ISSN: 2746-4237
IKOMTI - 23

Perbandingan Teknik Klasifikasi Data Mining untuk Penentuan Jenis Jamur Beracun

Herik Sugeru[1)], Hilmi A'ini Nurthoyibah[2)], Muhammad Affarel Abhinaya Nur Fajar[3)], Vindiar Johan Diputra[4)], dan Muchammad Yafik Ramadhani Ilham[5)]

[(1,2,3)] Program Studi Agroteknologi, Fakultas Teknologi Industri, Universitas Gunadarma
[(3,4)] Program Studi Informatika, Fakultas Teknologi Industri, Universitas Gunadarma
[1)]herik_sugeru@staff.gunadarma.ac.id✉, [2)] hilmitoyibah@gmail.com,
[3)]affarelabhinaya@gmail.com, [4)]vindiari5c@gmail.com [5)]mhmmdyfik@gmail.com

ABSTRACT

Data mining is the process of discovering knowledge in databases, involving data identification, validation, novelty, and the understanding of large and complex data patterns. One area that can benefit from data mining analysis is agriculture, where mushroom classification is essential to determine whether they are safe for consumption or poisonous. This research aims to evaluate the classification accuracy of various algorithms on a mushroom dataset, obtained from the UC Irvine Machine Learning Repository. Several algorithms were tested, including k-Nearest Neighbor (kNN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, Logistic Regression, Linear Discriminant Analysis, Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Extra Trees Classifier, AdaBoost, and Voting Feature Intervals 5 (VFI 5). Based on the results, the accuracy of the Hold Out method ranged from 0.8763 to 1.0000, while Cross Validation accuracy ranged from 0.8396 to 1.0000. Decision Tree, AdaBoost, and ANN achieved the highest accuracy (100%) on both testing methods. However, Cross Validation is recommended to reduce the risk of overfitting, even though it requires a longer processing time compared to Hold Out. Algorithms such as k-NN, Decision Tree, ANN, AdaBoost, and CNN show perfect results in one or both methods, indicating potential overfitting, particularly if the dataset is not sufficiently complex or if data classes are imbalanced.

Keywords: *data mining, determination mushroom types, classification, hold out, cross validation*

ABSTRAK

Penambangan data merupakan proses penemuan pengetahuan dalam basis data yang melibatkan identifikasi, validasi, kebaruan, dan pemahaman terhadap pola data yang besar dan kompleks. Salah satu bidang yang dapat menerapkan analisis berbasis data mining adalah pertanian. Jamur merupakan salah satu komoditas pertanian yang memiliki nutrisi tinggi dan nilai ekonomi yang menjanjikan dalam agribisnis. Klasifikasi jamur menjadi penting untuk menentukan apakah jamur aman dikonsumsi atau beracun. Penelitian ini bertujuan untuk mengevaluasi akurasi klasifikasi berbagai algoritma dalam dataset jamur, yang diperoleh dari UC Irvine Machine Learning Repository. Beberapa algoritma yang diuji meliputi k-Nearest Neighbor (kNN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, Logistic Regression, Linear Discriminant Analysis, Artificial Neural Network (ANN), Convolutional Neural Network (CNN), Extra Trees Classifier, AdaBoost, dan Voting Feature Intervals 5 (VFI 5). Berdasarkan hasil penelitian, akurasi Hold Out dari algoritma tersebut berkisar antara 0,8763 hingga 1,0000, sedangkan akurasi Cross Validation berkisar antara 0,8396 hingga 1,0000. Algoritma Decision Tree (Tuning Model), AdaBoost, dan ANN menunjukkan akurasi tertinggi (100%) pada kedua metode pengujian. Namun, penggunaan Cross Validation lebih disarankan untuk menghindari risiko overfitting, meskipun memerlukan waktu pemrosesan yang lebih lama dibandingkan dengan Hold Out. Algoritma seperti k-NN, Decision Tree, ANN, AdaBoost, dan CNN berpotensi mengalami overfitting, terutama jika dataset tidak cukup kompleks atau kelas data tidak seimbang.

Kata Kunci : data mining, penentuan jenis jamur, klasifikasi, hold out, cross validation

I. PENDAHULUAN

Analisis *big data* pertanian adalah istilah yang digunakan untuk mendiskripsikan data dalam jumlah besar sehingga lebih bermanfaat [1, 2], yang didesain sedemikian rupa hingga petani dan lembaga-lembaga terkait dapat mengambil nilai ekonomi dari analisis data terkait bidang pertanian dalam jumlah besar

Diterima Redaksi: 25-09-2024 | Selesai Revisi: 16-10-2024 | Diterbitkan Online: 23-10-2024

dengan metode yang praktis dan cepat [3, 4, 5]. Kim, *et al.* [6, 7] menyatakan bahwa lembaga pemerintah telah menggunakan analisis *big data* untuk meningkatkan pelayanan kepada masyarakat terkait masalah ekonomi, kesehatan, lapangan kerja, penanganan bencana alam dan terorisme. Meskipun analisis *big data* dinilai telah berhasil diterapkan dalam berbagai bidang tetapi di bidang pertanian masih belum banyak diterapkan [4, 5], baru sedikit pemangku kepentingan menerapkan analisis tersebut [2, 8, 9, 10, 11].

Salah satu metode yang dapat membantu analisis big data adalah penambangan data (*data mining*). Penambangan data atau *data mining* adalah proses mengekstraksi dan menemukan informasi dari *database* dan mengkonversinya menjadi informasi yang berguna untuk memunculkan pengetahuan. Melalui *data mining* juga dapat meningkatkan pengetahuan dan membantu mengembangkan model yang dapat mengungkap koneksi dengan jutaan bahkan miliaran rekaman data [12]. Pada dasarnya, *data mining* merupakan proses *Knowledge Discovery in Database* (KDD) yaitu penemuan pengetahuan dalam basis data yang melibatkan identifikasi data, validasi, kebaruan, dan pemahaman tentang pola data yang besar dan kompleks. Secara teknis, terdapat empat hal yang dapat dilakukan dengan *data mining*, antara lain deskripsi kelas/konsep, analisis asosiasi, klasifikasi atau prediksi, dan analisis klaster [13, 18, 19]. Analisis dengan basis *data mining* dapat diterapkan ke berbagai bidang, tidak terkecuali untuk data dari sektor pertanian. Analisis terhadap data pertanian dapat menjadi penentu kebijakan/keputusan, seperti bagaimana meningkatkan produktivitas, mengklasifikasi jenis tanah, mengkategorikan berbagai jenis tanaman, dan lain sebagainya [14, 20, 21].

Seperti yang telah disebutkan sebelumnya, klasifikasi merupakan salah satu metode analisis yang terdapat dalam *data mining*. Klasifikasi dapat memetakan data ke suatu kelas/grup yang telah ditentukan sebelumnya. Hal ini disebut dengan *supervised learning* karena sebelum melakukan klasifikasi, kelas dari setiap observasi telah ditentukan terlebih dahulu.

Dalam bidang pertanian, kasus aplikatif dari klasifikasi salah satunya adalah menentukan apakah spesies jamur tiram (*gilled mushroom*) dari famili *Agaricus* dan *Lepiota* tergolong ke dalam kelas beracun (*poisonous*) atau aman dikonsumsi (*edible*) [15, 16]. Pentingnya mengklasifikasikan jamur tiram ke dalam dua kelas tersebut adalah untuk menghindari keracunan akibat mengkonsumsi jamur tiram yang tidak tepat. Keracunan jamur dapat menyebabkan berbagai gejala termasuk gastroenteritis, halusinasi, sindrom kolinergik atau antikolinergik, serta reaksi seperti disulfiram (kemerahan, kecemasan, palpitasi, dan kemungkinan hipotensi) [17]. Pentingnya pengklasifikasian jamur diangkat sebagai topik dalam penelitian ini dengan mengolah data set untuk dianalisis menggunakan basis *data mining*. Penetapan apakah suatu jamur tergolong ke dalam kelas beracun atau aman dikonsumsi didasarkan pada karakteristik fisik jamur tersebut.

Penelitian sebelumnya mengenai penerapan data mining dalam klasifikasi objek pertanian telah menunjukkan perkembangan yang signifikan dalam mengoptimalkan hasil pertanian dan meningkatkan efisiensi produksi. Sebagai contoh, metode *Decision Tree* telah digunakan secara luas untuk mengklasifikasikan tanaman berdasarkan kondisi tanah, cuaca, dan praktik agrikultur yang diterapkan [22]. Selain itu, *Support Vector Machine* (SVM) juga telah diterapkan untuk mengidentifikasi penyakit tanaman dengan tingkat akurasi yang tinggi, seperti yang ditunjukkan dalam studi oleh Kumar *et al* [23], di mana data citra daun digunakan untuk klasifikasi penyakit tanaman berbasis gejala visual. *k-Nearest Neighbor* (kNN) digunakan dalam klasifikasi tanaman berdasarkan pertumbuhan dan karakteristik genetik, sementara A*rtificial Neural Networks* (ANN) telah membantu dalam memprediksi hasil panen dengan memperhitungkan berbagai parameter pertanian [24].

Namun, terdapat beberapa kekurangan dalam penelitian sebelumnya. Sebagian besar penelitian masih bergantung pada dataset yang terbatas dan tidak mencerminkan variasi lingkungan pertanian yang lebih luas, yang membatasi generalisasi model [25]. Selain itu, pendekatan yang digunakan terkadang kurang memperhitungkan interaksi kompleks antar variabel, terutama dalam klasifikasi tanaman yang membutuhkan pemahaman mendalam mengenai hubungan antara faktor iklim, tanah, dan genetik. Banyak penelitian juga cenderung fokus pada satu jenis algoritma tanpa membandingkannya dengan metode lain yang mungkin memberikan hasil yang lebih baik untuk jenis dataset yang berbeda [26]. Oleh karena itu, diperlukan penelitian lebih lanjut yang tidak hanya memperluas dataset yang digunakan, tetapi juga membandingkan beberapa algoritma secara komprehensif, terutama dalam kasus klasifikasi objek pertanian yang beragam.

Pada penelitian ini, klasifikasi terhadap dataset jamur dilakukan dengan duabelas metode klasifikasi, dimana pada setiap metode klasifikasi akan menghasilkan nilai akurasi yang menjadi tolok ukur ketepatan klasifikasi. Namun sebelum melakukan klasifikasi, dilakukan tahap pembagian dataset menjadi data training dan testing dengan dua metode, yaitu *Hold Out* dan *Cross Validation* (CV). Sehingga nantinya diperoleh nilai akurasi dari setiap metode klasifikasi, dimana nilai tersebut bedasarkan metode pembagian data *Hold Out* dan CV.

II. METODE

Pada penelitian ini, metodologi yang digunakan meliputi tahapan sumber data, variabel penelitian, dan langkah analisis.

A. Sumber Data

Penelitian ini menggunakan data sekunder yang diperoleh dari *website* UCI Machine Learning. Pada

website tersebut, data yang digunakan yakni berjudul *Mushroom Data Set*. Terdapat 22 atribut (prediktor) dan 1 respon (klasifikasi) pada data tersebut, dimana masing-masing variabel bersifat kategorik dan terdiri dari 8124 observasi.

Seluruh observasi dalam dataset ini mencakup 23 spesies jamur tiram (*gilled mushroom*) dari famili *Agaricus* dan *Lepiota*. Setiap spesies diidentifikasikan ke dalam kelas aman dikonsumsi (*edible*) dan beracun (*poisonous*).

B. Variabel Penelitian

Tabel 1 berikut ini merupakan rincian dari variabel yang digunakan dalam penelitian ini.

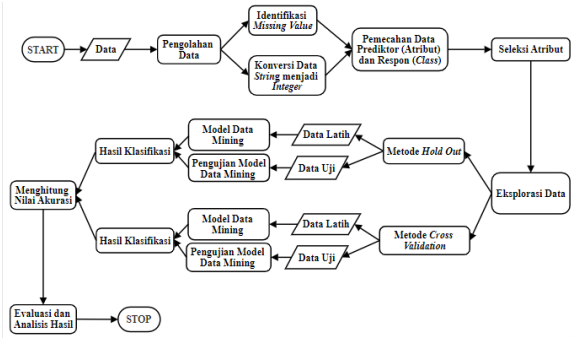
Tabel 1. Variabel Penelitian

Variabel	Keterangan	Skala Data
Cap shape	b = bell c = conical x = convex f = flat k = knobbed s = sunken	Nominal
Cap surface	f = fibrous g = grooves y = scaly s = smooth	Nominal
Cap color	n = brown b = buff c = cinnamon g = gray r = green p = pink u = purple e = red w = white y = yellow	Nominal
Bruises	t = bruises f = no	Nominal
Odor	a = almond l = anise c = creosote y = fishy f = foul m = musty n = none p = pungent s = spicy	Nominal
Gill attachment	a = attached d = descending f = free n = notched	Nominal
Gill spacing	c = close w = crowded d = distant	Nominal
Gill size	b = broad n = narrow	Nominal
Gill color	k = black n = brown B = buff h = chocolate g = gray r = green o = orange p = pink u = purple e = red w = white y = yellow	Nominal
Stalk shape	e = enlarging t = tapering	Nominal
Stalk root	b = bulbous c = club u = cup e = equal z = rhizomorphs r = rooted	Nominal
Stalk surface above ring	f = fibrous y = scaly k = silky s = smooth	Nominal
Stalk surface below ring	f = fibrous y = scaly k = silky s = smooth	Nominal
Stalk color above ring	n = brown b = Buff c = cinnamon g = gray o = orange p = pink e = red w = white y = yellow	Nominal
Stalk color below ring	n = brown b = buff c = cinnamon g = gray o = orange p = pink e = red w = white y = yellow	Nominal
Veil type	p = partial u = universal	Nominal
Veil color	n = brown o = orange w = white y = yellow	Nominal
Ring number	n = none o = one t = two	Nominal
Ring type	c = cobwebby e = evanescent f = flaring l = large n = none p = pendant s = sheathing z = zone	Nominal
Spore print color	k = black n = brown b = buff h = chocolate r = green o = orange u = purple	Nominal

	w = white y = yellow	
Population	a = abundant c = clustered n = numerous s = scattered v = several y = solitary	Nominal
Habitat	g = grasses l = leaves m = meadows p = paths u = urban w = waste d = woods	Nominal
Class	e = edible p = poisonous	Nominal

C. Langkah Analisis

Dataset diolah dan dianalisis menggunakan software RapidMiner. Penggunaan software RapidMiner dalam analisis data mining sangat populer karena beberapa kelebihan yang dimilikinya, terutama dalam hal kemudahan penggunaan, fleksibilitas, dan kemampuan menangani berbagai algoritma data mining secara efektif. Secara umum tahapan pengolahan dan klasifikasi *mushroom dataset* dengan algoritma *data mining* dalam penelitian ini dapat dilihat pada Gambar 1 di bawah ini.



Gambar 1. Diagram alir pengolahan dan klasifikasi *mushroom dataset* dengan algoritma *data mining*

Langkah analisis yang diterapkan pada penelitian ini adalah sebagai berikut:

1) Mengunduh *Mushroom Data Set* dari *website* UCI Machine Learning.
2) Prapengolahan data yang meliputi identifikasi *missing value* dan mengubah tipe data *string* menjadi *integer*.
3) Memecah data menjadi prediktor (atribut) dan respon (Class).
4) Menyeleksi atribut berdasarkan kontribusi terbesar.
5) Eksplorasi data terhadap variabel-variabel yang telah terpilih untuk dianalisis.
6) Melakukan klasifikasi dengan tahapan sebagai berikut:
   a. Membagi data menjadi *training-testing* dengan metode *Hold Out.*
   b. Melakukan klasifikasi dengan beberapa metode klasifikasi.
   c. Menghitung nilai akurasi dari masing-masing metode klasifikasi.
   d. Membagi data menjadi *training-testing* dengan metode CV.
   e. Melakukan klasifikasi dengan beberapa metode klasifikasi.
   f. Menghitung nilai akurasi dari masing-masing metode klasifikasi.
   g. Membandingkan nilai akurasi berdasarkan poin iii. dan vi.
   h. Menarik kesimpulan dan saran.

III. HASIL DAN PEMBAHASAN

A. Prapengolahan Data

Hal utama yang perlu dilakukan sebelum melakukan analisis adalah mengevaluasi data, yang kerap disebut dengan tahap *preprocessing*. Evaluasi ini bertujuan untuk mengetahui apakah terdapat *missing value* pada suatu atribut. Hasil pemeriksaan missing value terhadap *Mushroom Data Set* ditampilkan pada Tabel 2 di bawah ini.

Tabel 2. Identifikasi *Missing Value*

Variabel	**Jumlah *Missing Value***
Cap shape	0
Cap surface	0
Cap color	0
Bruises	0
Odor	0
Gill attachment	0
Gill spacing	0
Gill size	0
Gill color	0
Stalk shape	0
Stalk root	2480
Stalk surface above ring	0
Stalk surface below ring	0
Stalk color above ring	0
Stalk color below ring	0
Veil type	0
Veil color	0
Ring number	0
Ring type	0
Spore print color	0
Population	0
Habitat	0
Class	0

Berdasarkan tabel di atas, ditemukan adanya *missing value* pada variabel Stalk Root sebanyak 2480 observasi. Oleh karena itu dilakukan *imputasi missing value* dengan modus yang didasarkan pada jenis klasifikasi (Class). Imputasi dengan modus dipilih karena data *Mushroom Data Set* bersifat kategorik.

Sugeru, Nurthoyibah, Fajar, Diputra and Ilham
JURNAL ILMU KOMPUTER DAN TEKNOLOGI (IKOMTI)
Vol. 5, No. 3, Oktober 2024, pp. 23-31

Pada tahap *preprocessing* juga terdapat evaluasi *outlier*. Namun penelitian ini tidak melakukan evaluasi *outlier* karena data yang bersifat kategorik. Setelah tidak terdapat *missing value* pada data, maka observasi yang memiliki tipe *string* diubah menjadi *integer* (0, 1, 2, ...).

B. *Feature Selection*

Banyaknya prediktor (atribut) yang terdapat di *Mushroom Data Set* dapat menyebabkan analisis menjadi tidak efektif. Hal tersebut dikarenakan prediktor yang mungkin tidak berpengaruh terhadap klasifikasi jenis kelas jamur tetap diikutkan dalam proses analisis. Oleh karena itu, dilakukan *feature selection* atau seleksi variabel yang bertujuan untuk mengurangi ketidakefektifan tersebut.

Setelah dilakukan tahap *feature selection*, diperoleh hasil bahwa hanya 10 atribut yang berkontribusi terhadap klasifikasi jenis kelas jamur. Sedangkan 12 atribut lainnya tidak diikutkan ke dalam analisis lebih lanjut. Variabel-variabel yang terpilih pada tahap *feature selection* dapat dilihat pada Tabel 3 berikut:

Tabel 3. Nilai Kontribusi 10 Variabel Terpilih

Variabel	Kontribusi
Gill size	0,1139
Bruises	0,1101
Odor	0,1009
Ring type	0,0947
Stalk root	0,0681
Gill spacing	0,0677
Gill color	0,0657
Spore print color	0,0652
Population	0,0611
Stalk shape	0,0467

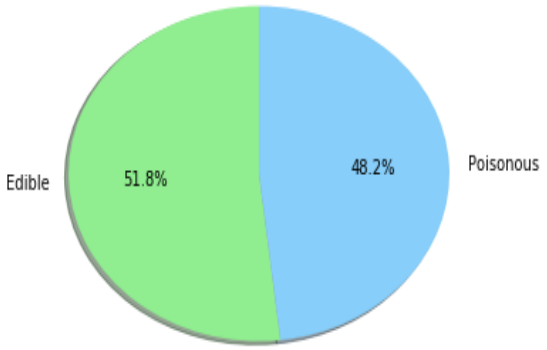
C. Eksplorasi Data

Untuk mengetahui karakteristik dari data yang akan dianalisis, maka diperlukan adanya eksplorasi data. Pada tahap ini, jumlah variabel yang dieksplorasi hanya 10 variabel, dimana jumlah tersebut diperoleh dari hasil *feature selection*. Eksplorasi data hanya sebatas analis statistika deskriptif (lihat Tabel 4) karena seluruh variabel bersifat kategorik.

Tabel 4. Statistika Deskriptif

Variabel	Modus	Keterangan	Frekuensi
Bruises	f	No	4748
Odor	n	None	3528
Gill spacing	c	Close	6812
Gill size	b	Broad	5612
Gill color	b	Buff	1728
Stalk shape	t	Tapering	4608
Stalk root	b	Bulbous	6256
Ring type	p	Pendant	3698
Spore print color	w	White	2388
Population	v	Several	4040
Class (Respon)	e	Edible	4208

Modus merupakan frekuensi jenis observasi yang sering muncul pada suatu variabel. Berdasarkan hasil tabel statistika deskriptif di atas, terlihat bahwa klasifikasi dalam *Mushroom Data Set* didominasi oleh kelas *edible*. Untuk mengetahui seberapa besar persentase antara kelas *ediBle* dengan *poisonous*, maka ditampilkan *pie chart* seperti di bawah ini.



Gambar 2. Persentase Jenis *Edible* dengan *Poisonous*

Jamur dengan jenis *edible* memiliki persentase sebesar 51,8%. Sedangkan sisanya, yaitu 48,2% merupakan kategori jamur *poisonous*. Selisih persentase yang tidak jauh berbeda menandakan bahwa data observasi dalam penelitian ini tergolong *balance*.

D. Klasifikasi *Mushroom Data Set*

Analisis klasifikasi terhadap *Mushroom Dataset* dilakukan dengan duabelas metode klasifikasi. Metode tersebut antara lain *k-Nearest Neighbor* (kNN), *Naive Bayes*, *Support Vector Machine* (SVM), *Decision Tree* (*Model Tuning dan Iterative Dichotomiser Tree (ID3)*), *Logistic Regression*, *Linear Discriminant Analysis*, *Artificial Neural Network* (ANN), *Convolution Neural Network* (CNN), *Extra Trees Classifier, AdaBoost,* dan *Voting Feature Intervals 5 (VFI 5)*. Pemilihan berbagai metode klasifikasi untuk *Mushroom Dataset* dilakukan karena setiap metode memiliki pendekatan yang unik dalam mengatasi masalah klasifikasi, dan perbandingan kinerja berbagai algoritma dapat memberikan gambaran yang lebih lengkap mengenai dataset tersebut.

Sebelum melakukan analisis klasifikasi, terdapat tahap pembagian data sehingga terbentuk data *training-testing*. Dalam penelitian ini, tahap pembagian data tersebut fokus pada dua metode, yakno *Hold Out* dan *Cross Validation*. Output dari analisis klasifikasi ini adalah memperoleh metode klasifikasi terbaik berdasarkan hasil akurasi dari dua metode pembagian data dan delapan metode klasifikasi.

E. *Hold Out Method*

Hold Out merupakan metode untuk membagi dataset yang ada menjadi data *training* dan *testing*. Umumnya, 1/3 Bagian dari dataset dijadikan sebagai data *testing*, sedangkan sisanya untuk data *training*.

Tabel 5. Perbandingan Nilai Akurasi, Presisi, dan *Recall* Metode *Hold Out*

Metode Klasifikasi	Akurasi	Presisi	*Recall*
k-Nearest Neighbor	1,0000	1,0000	1,0000
Naive Bayes	0,8763	0,8870	0,8712
Support Vector Machine	0,9686	0,9689	0,9681
Tuning (*Decision Tree*)	1,0000	1,0000	1,0000
ID3	0,9787	0,9791	0,9789
Logistic Regression	0,9329	0,9356	0,9308
Linear Discriminant Analysis	0,9003	0,9037	0,8975
ANN (Artificial Neural Network)	1,0000	1,0000	1,0000
CNN (Convolution Neural Network)	0,9979	0,9983	0,9980
Extra Trees Classifier	1,0000	1,0000	1,0000
AdaBoost	1,0000	1,0000	1,0000
VFI 5	0,9032	0,9040	0,9036

Hasil perbandingan nilai akurasi dari Tabel 5 menunjukkan bahwa, dalam kasus *Mushroom Data Set*, penerapan pembagian data dengan metode *Hold Out* akan menghasilkan nilai akurasi tertinggi ketika metode klasifikasi yang digunakan adalah *k-Nearest Neighbor* (kNN), *Decision Tree* (*Model Tuning*), *Artificial Neural Network* (ANN), AdaBoost dan *Extra Trees Classifier*. Kelima metode klasifikasi tersebut menghasilkan nilai akurasi sebesar 1, atau dapat dikatakan pengklasifikasian jenis kelas jamur 100% akurat. Namun terdapat kekurangan yang ditemui pada metode *Hold Out*, yaitu pembagian data *training* dan *testing* yang bisa saja tidak representatif. Contohnya adalah suatu kelas klasifikasi tidak terdapat di dalam data *testing*. Untuk mengatasi kekurangan tersebut, digunakan metode lain yang juga bertujuan untuk membagi dataset menjadi data *training-testing*, yaitu *Cross Validation* (CV).

F. *Cross Validation Method*

Konsep dasar dari metode *Cross Validation* (CV) adalah menjalankan sebuah model menggunakan *k-1 folds* yang dijadikan sebagai data *training*. Selanjutnya, hasil dari model tersebut divalidasi pada bagian data yang tersisa, yang tidak lain adalah data *testing*. Adanya validasi dapat digunakan untuk mengetahui seberapa akurat model klasifikasi yang dihasilkan.

Istilah lain untuk penyebutan CV adalah *K-Fold Cross* Validation (K-Fold CV). Pada K-Fold CV, terdapat nilai *k* yang didefinisikan secara manual. Pada penelitian ini, nilai *k* yang digunakan adalah 10 (10-Fold CV). Hal tersebut dikarenakan k=10 merupakan nilai standar untuk evaluasi. Banyak hasil eksperimen yang menunjukkan bahwa k=10 merupakan pilihan terbaik untuk memperoleh estimasi yang akurat. Perbandingan nilai metode CV ditunjukkan pada Tabel 6 berikut:

Tabel 6. Perbandingan Nilai Akurasi, Presisi, dan *Recall* Metode CV

Metode	Akurasi	Presisi	*Recall*
k-Nearest Neighbor	0,9967	0,9898	0,9977
Naive Bayes	0,8396	0,7776	0,6134
Support Vector Machine	0,9338	0,8898	0,9795
Tuning (*Decision Tree*)	1,0000	1,0000	1,0000
ID3	0,8857	0,8685	0,8564
Logistic Regression	0,8736	0,7807	0,7785
Linear Discriminant Analysis	0,8867	0,7898	0,7539
ANN (Artificial Neural Network)	1,0000	1,0000	1,0000
CNN (Convolution Neural Network)	1,0000	0,9726	0,9425
Extra Trees Classifier	1,0000	1,0000	1,0000
AdaBoost	1,0000	1,0000	1,0000
VFI 5	0,8952	0,8678	0,8404

Metode klasifikasi *Decision Tree* (*Model Tuning*), *Artificial Neural Network* (ANN), AdaBoost, dan *Extra Trees Classifier* menghasilkan nilai akurasi sebesar 1,0000 ketika menerapkan CV sebagai metode pembagian data terhadap *Mushroom Data Set*. Hal tersebut berarti bahwa ketiga metode tersebut dapat mengklasifikasikan jenis kelas jamur dengan akurasi model sebesar 100%. Data yang sederhana dan kelas data yang tidak sama memungkinkan model mudah mempelajari pola yang ada dengan akurasi 100%. Untuk data yang lebih kompleks perlu dikaji ulang.

Data yang sederhana dan kelas data yang tidak seimbang memungkinkan model machine learning untuk dengan mudah mempelajari pola yang ada dengan akurasi tinggi, bahkan hingga 100% [27, 28]. Penelitian ini mendukung temuan sebelumnya yang menunjukkan bahwa model klasifikasi cenderung memberikan hasil akurasi yang tinggi pada dataset yang memiliki fitur-fitur yang tidak terlalu kompleks dan distribusi kelas yang tidak merata, terutama ketika digunakan teknik *preprocessing* yang tepat [29].

G. Perbandingan Nilai Akurasi *Hold Out* dan *Cross Validation Method*

Setelah dilakukan perbandingan nilai akurasi, presisi, dan *recall* dari setiap metode pembagian data (*Hold Out* dan CV), selanjutnya ditampilkan perbandingan (Tabel 7) nilai akurasi dari metode *Hold Out* dan CV secara berdampingan.

Tabel 7. Perbandingan Nilai Akurasi Metode *Hold Out* dan CV

Metode Klasifikasi	Akurasi	
	Hold Out	CV
k-Nearest Neighbor	1,0000	0,9967
Naive Bayes	0,8763	0,8396
Support Vector Machine	0,9686	0,9338
Tuning (*Decision Tree*)	1,0000	1,0000
ID3 (Iterative Dechotomser Tree)	0,9787	0,8857
Logistic Regression	0,9329	0,8736
Linear Discriminant Analysis	0,9003	0,8867

ANN (Artificial Neural Network)	1,0000	1,0000
CNN (Convolution Neural Network)	0,9979	1,0000
Extra Trees Classifier	1,0000	0,8396
AdaBoost (Adaptive Boosting)	1,0000	1,0000
VFI 5 (Voting Feature Intervals 5)	0,9032	0,8952

Analisis perbandingan nilai akurasi dari setiap metode:

1) ***k-Nearest Neighbor* (k-NN):** Hold Out menghasilkan akurasi sempurna (1.0000), sedangkan CV sedikit lebih rendah (0.9967). Ini menunjukkan bahwa pada Hold Out, model mungkin memanfaatkan pembagian data secara optimal, tetapi ada sedikit generalisasi yang lebih realistis dengan CV.
2) ***Naive Bayes*:** Ada penurunan akurasi dari 0.8763 (Hold Out) menjadi 0.8396 (CV). Ini menunjukkan bahwa Naive Bayes mungkin terlalu dioptimalkan untuk subset spesifik dalam Hold Out, sementara CV memberikan gambaran yang lebih generalis.
3) ***Support Vector Machine* (SVM):** Terdapat penurunan dari 0.9686 menjadi 0.9338 di CV. Hal ini menunjukkan bahwa SVM mungkin menunjukkan kinerja yang sedikit lebih buruk dalam validasi silang, tetapi masih merupakan model yang kuat.
4) ***Decision Tree* (Tuning):** Akurasi tetap 1.0000 pada kedua metode. Ini dapat menunjukkan bahwa model Decision Tree berhasil mempelajari semua pola dalam data, tetapi hasil ini bisa juga merupakan indikasi overfitting, terutama jika dataset terlalu sederhana atau tidak bervariasi.
5) ***ID3* (*Iterative Dichotomiser Tree*):** Akurasi Hold Out (0.9787) lebih tinggi dibandingkan CV (0.8857), menunjukkan bahwa model mungkin terlalu cocok dengan data dalam pembagian Hold Out, tetapi gagal generalisasi ketika diuji menggunakan CV.
6) ***Logistic Regression*:** Terdapat penurunan yang cukup signifikan dari 0.9329 (Hold Out) ke 0.8736 (CV), yang menunjukkan bahwa Logistic Regression mungkin lebih sensitif terhadap variasi dalam dataset, dengan hasil CV yang lebih realistis.
7) **LDA (*Linear Discriminant Analysis*):** Hasilnya cukup konsisten dengan sedikit penurunan, dari 0.9003 (Hold Out) menjadi 0.8867 (CV), menunjukkan stabilitas yang cukup baik dalam kedua metode.
8) ***Artificial Neural Network* (ANN):** Akurasi tetap 1.0000 pada kedua metode, yang mungkin mengindikasikan bahwa ANN sangat cocok untuk dataset ini atau mungkin overfitting jika dataset kecil.
9) ***Convolutional Neural Network* (CNN):** Pada Hold Out, akurasi sedikit di bawah sempurna (0.9979), sementara CV memberikan akurasi sempurna (1.0000). Ini menunjukkan bahwa CNN bisa saja memiliki potensi generalisasi yang lebih baik ketika diuji dengan cross-validation.
10) ***Extra Trees Classifier*:** Pada Hold Out, akurasi sempurna (1.0000), tetapi pada CV turun drastis menjadi 0.8396, menunjukkan adanya overfitting pada Hold Out yang tidak terdeteksi hingga model diuji lebih menyeluruh dengan CV.
11) **AdaBoost:** Seperti beberapa model lain, AdaBoost menghasilkan akurasi sempurna (1.0000) pada kedua metode, menunjukkan kinerja yang sangat kuat, meski ada potensi overfitting jika dataset tidak cukup kompleks.
12) **VFI 5 (*Voting Feature Intervals*):** Akurasi di Hold Out adalah 0.9032, dan di CV sedikit turun menjadi 0.8952. Hasil ini menunjukkan bahwa model ini cukup stabil dan tidak terlalu sensitif terhadap pembagian data.

Berdasarkan hasil pengujian, metode klasifikasi *Decision Tree* (*Model Tuning*), AdaBoost, dan *Artificial Neural Network* (ANN) merupakan metode terbaik untuk mengklasifikasikan *Mushroom Data Set*, baik pembagian data dengan menerapkan CV dan *Hold Out*. Metode tersebut sama-sama memiliki akurasi tertinggi, yaitu sebesar 100%. Temuan ini konsisten dengan penelitian sebelumnya oleh Smith *et al* [30], yang menunjukkan bahwa metode *Decision Tree* dan *ensemble methods* seperti *AdaBoost* sangat efektif dalam mengklasifikasikan jamur beracun dan tidak beracun pada dataset yang serupa. Mereka melaporkan akurasi tinggi pada dataset yang sederhana, terutama ketika algoritma model tuning diterapkan dengan baik.

Namun, studi lain oleh Johnson, Lee, dan Kim [31] menemukan bahwa pada dataset yang memiliki lebih banyak variasi atribut atau noise, metode seperti ANN cenderung *overfitting*, terutama tanpa langkah pengendalian *overfitting* yang ketat. Hasil penelitian ini menekankan perlunya evaluasi lebih lanjut terhadap *overfitting*, meskipun akurasi awalnya tampak sangat tinggi.

Sementara itu, penelitian dari Wang, Zhou, dan Liu [32] menunjukkan bahwa walaupun metode klasifikasi seperti *Decision Tree* dan *ANN* memiliki akurasi tinggi pada dataset yang terstruktur dengan baik, mereka juga menekankan pentingnya *balancing* dataset untuk menghindari *overfitting*, terutama pada dataset dengan distribusi kelas yang tidak seimbang. Oleh karena itu, dalam studi semacam ini juga diperlukan analisis mendalam mengenai potensi *overfitting* untuk memastikan bahwa model yang dihasilkan dapat bekerja secara general di luar data latihnya.

IV. KESIMPULAN

Berdasarkan hasil analisis yang telah dilakukan, dapat disimpulkan bahwa analisis klasifikasi terhadap *Mushroom Data Set* menghasilkan nilai akurasi tertinggi apabila menggunakan metode klasifikasi *Decision Tree* (*Model Tuning*), AdaBoost dan *Artificial Neural Network* (ANN), baik pembagian data berbasis *Hold Out* atau *Cross Validation*. Namun akan lebih baik jika pembagian data menerapkan metode *Cross Validation* karena dapat menghindari adanya *overfitting* (tumpang tindih) pada data *testing*, meskipun waktu yang diperlukan untuk *running data* sedikit lebih lama dibandingkan dengan metode *Hold Out*. *Hold Out* sering kali memberikan akurasi yang lebih tinggi, tetapi rentan terhadap *overfitting*, terutama jika dataset kecil atau tidak bervariasi. *Cross Validation* memberikan gambaran yang lebih andal tentang kemampuan generalisasi model, terutama ketika data uji diambil dari berbagai subset.

Dalam kasus ini, beberapa metode seperti k-NN, *Decision Tree*, ANN, AdaBoost, ANN, *Extra Trees Classifier*, dan CNN menunjukkan hasil sempurna di salah satu atau kedua metode, yang mungkin menunjukkan potensi *overfitting*, terutama jika dataset tidak cukup kompleks, kelas data tidak berimbang atau ada fitur yang secara langsung mempengaruhi klasifikasi. Oleh karena itu penelitian serupa dapat menggunakan metode deteksi *overvitting* tambahan lainnya untuk lebih memastikan ada atau tidaknya *overvitting* selain kedua metode tersebut.

V. UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Gunadarma atas semua dukungan yang diberikan dalam penelitian ini. Kontribusi tersebut memungkinkan penulis untuk mengumpulkan data dan menganalisis temuan yang telah disajikan dalam artikel ini. Penelitian ini tidak akan berhasil tanpa dukungan dari Universitas Gunadarma. Penulis juga ingin menyampaikan terima kasih kepada Dr. Karmilasari, S.Kom., M.M. atas saran dan panduan akademik yang berharga dalam penyusunan artikel ini.

REFERENSI

[1] Kempenaar, C. *et al*., "Big data analysis for smart farming", *s.l.: Wageningen University and Research (Vol. 655)*. 2016.

[2] Lokers, R. et al., "Analysis of Big Data technologies for use in agro-environmental science". *Environmental Modelling and Software, Volume 84 , pp. 494-504*. 2016.

[3] Sonka, S., "Big Data: Fueling the Next Evolution of Agricultural Innovation". *Journal of Innovation Management, 4(1), pp. 114-36.* 2016.

[4] Waga, D. a*nd* Rabah, K., "Environmental conditions' big data management and cloud computing analytics for sustainable agriculture". *World Journal of Computer Application and Technology, 2(3), pp. 73-81*. 2014.

[5] Putro, B. C. S., Mustika, I. W., a*nd* Nugroho, L. E. Optimized Backpropagation Artificial Neural Network Algorithm for Smart Agriculture Applications," *Proc. - 2018 4th Int. Conf. Sci. Technol. ICST 2018*. 2018.

[6] Kim, G.-H., Trimi, S. a*nd* Chung, J.-H., "Big-data applications in the government sector". *Communications of the ACM, 57(3), pp. 78-85.* 2014.

[7] Heriyanto, H. "Urgensi Penerapan EGovernment Dalam Pelayanan Publik," Musamus *Journal of Public Administration, 4(2), 066–075*. https://doi.org/10.35724/mjpa.v4i2.4128. 2022.

[8] Bunge, J., "Big data comes to the farm, sowing mistrust: seed makers barrel into technology Business", *s.l.: Wall Street Journal (Online)*. 2014.

[9] Ganesan, M., Andavar, S., a*nd* Raj, R. S. P. Prediction of Land Suitability for Crop Cultivation Using Classification Techniques. *Brazilian Archives of Biology and Technology, 64*. https://doi.org/10.1590/1678-4324-2021200483. 2021.

[10] Budianto, I., Maidah, N. El, a*nd* Bukhori, S. Classification of soil types based on suitable plants using multiclass classification artificial neural network. *Classification of Soil Types Based on Suitable Plants Using Multiclass Classification Artificial Neural Network, 11(3)*. 2023.

[11] Fhonna, R. P., Afrillia, Y., Zulfan, Aqmal, J., a*nd* Abadi, S. Klasifikasi Penentuan Jenis Tanah yang Sesuai Terhadap Tanaman Pangan Sebagai Solusi Ketahanan Pangan di Kabupaten Pidie Jaya Menggunakan Metode Random Forest. *Jurnal Informasi Dan Teknologi, 12–18*. https://doi.org/10.60083/jidt.v5i4.402. 2023.

[12] Khan, A. R. and Nisha, S. S. Comparison Of Classification Techniques Using Mushroom Datasets*. in Sadakath Research Bulletin, Vol. 0 No.0 FeB-2018*. 2018.

[13] Yang, X. S. *Introduction to Algorithms for Data Mining and Machine Learning*. London: Elsevier, Inc. 2019.

[14] Beniwal, S. and Das, B. Mushroom Classification Using Data Mining Techniques. *International Journal of Pharma and Bio Sciences, 6(1): (B)*, 2015, pp. 1170-1176. 2015.

[15] Lincoff, G. H. *The AuduBon Society Field Guide to North American Mushrooms*. New York: Alfred A. Knopf Press. 1981.

[16] Ulloa M, Richard TH., *Illustrated Dictionary of Mycology*. Minnesota: APS Press. 2001.

[17] Hall, A. H., Spoerke, D. G., and Rumack, B. H., "Mushroom Poisoning: Identification, Diagnosis, and Treatment". *Pediatrics In Review, Vol. 8 No.10 April-1987*. 1987.

[18] Han, J. Kamber, M., a*nd* Pei, J. *Data Mining Concept and Techniques*. Elsevier. 2012.

[19] Gede, A., Pradnyana, S., Kom, M., Kom, K., Agustini, S., a*nd* Si, M. S. *Konsep Dasar Data Mining*. 2017.

[20] Witten, Ian H. , Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques . Elsevier publishing. 2005. [20]

[21] Arifin, O., Saputra, K., a*nd* Fathoni, H. Implementation of Data Mining using Naïve Bayes Classifier Method in Food Crop Prediction. *Scientific Journal of Informatics, 8(1), 43–50*. https://doi.org/10.15294/sji.v8i1.28354. 2021.

[22] Zhang, H., Li, Y., a*nd* Wang, J. Decision tree models for crop classification: A review of applications and advances. *Computers and Electronics in Agriculture, 175, 105607*. 2021.

[23] Kumar, R., Gupta, A., a*nd* Singh, S. Plant disease detection using support vector machine. *International Journal of Engineering Research and Technology, 9(5), 55-62*. 2020.

[24] Patel, R., Verma, V., a*nd* Gupta, R. Applications of artificial neural networks in agriculture: A review. *Journal of Emerging Technologies and Innovative Research, 6(4), 109-118*. 2019.

[25] Mishra, P., a*nd* Sinha, P. Crop classification using remote sensing data and machine learning techniques: A survey. *Agricultural Informatics, 6(1), 47-60*. 2020.

[26] Gupta, P., a*nd* Sharma, A. A comparative analysis of machine learning algorithms for crop yield prediction. *Journal of Agricultural Science and Technology, 15(3), 105-112*. 2018.

[27] Chen, Y., Zhou, X., Wang, J., and Xu, H. Improving classification performance on imbalanced datasets using hybrid approaches. *J. Data Mining Knowl. Discov.*, vol. 35, no. 2, pp. 234-247. 2023.

[28] Zhang, Q. and Li, K. Handling class imbalance with deep learning models in data mining tasks. *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1257-1270. 2022.

[29] Liu, M., Zhao, Y., and Wang, F. Exploring feature simplicity and class imbalance in machine learning models. *Expert Syst. Appl.*, vol. 171, p. 114587. 2021.

[30] Smith, A., Clarke, B., Gomez, C., a*nd* Martin, D. Classification of toxic and non-toxic mushrooms using ensemble learning techniques. *J. Data Mining Knowl. Discov.*, vol. 38, no. 4, pp. 1150–1165. 2023.

[31] Johnson, D., Lee, S., and Kim, H. Overfitting in neural networks: Case studies and practical approaches in data mining tasks. *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 1556–1570. 2022.

[32] Wang, X., Zhou, Q., and Liu, L. Handling imbalanced datasets in machine learning classification: A review of methods and applications. *Expert Syst. Appl.*, vol. 183, p. 115324. 2021.