Roundness* (R), dan *Compactness* (CO) memberikan gambaran geometris dan sifat fisik kacang.



Gambar 1. Representasi Fitur spasial dasar dari citra biner kacang.

Untuk fitur bentuk, digunakan beberapa faktor bentuk: Shape Factor 1 (SF1) dan Shape Factor 2 (SF2) mengukur rasio antara panjang sumbu terhadap area, sedangkan Shape Factor 3 (SF3) dan Shape Factor 4 (SF4) melibatkan perhitungan area dalam kaitannya dengan elips yang berorientasi pada sumbu tertentu. Fitur-fitur ini menjadi dasar dalam klasifikasi kacang secara otomatis, seperti dijelaskan pada penelitian [1].

3.2. Normalisasi

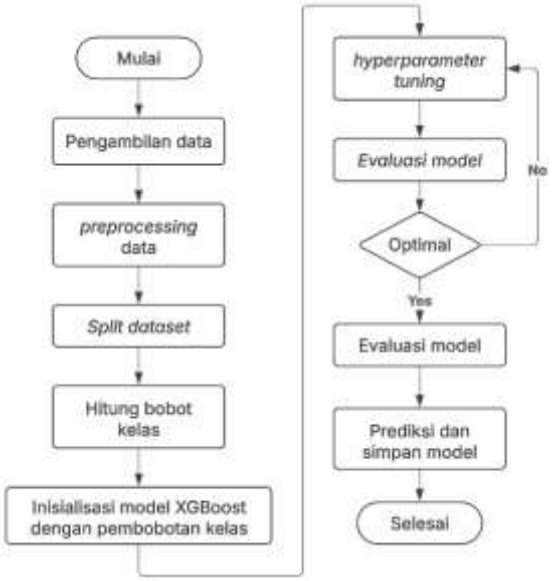
Min-max normalization adalah teknik normalisasi yang digunakan untuk menyelaraskan data dengan cara mengubah skala data mentah menjadi rentang tertentu secara linear [23]. Metode ini mentransformasi data ke dalam rentang 0 hingga 1 atau -1 hingga 1 menggunakan Persamaan 2 berikut:

$$v_i' = \frac{(v_i - \min_A)}{(\max_A - \min_A)} \times (\max_{A\,new} - \min_{A\,new}) + \min_{A\,new} \quad (2)$$

Dengan min_A dan max_A adalah nilai minimum dan maksimum dari data ke-i secara berurutan. Batas bawah dan batas atas untuk menyekalakan ulang data dinotasikan dengan $min_{A\,new}$ dan $max_{A\,new}$ [24]. Dalam penelitian ini, skala yang digunakan untuk menganalisis kinerja klasifikasi adalah [0,1] dan [-1,1].

3.3. Pelatihan Model XGBoost

Setelah proses normalisasi, dataset dibagi menjadi dua bagian: 70% sebagai data latih untuk membangun model klasifikasi, dan 30% sebagai data uji untuk menguji model. Proses pemodelan dilakukan menggunakan metode Extreme Gradient Boosting (XGBoost), yang pertama kali diperkenalkan oleh Chen dan Guestrin pada tahun 2016 [3]. XGBoost dikenal sebagai estimator canggih dengan performa luar biasa dalam tugas klasifikasi dan regresi, menawarkan berbagai peningkatan dibandingkan algoritma gradient boosting tradisional. Salah satu keunggulannya adalah penggunaan fungsi loss dengan regularisasi untuk mencegah overfitting, serta penerapan deret Taylor orde kedua sebagai fungsi objektif.



Gambar 2. Diagram alir penelitian

Parameter XGBoost yang digunakan dalam penelitian ini disajikan dalam Tabel 1.

Tabel 1. Parameter XGBoost yang Digunakan

Parameter	Nilai yang Dicoba
learning_rate	[0.001, 0.01, 0.1, 0.2]
n_estimators	[50, 100, 200]
max_depth	[3, 4, 6, 8, 10]
subsample	[0.5, 1.0]
colsample_bytree	[0.3, 1.0]

- *learning_rate*: Mengontrol ukuran langkah setiap iterasi boosting. Nilai yang lebih kecil membuat model lebih robust terhadap overfitting tetapi memerlukan lebih banyak iterasi.
- *n_estimators*: Menentukan jumlah pohon keputusan yang akan dibangun oleh model. Lebih banyak pohon dapat meningkatkan akurasi tetapi juga meningkatkan risiko overfitting.
- *max_depth*: Kedalaman maksimum setiap pohon keputusan. Kedalaman yang lebih besar memungkinkan model menangkap pola yang lebih kompleks tetapi dapat menyebabkan overfitting.
- *subsample*: Proporsi sampel pelatihan yang digunakan untuk membangun setiap pohon. Nilai kurang dari 1.0 memperkenalkan randomisasi dan membantu mencegah overfitting.

- *colsample_bytree*: Proporsi fitur yang digunakan untuk membangun setiap pohon. Pengurangan jumlah fitur per pohon dapat membantu mengurangi overfitting, terutama pada dataset berdimensi tinggi.

Proses tuning dilakukan menggunakan teknik pencarian grid (*grid search*) untuk menemukan kombinasi parameter optimal yang memaksimalkan kinerja model pada data validasi. Setelah model terbaik diperoleh, evaluasi akhir dilakukan pada data uji untuk menilai generalisasi model terhadap data yang tidak terlihat selama pelatihan.

Proses penganan kelas yang tidak seimbang dilakuka dengan menerapkan teknik pembobotan kelas dalam algoritma XGBoost. Pendekatan ini memberikan bobot lebih tinggi pada sampel dari kelas yang kurang terwakili, memastikan model memberikan perhatian yang proporsional selama pelatihan. Bobot untuk setiap kelas dihitung menggunakan Persamaan 3.

$$\text{class\_weight}_i = \frac{N}{k \times n_i} \quad (3)$$

dimana N adalah total jumlah sampel dalam dataset, k adalah jumlah total kelas, dan n_i adalah jumlah sampel dalam kelas ke-i.

Dengan demikian, sampel dari kelas minoritas akan memiliki bobot lebih tinggi, membantu model belajar pola yang lebih representatif dari kelas-kelas tersebut. Secara umum, proses pelatihan dan pengujian model diilustrasikan pada Gambar 2

3.4. Evaluasi Model

Dalam penelitian ini, proses evaluasi kinerja model **XGBoost** yang telah dilatih untuk mengklasifikasikan data kacang kering ke dalam tujuh kelas. Menghadapi tantangan ketidakseimbangan kelas, kami memilih metrik evaluasi yang tepat untuk memastikan bahwa model tidak hanya unggul dalam akurasi keseluruhan tetapi juga efektif dalam mengenali kelas minoritas.

3.5. Matriks Kebingungan (*Confusion Matrix*)

Matriks kebingungan digunakan untuk menganalisis performa model dengan menampilkan jumlah prediksi benar dan salah untuk setiap kelas. Dalam konteks klasifikasi multi-kelas, matriks ini berbentuk matriks $k \times k$, dimana k adalah jumlah kelas. Setiap entri M_{ij} menunjukkan jumlah sampel yang sebenarnya berasal dari kelas i tetapi diprediksi sebagai kelas j. Analisis matriks ini membantu dalam mengidentifikasi pola kesalahan prediksi yang spesifik untuk setiap kelas.

3.6. Metrik Evaluasi

Akurasi (*Accuracy*) Mengukur proporsi prediksi yang benar terhadap total sampel. Namun, pada dataset yang tidak seimbang, akurasi saja tidak cukup representatif karena model dapat cenderung mengabaikan kelas minoritas.

Presisi (*Precision*): Mengukur proporsi prediksi positif yang benar terhadap total prediksi positif. Untuk kelas i, presisi dihitung dengan Persamaan 4.

$$\text{Presisi}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (4)$$

dimana TP_i adalah *True Positives* untuk kelas i dan FP_i adalah *False Positives* untuk kelas i.

Recall (Sensitivitas) Mengukur proporsi sampel aktual positif yang diprediksi dengan benar. Untuk kelas i, *recall* dihitung dengan Persamaan 5.

$$\text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \quad (5)$$

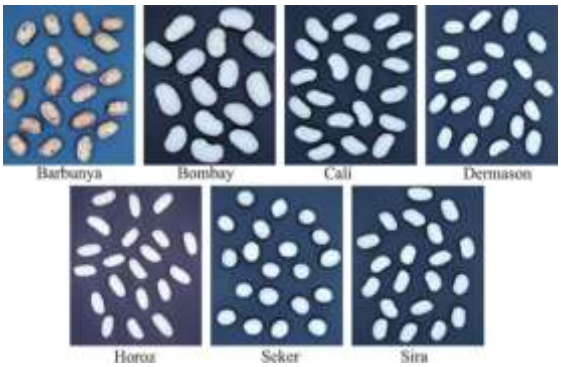
di mana FN_i: *False Negatives* untuk kelas i.

Skor F1 (F1-*Score*) merupakan rata-rata harmonis dari presisi dan *recall*, memberikan keseimbangan antara keduanya. Untuk kelas i, skor F1 dihitung dengan Persamaan 6.

$$\text{F1}_i = 2 \times \frac{\text{Presisi}_i \times \text{Recall}_i}{\text{Presisi}_i + \text{Recall}_i} \quad (6)$$

4. HASIL DAN PEMBAHASAN

Pada bagian ini, pertama-tama disajikan Gambar 3 yang menunjukkan contoh gambar dari setiap kelas kacang kering yang terdapat dalam dataset. Gambar-gambar ini menggambarkan ciri-ciri visual dari setiap jenis kacang kering yang digunakan dalam penelitian.

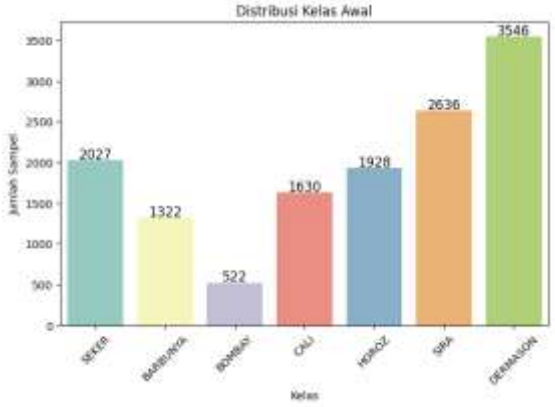


Gambar 3. Contoh citra kacang kering

Kemudian, Gambar 3 menyajikan distribusi awal jumlah sampel untuk masing-masing kelas. Berdasarkan Gambar 4, terlihat bahwa terdapat ketidakseimbangan yang signifikan dalam distribusi kelas pada dataset. Kelas "Dermason" memiliki jumlah sampel terbanyak, yakni 3.546 sampel, sedangkan kelas "Bombay" hanya memiliki 522 sampel. Ketidakseimbangan ekstrem ini berpotensi menyebabkan bias dalam proses pembelajaran model klasifikasi, yang dapat menurunkan akurasi prediksi keseluruhan. Oleh karena itu, penting untuk melakukan penyeimbangan data terlebih dahulu guna memastikan performa model yang optimal.

Tabel 2 menyajikan statistik deskriptif untuk 16 fitur yang diukur dalam satuan piksel, yang menggambarkan penyebaran data pada masing-masing fitur kacang kering. Statistik deskriptif ini termasuk nilai minimum, maksimum, rata-rata, dan standar deviasi, yang memberikan gambaran

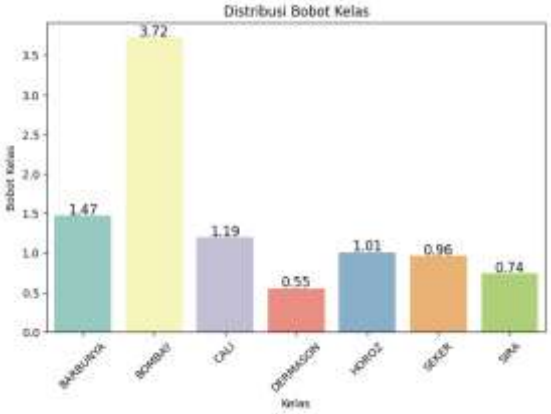
mengenai variasi data untuk setiap fitur. Secara keseluruhan, statistik ini menunjukkan distribusi nilai yang cukup bervariasi di antara fitur-fitur yang diukur, yang nantinya akan digunakan dalam proses normalisasi dan pemodelan.



Gambar 4. Distribusi kelas awal

Gambar 4 menunjukkan Distribusi Bobot Kelas untuk dataset kacang kering. Berdasarkan grafik tersebut, kelas ”Bombay” memiliki bobot yang paling tinggi, yaitu 3.72, yang mencerminkan ketidakseimbangan kelas yang signifikan dalam dataset ini. Sebaliknya, kelas "Dermason" memiliki bobot paling rendah, yakni 0.55, yang sesuai dengan jumlah sampel yang lebih banyak dalam kelas tersebut. Bobot yang lebih tinggi diberikan kepada kelas-kelas dengan jumlah sampel yang lebih sedikit untuk memastikan model lebih memfokuskan perhatian pada kelas minoritas selama pelatihan.

Grafik pada Gambar 5 menggambarkan bagaimana pembobotan kelas dapat menyeimbangkan kontribusi setiap kelas terhadap proses pelatihan model, membantu mengatasi masalah ketidakseimbangan kelas dan meningkatkan akurasi prediksi untuk kelas-kelas yang lebih jarang.



Gambar 5. Distribusi bobot Kelas

Untuk meningkatkan kinerja model XGBoost, dilakukan pencarian *hyperparameter* menggunakan metode *Grid Search* dengan 5-*fold cross-validation*. Proses ini bertujuan untuk menemukan kombinasi terbaik dari *hyperparameter* model, yang meliputi learning_rate, n_estimators, max_depth, subsample, dan colsample_bytree. *Grid search* mencoba semua kombinasi yang mungkin dari nilai-nilai yang telah didefinisikan dalam param_grid dan mengevaluasi setiap kombinasi berdasarkan skor akurasi yang diperoleh selama validasi silang. Adapun parameter yang diuji sesuai dengan Tabel 1.

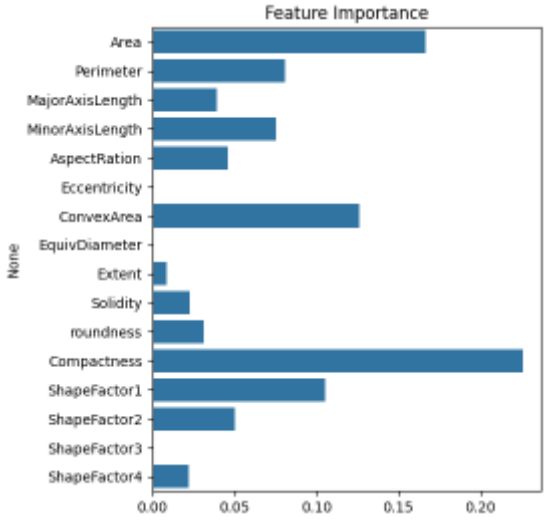
Setelah melakukan pencarian *hyperparameter* menggunakan *Grid Search* dengan 5-*fold cross-validation*, model XGBoost menunjukkan hasil yang sangat baik. Akurasi *cross-validation* yang diperoleh adalah 92.5%, yang mengindikasikan model memiliki kemampuan generalisasi yang baik pada data latih. *Hyperparameter* terbaik yang ditemukan adalah:

- colsample_bytree: 1.0
- learning_rate: 0.1
- max_depth: 6
- n_estimators: 100
- subsample: 0.5

Tabel 2. Statistik deskriptif untuk fitur kacang kering (dalam pixel)

No	**Fitur**	***Min.***	***Max.***	***Mean***	***Std. Dev.***
1	Area	20.420	254.616	53.048	29.324,1
2	Perimeter	524,7	1.985,4	855,3	214,2897
3	Major Axis Length	183,6	738,9	320,1	85,6941
4	Minor Axis Length	122,5	460,2	202,3	44,9700
5	Aspect Ration	1,025	2,430	1,583	0,2466
6	Eccentricity	0,2190	0,9114	0,7509	0,0920
7	Convex Area	20.684	263.261	53.768	29.774,92
8	Equivalent Diameter	161,2	569,4	253,1	59,1771
9	Extent	0,5553	0,8662	0,7497	0,0490
10	Solidity	0,9192	0,9947	0,9871	0,0046
11	Roundness	0,4896	0,9907	0,8733	0,0595
12	Compactness	0,6406	0,9873	0,7999	0,0617
13	Shape Factor 1	0,0027	0,0104	0,0065	0,0011
14	Shape Factor 2	0,0005	0,0036	0,0017	0,0005
15	Shape Factor 3	0,4103	0,9748	0,6436	0,0989
16	Shape Factor 4	0,9477	0,9997	0,9951	0,0043

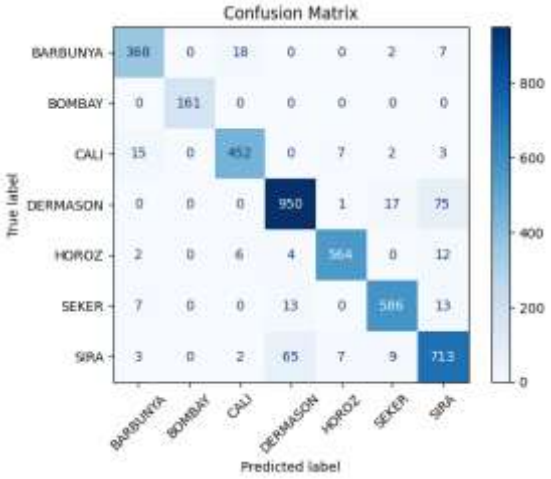
Dengan kombinasi *hyperparameter* tersebut, skor terbaik pada *cross-validation* yang diperoleh adalah 92.8%.



Gambar 6. *Feature Importance*

Gambar 6 menunjukkan *Feature Importance* untuk model klasifikasi kacang kering yang menggunakan algoritma XGBoost. Berdasarkan grafik, fitur *Area* dan *Compactness* memiliki pengaruh paling besar terhadap keputusan model, dengan nilai importance yang jauh lebih tinggi dibandingkan fitur lainnya. Fitur *Area* dan *Compactness* secara jelas menunjukkan bahwa ukuran dan bentuk fisik kacang, seperti luas area dan tingkat kekompakan, adalah faktor dominan dalam klasifikasi jenis kacang kering. Selain itu, fitur *Perimeter*, *MajorAxisLength*, dan *ConvexArea* juga menunjukkan tingkat kepentingan yang signifikan, mengindikasikan bahwa dimensi morfologis kacang berperan penting dalam membedakan setiap kelas. Sementara itu, fitur-fitur seperti *ShapeFactor1*, *ShapeFactor2*, *ShapeFactor3*, dan *ShapeFactor4* memiliki kontribusi yang lebih kecil, menunjukkan bahwa faktor bentuk yang lebih kompleks mungkin kurang berpengaruh dalam klasifikasi dibandingkan dengan dimensi dasar seperti area dan perimeter.

Matriks kebingungan pada Gambar 7 menunjukkan hasil evaluasi model klasifikasi untuk dataset kacang kering, dengan tujuh kelas yang berbeda. Dari matriks ini, terlihat bahwa model memiliki akurasi yang sangat baik dalam mengklasifikasikan sebagian besar kelas, dengan nilai diagonal utama yang tinggi, menunjukkan banyaknya prediksi yang benar (*True Positives*) untuk setiap kelas. Kelas dengan jumlah sampel terbesar, seperti Dermason dan Sira, memiliki prediksi yang sangat tepat, dengan 950 dan 713 sampel yang benar, masing-masing. Namun, kesalahan prediksi juga terjadi pada beberapa kelas minoritas, seperti Bombay dan Barbunya, yang memiliki lebih banyak *False Positives* dan *False Negatives*, mencerminkan ketidakseimbangan kelas dalam dataset. Hal ini menunjukkan bahwa model lebih cenderung memprediksi kelas mayoritas dengan benar dan cenderung kesulitan dalam mengenali kelas yang lebih sedikit terwakili.



Gambar 7. Confusion Matrix

Kesalahan prediksi yang terjadi pada kelas minoritas ini dapat dijelaskan oleh ketidakseimbangan kelas yang ada pada dataset, di mana kelas seperti Bombay dengan 522 sampel cenderung lebih sulit dikenali oleh model dibandingkan dengan kelas yang memiliki jumlah sampel lebih banyak, seperti "Dermason" yang memiliki 3.546 sampel. Meskipun demikian, model menunjukkan performa yang cukup baik secara keseluruhan dengan akurasi yang tinggi pada kelas mayoritas.

5. KESIMPULAN DAN SARAN

Penelitian ini berhasil menunjukkan efektivitas penggunaan algoritma XGBoost dalam klasifikasi varietas kacang kering dengan mengatasi ketidakseimbangan kelas melalui pembobotan kelas. Model yang dihasilkan mencapai akurasi 92,8% pada data uji dan menunjukkan performa yang baik dalam menangani kelas minoritas. Hasil ini mengindikasikan bahwa XGBoost dapat menjadi model yang efektif untuk aplikasi klasifikasi di industri pertanian, khususnya dalam klasifikasi produk dengan distribusi data yang tidak seimbang. Untuk penelitian selanjutnya, disarankan untuk mengeksplorasi penggunaan teknik lainnya, seperti SMOTE untuk oversampling, serta menguji model pada dataset yang lebih besar atau lebih beragam untuk meningkatkan generalisasi model.

DAFTAR PUSTAKA

[1] M. Koklu and I. A. Ozkan, "Multiclass classification of dry beans using computer vision and machine learning techniques," *Comput. Electron. Agric.*, vol. 174, p. 105507, Jul. 2020, doi: 10.1016/j.compag.2020.105507.

[2] J. C. Macuácua, J. A. S. Centeno, and C.

Amisse, "Data mining approach for dry bean seeds classification," *Smart Agric. Technol.*, vol. 5, p. 100240, Oct. 2023, doi: 10.1016/j.atech.2023.100240.

[3] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[5] Haibo He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[6] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-98074-4.

[7] D. Prakash, A. Niranjan, S. K. Tewari, and P. Pushpangadan, "Underutilised legumes: potential sources for low-cost protein," *Int. J. Food Sci. Nutr.*, vol. 52, no. 4, pp. 337–341, Jan. 2009, doi: 10.1080/09637480120057521.

[8] J. Popoola, O. Ojuederie, C. Omonhinmin, and A. Adegbite, "Neglected and Underutilized Legume Crops: Improvement and Future Prospects," in *Recent Advances in Grain Crops Research*, IntechOpen, 2020. doi: 10.5772/intechopen.87069.

[9] C. Wang, C. Deng, and S. Wang, "Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost," *Pattern Recognit. Lett.*, vol. 136, pp. 190–197, Aug. 2020, doi: 10.1016/j.patrec.2020.05.035.

[10] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, Sep. 2017, pp. 79–85. doi: 10.1109/ICACCI.2017.8125820.

[11] D. A.-L. Mariadass, E. G. Moung, M. M. Sufian, and A. Farzamnia, "Extreme Gradient Boosting (XGBoost) Regressor and Shapley Additive Explanation for Crop Yield Prediction in Agriculture," in *2022 12th International Conference on Computer and Knowledge Engineering (ICCKE)*, IEEE, Nov. 2022, pp. 219–224. doi: 10.1109/ICCKE57176.2022.9960069.

[12] H. I. Yoon *et al.*, "Predicting Models for Plant Metabolites Based on PLSR, AdaBoost, XGBoost, and LightGBM Algorithms Using Hyperspectral Imaging of Brassica juncea," *Agriculture*, vol. 13, no. 8, p. 1477, Jul. 2023, doi: 10.3390/agriculture13081477.

[13] O. M'hamdi, S. Takács, G. Palotás, R. Ilahy, L. Helyes, and Z. Pék, "A Comparative Analysis of XGBoost and Neural Network Models for Predicting Some Tomato Fruit Quality Traits from Environmental and Meteorological Data," *Plants*, vol. 13, no. 5, p. 746, Mar. 2024, doi: 10.3390/plants13050746.

[14] C. A. E. Piter, S. Hadi, and I. N. Yulita, "Multi-Label Classification for Scientific Conference Activities Information Text Using Extreme Gradient Boost (XGBoost) Method," in *2021 International Conference on Artificial Intelligence and Big Data Analytics*, IEEE, Oct. 2021, pp. 1–5. doi: 10.1109/ICAIBDA53487.2021.9689699.

[15] R. Sharma, N. K. Valivati, G. K. Sharma, and M. Pattanaik, "A New Hardware Trojan Detection Technique using Class Weighted XGBoost Classifier," in *2020 24th International Symposium on VLSI Design and Test (VDAT)*, IEEE, Jul. 2020, pp. 1–6. doi: 10.1109/VDAT50263.2020.9190603.

[16] E. Ramirez-Asis, J. Vilchez-Carcamo, C. M. Thakar, K. Phasinam, T. Kassanuk, and M. Naved, "A review on role of artificial intelligence in food processing and manufacturing industry," *Mater. Today Proc.*, vol. 51, pp. 2462–2465, 2022, doi: 10.1016/j.matpr.2021.11.616.

[17] S. Ghazal, A. Munir, and W. S. Qureshi, "Computer vision in smart agriculture and precision farming: Techniques and applications," *Artif. Intell. Agric.*, vol. 13, pp. 64–83, Sep. 2024, doi: 10.1016/j.aiia.2024.06.004.

[18] I. Wardhana, M. Ariawijaya, V. A. Isnaini, and R. P. Wirman, "Gradient Boosting Machine, Random Forest dan Light GBM untuk Klasifikasi Kacang Kering," *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 6, no. 2, pp. 92–99, 2022, doi: https://doi.org/10.29207/resti.v6i1.3682.

[19] N. T. Tran, T. T. G. Tran, T. A. Nguyen, and M. B. Lam, "A new grid search algorithm based on XGBoost model for load forecasting," *Bull. Electr. Eng. Informatics*, vol. 12, no. 4, pp. 1857–1866, Aug. 2023, doi: 10.11591/eei.v12i4.5016.

[20] D. Safitri, Susanti, Rahmaddeni, and T. A. Fitri, "Perbandingan Algoritma XGBoost dan SVM Dalam Analisis Opini Publik Pemilihan Presiden 2024," *Indones. J. Comput. Sci.*, vol. 13, no. 3, Jun. 2024, doi: 10.33022/ijcs.v13i3.4041.

[21] T.-T.-H. Le, Y. E. Oktian, and H. Kim, "XGBoost for Imbalanced Multiclass

Classification-Based Industrial Internet of Things Intrusion Detection Systems," *Sustainability*, vol. 14, no. 14, p. 8707, Jul. 2022, doi: 10.3390/su14148707.

[22] N. S. Visen, J. Paliwal, D. S. Jayas, and N. D. G. White, "AE—Automation and Emerging Technologies," *Biosyst. Eng.*, vol. 82, no. 2, pp. 151–159, Jun. 2002, doi: 10.1006/bioe.2002.0064.

[23] J. Han, M. Kamber, and J. Pei, *Data Mining*. Elsevier, 2012. doi: 10.1016/C2009-0-61819-5.

[24] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Appl. Soft Comput.*, vol. 97, p. 105524, Dec. 2020, doi: 10.1016/j.asoc.2019.105524.