JSI : Jurnal Sistem Informasi *(E-Journal), VOL. 17, NO. 1, April 2025*
ISSN Print : 2085-1588
ISSN Online :2355-4614
LINK: https://jsi.ejournal.unsri.ac.id/index.php/jsi/index

Algoritma Iterative Dichotomizer 3 (ID3) dan Extreme Gradient Boosting (XGBOOST) untuk Memprediksi Penyakit Polycystic Ovarian Syndrome (PCOS)

Dina Mustaqima[1] , Siti Husnul Hotimah[2] , Anita Desiani[3], Indri Ramayanti[4]
[1,2,3] Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Sriwijaya

e-mail: 08011182227052@student.unsri.ac.id[1] , 08011282227028@student.unsri.ac.id[2], anita_desiani@unsri.ac.id[3] , indri_ramayanti@um-palembang.ac.id[4]

Abstrak

Polycystic Ovary Syndrome (PCOS) adalah kelainan hormon reproduksi wanita yang dapat dideteksi dengan membandingkan algoritma iterative dichotomiser 3 (ID3) dan extreme gradient boosting (XGBoost). Perbandingan antara kedua algoritma ini memiliki tujuan untuk mengidentifikasi algoritma yang paling efektif dalam mengklasifikasikan penyakit PCOS. Pengujian kedua algoritma dilakukan menggunakan dua metode, yaitu teknik pemisahan persentase dan validasi silang k-fold. Teknik pemisahan persentase, dataset dipecah menjadi 80% untuk proses pelatihan serta 20% untuk proses pengujian. Teknik validasi silang k-fold menggunakan nilai k sebanyak 10, dimana data dibagi menjadi sepuluh bagian yang digunakan secara bergantian untuk pelatihan dan pengujian. Hasil perbandingan menujukkan bahwa teknik validasi silang k-fold memberikan kinerja yang lebih unggul dibandingkan dengan teknik pemisahan persentase, karena teknik ini secara signifikan meningkatkan nilai presisi, sensitivitas, dan ketepatan akurasi dari setiap algoritma yang dievaluasi. XGBoost mencapai presisi sebesar 88%, sensitivitas 87%, dan tingkat akurasi 89% setiap algoritma yang diuji. Berdasarkan hasil ini dapat disimpulkan bahwa penggunaan validasi silang k-fold menjadikan XGBoost sebagai algoritma yang lebih efektif dalam mengindetifikasi PCOS, dibandingkan dengan algoritma dan teknik pengujian lainnya.

Abstract

Polycystic Ovary Syndrome (PCOS) is a female reproductive hormone disorder that can be detected by comparing the iterative dichotomiser 3 (ID3) and extreme gradient boosting (XGBoost) algorithms. The comparison between these two algorithms aims to identify the most effective algorithm in classifying PCOS disease. The testing of the two algorithms was carried out using two methods, namely the percentage split technique and k-fold cross-validation. Percentage split technique, the dataset is broken down into 80% for the training process and 20% for the testing process. The k-fold cross-validation technique uses a k-value of 10, where the data is divided into ten parts that are used alternately for training and testing. The comparison results show that the k-fold cross-validation technique provides superior performance compared to the percentage separation technique, as it significantly improves the precision, sensitivity, and accuracy accuracy values of each evaluated algorithm. XGBoost achieved 88% precision, 87% sensitivity, and 89% accuracy rate of each algorithm tested. Based on these results, it can be concluded that the use of k-fold cross-validation makes XGBoost a more effective algorithm in identifying PCOS, compared to other algorithms and testing techniques.

Keywords*: PCOS; ID3; XGBoost; pemisahan persentase; validasi silang k-fold.*

1. PENDAHULUAN

Polycystic Ovary Syndrome (PCOS) adalah kelainan hormonal yang paling sering dialami oleh wanita di usia subur [1]. PCOS umumnya disebabkan oleh ketidakseimbangan hormon reproduksi, yang mengakibatkan terbentuknya banyak kista kecil di tepi ovarium. Faktor-faktor yang memicu terjadinya PCOS dapat melibatkan

faktor genetik dan gaya hidup, pola hidup yang tidak sehat serta pola makan yang tidak teratur dan tidak sehat juga dapat meningkatkan risiko munculnya PCOS. Di Indonesia, prevalensi PCOS berkisar antara 1,8% hingga 15%, tergantung pada faktor etnis, latar belakang individu, serta kriteria diagnostik yang diterapkan [2]. Mengingat tingginya jumlah kasus PCOS dan dampak PCOS terhadap kesehatan reproduksi, metabolisme, dan psikologis wanita, deteksi dini dan klasifikasi untuk penyakit PCOS menjadi sangat penting. Klasifikasi penyakit PCOS dapat dilakukan dengan berbagai metode dalam *data mining*, diantara metode yang dapat digunakan yaitu algoritma *Iterative Dichotomiser* 3 (ID3) dan *algoritma Extreme Gradient Boosting* (XGBoost).

Algoritma *Iterative Dichotomiser* 3 (ID3) merupakan metode yang berfungsi untuk menghasilkan pohon keputusan dalam menganalisis data [3]. Kelebihan dari algoritma ID3 mampu membangun pohon keputusan secara efisien, dengan hanya memerlukan sedikit pengujian atribut hingga seluruh data berhasil terklasifikasi [4]. Algoritma ID3 juga mampu mengeliminasi perhitungan-perhitungn yang tidak diperlukan, sehingga pohon keputusan yang terbentuk oleh algortima ID3 menjadi lebih sederhana dan efektif dalam melakukan klasifikasi [5]. Algoritma ID3 mempunyai kelemahan ketika dilakukan pembentukan cabang atau node yang tidak diperlukan dapat membuat ukuran pohon keputusan menjadi terlalu besar, yang dikenal dengan istilah *over-fitting*. *Over-fitting* ini dapat menghasilkan galat klasifikasi yang dapat menurunkan tingkat keakuratan klasifikasi [6]. Klasifikasi PCOS menggunakan ID3 belum pernah dilakukan, penelitian yang pernah dilakukan mengguakan algoritma ID3 adalah klasifikasi penyakit campak [7] dengan hasil akurasi sebesar 89,83%. Penelitian lain mengenai klasifikasi penyakit diabetes melitus [8] menunjukkan hasil akurasi sebesar 90%. Penelitian lainnya melakukan klasifikasi diagnosis penyakit COVID-19 [9] dengan hasil nilai akurasi sebesar 90%.

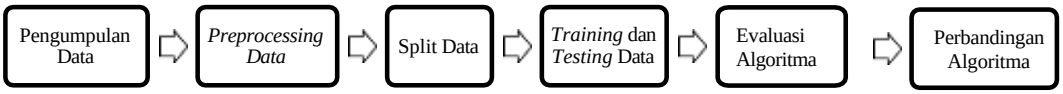
Algoritma *Extreme Gradient Boosting* (XGBoost) merupakan algoritma *machine learning* yang menggunakan pohon keputusan sebagai pengklasifikasi dengan dasar dari metode *gradient boosting* [10]. Algoritma *Extreme Gradient Boosting* (XGBoost) memiliki formalisasi model yang lebih terstruktur untuk mengendalikan *over-fitting*, dengan demikian mampu menutupi kelemahan algortima ID3 [11]. Kelemahan pada algoritma pengklasifikasi seperti ID3 dapat diatasi dengan menggunakan algoritma pengklasifikasi ganda salah satunya adalah XGBoost. Kekurangan algoritma XGBoost adalah kompleksitas modelnya, yang berkaitan dengan banyaknya parameter hal tersebut membuat proses penentuan nilai optimal untuk setiap parameter menjadi sulit dan memakan waktu [12]. Klasifikasi PCOS menggunakan XGBoost belum pernah dilakukan, berbagai penelitian yang telah dilakukan dengan memanfaatkan algoritma yaitu klasifikasi penyakit diabetes [13]dengan nilai akurasi sebesar 90.10%, klasifikasi penyakit parkinson [14] dengan nilai akurasi sebesar 93.66%, dan klasifikasi penyakit jantung [15] dengan nilai akurasi sebesar 84.74%.

Penelitian ini bertujuan untuk membandingkan kinerja algoritma ID3 dan XGBoost dalam klasifikasi penyakit PCOS, dalam pengujian kedua algoritma menggunakan dua metode yaitu teknik pemisahan persentase dan teknik validasi silang k-fold. Teknik pemisahan persentase, data dipecah menjadi 80% untuk proses pelatihan dan 20% untuk proses pengujian. Teknik validasi silang k-fold diterapkan dengan nilai k sebanyak 10, dimana data dipecah menjadi sepuluh bagian yang digunakan secara

bergantian untuk pelatihan dan pengujian. Hasil dari penelitian mengidentifikasi algoritma dan teknik pengujian yang efektif dalam mengklasifikasi penyakit PCOS berdasarkan nilai akurasi, presisi, dan sensitivitas, sehingga dapat memberikan manfaat dalam diagnosis dan penanganan penyakit PCOS.

2. METODE PENELITIAN

Langkah-langkah yang dilaksanakan dalam penelitian disajikan pada Gambar 1.



Gambar 1. Rangkaian proses yang dilakukan dalam Penelitian

Pada Gambar 1 menjelaskan langkah-langkah yang dilaksanakan dalam penelitian yang terdiri dari pengumpulan data, *preprocessing* data, split data, *training* dan *testing* data, evaluasi algortima, dan perbandingan algoritma.

2.1 Deskripsi Data

Data yang digunakan dalam penelitian diambil dari dataset klasifikasi penyakit PCOS yang dapat diakses di situs kaggle [16]. Dataset tersebut mencakup 541 data dengan 44 atribut, dimana target pada dataset ini mencerminkan kondisi yang terbagi menjadi dua kelas yaitu YES (positif PCOS) dan NO (negatif PCOS). Atribut-atribut yang digunakan dalam penelitian ini mencakup berbagai parameter fisik dan klinis untuk menentukan masalah terkait PCOS. Penjelasan mengenai atribut-atribut yang digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Atribut dan Deskripsinya

Nama Atribut	**Tipe Atribut**	**Nilai**	***Missing Value***
Age	Numerik	Umur (20 - 48 tahun)	*Null*
Weight	Numerik	Berat Badan (31 - 108 kg)	*Null*
Height	Numerik	Tinggi Badan (137 - 180 cm)	*Null*
BMI	Numerik	Indeks Massa Tubuh (12.417882-38.9)	*Null*
Blood Group	Nominal	Golongan darah (11 = A+, 12 = A-, 13 = B+, 14 = B-, 15 = O+, 16 = O-, 17 = AB+, 18 = AB-)	*Null*
Pulse Rate	Numerik	Denyut nadi (13 - 82 bpm)	*Null*
RR	Numerik	Laju pernapasan (16 - 28 kali/menit)	*Null*
HB	Numerik	Hemoglobin (8.5 - 14.8 g/dL)	*Null*
Cycle	Numerik	Siklus menstruasi (2 - 5 hari)	*Null*

Cycle Length	Numerik	Panjang siklus menstruasi (0 - 12 hari)	*Null*
Marraige Status	Numerik	Lama pernikahan (0 - 30 tahun)	*Null*
Pregnant	Nominal	Kehamilan (0 = NO, 1= YES)	*Null*
No. Of Aborptions	Numerik	Jumlah aborsi (0 - 5)	*Null*
I Beta-HCG	Numerik	Kadar beta-HCG (1.3 - 32460.97 mIU/mL)	*Null*
FSH	Numerik	Follicle-Stimulating Hormone (0.21 - 5052.0 mIU/mL)	*Null*
LH	Numerik	Luteinizing Hormone (0.02 - 2018.0 mIU/mL)	*Null*
FSH/LH	Numerik	Rasio FSH/LH (0.002146 - 1372.826087)	*Null*
Hip	Numerik	Lingkar pinggul (26 - 48 inci)	*Null*
Waist	Numerik	Lingkar pinggang (24 - 47 inci)	*Null*
Waist : Hip Ratio	Numerik	Rasio lingkar pinggang ke pinggul (0.755556 - 0.979167)	*Null*
TSH	Numerik	Thyroid-Stimulating Hormone (0.04 - 65.0 mIU/L)	*Null*
PRL	Numerik	Prolaktin (0.4 - 128.24 ng/mL)	*Null*
Vit D3	Numerik	Vitamin D3 (0 - 6014.66 ng/mL)	*Null*
PRG	Numerik	Progesteron (0.047 - 85 ng/mL)	*Null*
RBS	Numerik	Random Blood Sugar (60 - 350 mg/dL)	*Null*
Weight Gain	Nominal	Kenaikan berat badan (0 = NO, 1= YES)	*Null*
Hair Growth	Nominal	Pertumbuhan rambut berlebih (0 = NO, 1= YES)	*Null*
Skin Darkening	Nominal	Penggelapan kulit (0 = NO, 1= YES)	*Null*
Hair Loss	Nominal	Rambut rontok (0 = NO, 1= YES)	*Null*
Pimples	Nominal	Jerawat (0 = NO, 1= YES)	*Null*
Fast Food	Nominal	Konsumsi makanan cepat saji (0 = NO, 1= YES)	*Null*
Reg. Exercise	Nominal	Olahraga teratur (0 = NO, 1= YES)	*Null*
BP Systolic	Numerik	Tekanan darah sistolik [12, 140]	*Null*
BP Diastolic	Numerik	Tekanan darah diastolik [8, 100]	*Null*
Follicle NO.	Numerik	Jumlah folikel (0 - 20)	*Null*

Follicle NO. .1	Numerik	Jumlah folikel (0 - 20)	*Null*
Avg. F Size	Numerik	Ukuran rata-rata folikel (0 - 24 mm)	*Null*
Avg. F Size. 1	Numerik	Ukuran rata-rata folikel (0 - 24 mm)	*Null*
Endometrium	Numerik	Ketebalan Endometrium (0 - 18 mm)	*Null*

2.2 Praproses Data

Praproses data adalah serangkaian langkah sebelum analisis dengan tujuan untuk membersihkan, menormalkan, dan mempersiapkan data agar optimal untuk algoritma analisis [17]. Tabel 1 menyajikan data tentang penyakit PCOS yang terdiri dari 44 atribut. Atribut - atribut yang tidak relevan akan dihapus, sehingga jumlah atribut yang semula 44 berkurang menjadi 40. Atribut atau kolom kondisi berfungsi sebagai label keputusan yang menunjukkan status penderita, sementara 39 kolom lainnya menggambarkan kriteria yang digunakan untuk menentukan apakah seseorang terdiagnosis PCOS atau tidak.

a. *Missing Value*

Missing value dapat terjadi karena berbagai alasan, seperti kesalahan dalam proses pengumpulan data, masalah teknis, atau responden yang tidak memberikan informasi. Pada dataset penyakit PCOS tidak ada atribut yang mempunyai data yang kosong didalamnya, sehingga tidak dilakukan *missing value*.

b. Normalisasi Data

Data penyakit PCOS yang terdapat pada Tabel 1 terlihat bahwa beberapa atribut memiliki rentang nilai yang sangat berbeda, sehingga memerlukan normalisasi data. Normalisasi dilakukan untuk menyamakan rentang nilai antar atribut sehingga hasil analisis lebih akurat dan waktu komputasi model dapat diminimalkan. Proses normalisasi data dapat dilakukan dengan menggunakan penskalaan *min-max* dengan rumus sebagai berikut [18].

$$Data\ (x) = \frac{(x - minValue)\ (maxRange - minRange)}{maxValue - minValue} + minRange \quad (1)$$

Dimana $Data\ (x)$ merujuk pada nilai data baru hasil normalisasi, x adalah data yang dinormalisasi, $min\ Value$ adalah nilai terkecil dari satu kolom atau baris, $max\ Value$ adalah nilai terbesar dari satu kolom atau baris, $min\ Range$ adalah batas nilai terkecil dari normalisasi, dan $max\ Range$ adalah batas nilai terbesar dari normalisasi.

c. Pengujian Data

Pada pengujian data diuji menggunakan teknik pemisahan persentase dan teknik validasi silang k-fold. Teknik pemisahan persentase dalam *data mining* merupakan metode untuk membagi dataset menjadi dua bagian yaitu pelatihan (*training set*) dan pengujian (*testing set*) [19]**.** Metode pemisahan persentase,

data dipecah menjadi 80% untuk pelatihan dan 20% untuk pengujian, sedangkan dalam teknik validasi silang k-fold, digunakan nilai k sebanyak 10, dimana data dipecah menjadi sepuluh bagian yang digunakan secara bergantian untuk pelatihan dan pengujian.

2.3 Algoritma ID3

Algoritma ID3 merupakan algoritma yang bertujuan untuk membnetuk pohon keputusan dengan menggunakan pencarian optimal untuk menguji atribut di setiap node, dengan menghitung nilai *entropy* dan *information gain*. Secara singkat, langkah-langkah kerja algoritma ID3 dapat dijelaskan sebagai berikut [20] :

1. Melakukan inputan sampel data untuk *training*, label *training*, dan atributnya.
2. Melakukan pemilihan atribut dengan nilai *information gain* tertinggi. Proses ini dimulai dengan menghitung *entropy* dan *information gain* untuk setiap atribut menggunakan rumus berikut:
   $$\text{Entropy}(S) = -P_+ \log 2\, P_+ - -P_- \log 2\, P_- \quad (2)$$
   Dimana S sebagai data sampel yang dipakai untuk *training*, P_+ adalah probabilitas dari kelas yang positif yang diperoleh dengan membagi jumlah kasus positif dengan total kasus, dan P_- adalah probabilitas dari kelas yang negatif yang diperoleh dengan membagi jumlah kasus negatif dengan total kasus.

   $$Gain(S, A) = Entropy(S) - \sum_{v \in nilai(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (3)$$
   Dengan S sebagai data sampel yang digunakan untuk *training*, A merupakan atribut dalam *dataset*, V adalah nilai yang mungkin untuk atribut A, nilai(A) adalah himpunan nilai yang dimilki oleh atribut A, |Sv| menunjukkan jumlah sampel untuk nilai V, |S| adalah jumlah total sampel dalam dataset, dan $Entropy(S_v)$ adalah *entropy* untuk sampel-sampel yang memiliki nilai V.
3. Pilih atribut dengan *information gain* terbesar, lalu buat simpul yang berisi atribut tersebut.
4. Proses ini diulangi dengan terus menghitung *information gain* hingga seluruh data terklasifikasikan ke dalam satu kelas yang sama. Atribut yang sudah dipilih tidak akan dihitung lagi dalam perhitungan *information gain* pada iterasi berikutnya.

2.4 Algoritma XGBoost

Algoritma *Extreme Gradient Boosting* (XGBoost) adalah model algoritma pembelajaran mesin yang kuat dan efisien yang digunakan untuk masalah klasifikasi [21]. Algoritma XGBoost menggunakan pendekatan *ensemble learning*, dimana sejumlah pohon keputusan dibangun secara berurutan dan setiap pohon keputusan memperbaiki kesalahan prediksi dari pohon sebelumnya. Secara ringkas langkah-langkah dalam membangun pohon XGBoost adalah sebagai berikut [22] :

1. Menginisialisasi probabilitas awal dari prediksi Pr_i^1, dengan $i = 1, 2, \ldots, n$

   Dimana Pr_i^1 merupakan probabilitas prediksi untuk sampel data ke- i pada iterasi ke- t.

2. Mengitung selisih (residual) menggunakan rumus berikut:

$$Residual_i^t = Y_i - Pr_i^t \quad (4)$$

Dimana $Residual_i^t$ merupakan residual atau selisih antara nilai target aktual Y_i dan nilai prediksi Pr_i^t pada iterasi ke- t.

3. Menghitung nilai cover dari atribut menggunakan rumus berikut:

$$Cover(A) = \sum_{i=1}^{n}\left(Pr_i^t\,(1 - Pr_i^t)\right) \quad (5)$$

Dimana $Cover(A)$ menggambarkan seberapa banyak variasi dalam probabilitas prediksi yang dihasilkan oleh atribut A. Pr_i^1 merupakan probabilitas prediksi untuk sampel data ke- i pada iterasi ke- t. n adalah jumlah total sampel dalam dataset. Dan A adalah atribut yang sedang dievaluasi.

4. Menghitung skor kemiripan (SS) menggunakan rumus berikut:

$$SS_{node} = \frac{(\sum_{i=1}^{n} Residual_i)^2}{\sum_{i=1}^{n}\left(Pr_i^t\,(1 - Pr_i^t)\right) + \lambda} \quad (6)$$

Dimana SS_{node} adalah skor kemiripan pada node tersebut. n adalah jumlah total sampel dalam dataset. Pr_i^1 merupakan probabilitas prediksi untuk sampel data ke- i pada iterasi ke- t. $Residual_i$ residual atau selisih antara nilai target aktual dan nilai prediksi saat ini untuk sampel ke- i. λ adalah parameter regularisasi yang digunakan untuk mengendalikan kompleksitas model.

5. Hitung nilai gain atribut dengan rumus berikut:

$$Gain\ (A) = SS_{left} + SS_{right} - SS_{root} \quad (7)$$

Dimana $Gain\ (A)$ menggambarkan peningkatan dalam pemisahan kelas target yang dihasilkan oleh pemisahan menggunakan atribut A. SS_{left} adalah skor kemiripan pada node anak kiri (setelah pemisahan menggunakan atribut A. SS_{right} adalah skor kemiripan pada node anak kanan (setelah pemisahan menggunakan atribut A. Dan SS_{root} adalah skor kemiripan pada node induk (sebelum pemisahan menggunakan atribut A.

6. Hitung nilai output dengan rumus berikut:

$$Output\ (A)_i = \frac{\sum_{i=1}^{n} Residual_i}{\sum_{i=1}^{n}(Pr_i\,(1 - Pr_i)) + \lambda} \quad (8)$$

Dimana $Output\ (A)_i$ memberikan gambaran tentang kontribusi atribut A terhadap pembaruan probabilitas pada setiap iterasi dalam pembangunan pohon XGBoost. n adalah jumlah total sampel dalam dataset. Pr_i^1 merupakan probabilitas prediksi untuk sampel data ke- i pada iterasi ke- t. $Residual_i$ residual atau selisih antara nilai target aktual dan nilai prediksi saat ini untuk sampel ke- i. λ adalah parameter regularisasi yang digunakan untuk mengendalikan kompleksitas model.

7. Hitung nilai log odds sebagai berikut

$$\log odds_i^t = \log\left(\frac{Pr_i^t}{1 - Pr_i^t}\right) \quad (9)$$

dimana $\log odds_i^t$ menggambarkan logaritma dari rasio antara probabilitas prediksi positif dan probabilitas prediksi negatif untuk setiap sampel. Pr_i^1 merupakan probabilitas prediksi untuk sampel data ke- i pada iterasi ke- t

8. Perbarui nilai probabilitas menjadi diNOrmalisasi dengan rumus berikut

$$Pr_i^{t+1} = \log odds_i^t + (\eta \times output\ (A)_i) \quad (10)$$

Dimana $\log odds_i^t$ adalah log odds untuk sampel data ke- i pada iterasi ke- t. η adalah learning rate, yang mengontrol seberapa besar pengaruh dari pembaruan probabilitas terhadap nilai probabilitas yang baru. $output\ (A)_i$ adalah nilai output dari atribut A untuk sampel data ke- i.

9. Normalisasikan nilai probabilitas menggunakan fungsi sigmoid biner sebagai berikut

$$Sigmoid\ (Pr_i^{t+1}) = \frac{exp^{Pr_i^{t+1}}}{1 + exp^{Pr_i^{t+1}}} \quad (11)$$

Dimana Pr_i^{t+1} adalah nilai probabilitas yang diperbarui untuk sampel data ke- i pada iterasi ke- $t + 1$.

10. Ulangi Langkah ke-2 sampai ke-8.
11. Evaluasi kinerja algoritma klasifikasi dengan *confusion matrix*.

2.5 Analisis Hasil

Dalam penelitian ini, hasil evaluasi ditampilkan dalam bentuk *confusion matrix*. *confusion matrix* merupakan tabel yang digunakan untuk menunjukkan jumlah data uji yang diklasifikasikan dengan benar dibandingkan dengan total jumlah data uji yang diklasifikasikan dengan salah [23]. *Confusion matrix* untuk dua kelas dalam bentuk umum dapat dilihat pada Tabel 2 berikut [24].

Tabel 2. *Confusion Matrix*

Confusion matrix		**Prediksi**	
		Positif	**Negaitf**
Aktual	Positif	TP (*True Positive*)	FN (*False Negative*)
	Negatif	FP (*False Positive*)	TN (*True Negative*)

TP merujuk pada jumlah data positif yang terklasifikasi sebagai positif. FP adalah total jumlah data negatif yang diklasifikasikan sebagai positif. FN menunjukkan jumlah data positif yang terklasifikasi sebagai negatif. TN adalah jumlah data negatif yang terklasifikasi sebagai negatif. Nilai TP, FP, FN, dan TN yang terdapat dalam *confusion*

matrix digunakan untuk mengevaluasi kinerja metode melalui pengukuran akurasi, presisi, dan sensitivitas.

Akurasi adalah nilai yang menunjukkan seberapa sering model membuat prediksi yang benar dibandingkan dengan keseluruhan jumlah prediksi yang dibuat. Presisi adalah nilai yang menunjukkan seberapa besar proporsi prediksi kelas positif dibandingkan dengan seluruh hasil yang diprediksi positif. Sensitivitas mengukur proporsi kelas positif yang benar dibandingkan dengan seluruh hasil yang sebenarnya positif. Berdasarkan [13], [25], dan [26] rumus untuk menghitung akurasi, presisi, dan sensitivitas dapat ditemukan pada persamaan (12), persamaan (13), dan persamaan (14) berikut.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

$$Presisi = \frac{TP}{FP+TP} \quad (13)$$

$$Sensitivitas = \frac{TP}{FN+TP} \quad (14)$$

3. HASIL DAN PEMBAHASAN

3.1. Algoritma ID3

Algoritma ID3 melakukan perhitungan menggunakan teknik pemisahan persentase dan teknik validasi silang k-fold untuk klasifikasi penyakit PCOS menghasilkan *confusion matrix* dengan nilai presisi, sensitivitas dan akurasi untuk setiap kelas target, yang hasilnya disajikan dalam Tabel 3.

Tabel 3. Perbandingan Teknik Pemisahan Persentase dan Teknik Validasi Silang K-fold pada ID3

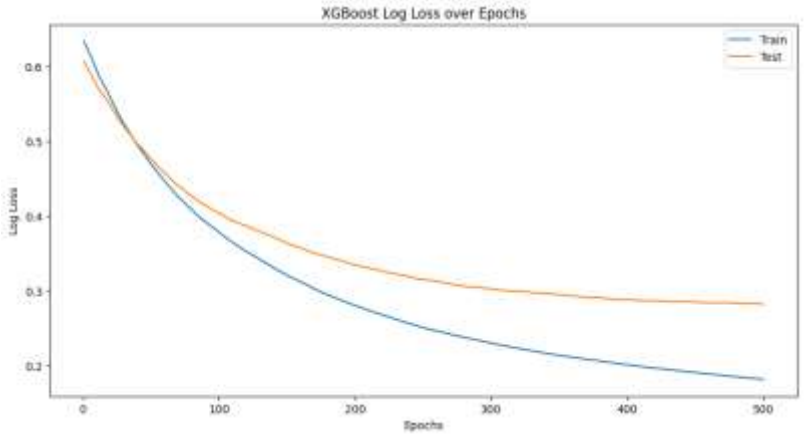
Kelas	Pemisahan Persentase			Validasi Silang K-fold		
	Presisi (%)	Sensitivitas (%)	Akurasi (%)	Presisi (%)	Sensitivitas (%)	Akurasi (%)
YES (Positif PCOS)	72	66	83	77	72	84
NO (Negatif PCOS)	86	90		87	90	

Berdasarkan Tabel 3. dapat disimpulkan bahwa teknik pengujian validasi silang k-fold menghasilkan kinerja yang lebih baik dibandingkan dengan teknik pengujian pemisahan persentase. Hal ini terlihat dari nilai presisi, sensitivitas, dan tingkat akurasi yang lebih optimal saat pengujian dengan teknik validasi silang k-fold.

3.2. Algoritma XGBoost

Penerapan algoritma XGBoost menggunakan teknik pengujian pemisahan persentase dan validasi silang k-fold. Evaluasi metrik yang digunakan pada

algoritma ini menggunakan *log loss*. Dari proses *training* dan validasi, dapat dilihat perbandingan kurva *log loss* seperti pada Gambar 2.



Gambar 2. *Kurva Log Loss Algortima XGBoost*

Dari Gambar 2 terlihat bahwa nilai log loss untuk data pelatihan dan pengujian terus menurun seiring bertambahnya jumlah epoch. Penurunan ini menunjukkan bahwa algoritma mampu belajar dari data secara konsisten tanpa gejala overfitting yang signifikan. Kurva data pengujian yang mendekati data pelatihan menunjukkan model memiliki kemampuan generalisasi yang baik, meskipun laju penurunan log loss pada data pengujian cenderung melambat setelah jumlah epoch tertentu. Dari data yang sudah dilatih dilakukan proses untuk mengukur keberhasilan klasifikasi algoritma XGBoost menggunakan menggunakan teknik pengujian pemisahan persentase dan validasi silang k-fold dengan *confusion matrix*. Hasil akhir *confusion matrix* dari algoritma XGBoost dapat dilihat pada Tabel 4 berikut.

Tabel 4. Perbandingan Teknik Pemisahan Persentase dan Teknik Validasi Silang K-fold pada XGBoost

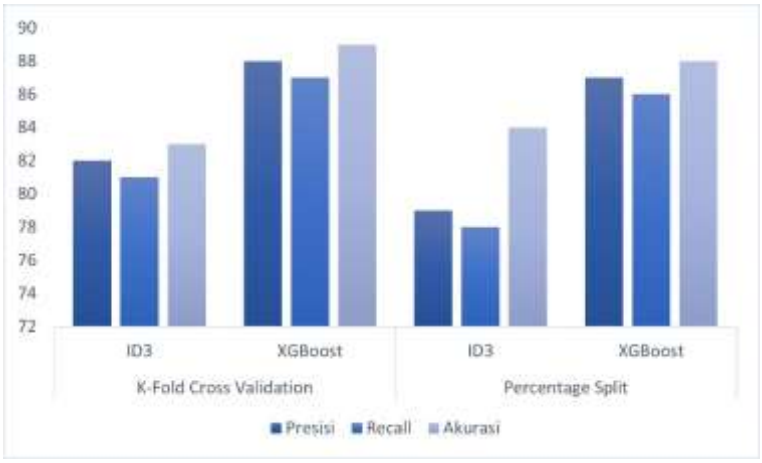
Kelas	Pemisahan Persentase			Validasi Silang K-fold		
	Presisi (%)	Sensitivitas (%)	Akurasi (%)	Presisi (%)	Sensitivitas (%)	Akurasi (%)
YES (Positif PCOS)	85	95	88	86	80	89
NO (Negatif PCOS)	89	72		90	94	

Dari Tabel 4, terlihat bahwa teknik validasi silang k-fold menunjukkan kinerja yang lebih unggul daripada teknik pemisahan persentase. Hal tersebut terlihat dari hasil nilai presisi, sensitivitas, dan akurasi yang lebih tinggi pada teknik teknik validasi silang k-fold.

JSI : Jurnal Sistem Informasi *(E-Journal), VOL. 17, NO. 1, April 2025*
ISSN Print : 2085-1588
ISSN Online :2355-4614
LINK: https://jsi.ejournal.unsri.ac.id/index.php/jsi/index

3.3 Perbandingan Hasil Kedua Algortima

Perbandingan hasil kedua algoritma berdasarkan rata-rata nilai presisi, sensitivitas, dan akurasi dari seluruh kelas dengan menggunakan teknik pengujian pemisahan persentase dan validasi silang k-fold ditampilkan pada Gambar 3.



Gambar 3. Nilai Rata-rata Presisi, Recall, Akurasi metode ID3 dan XGBoost dengan Teknik Pengujian Pemisahan Persentase dan Validasi Silang K-fold

Pada Gambar 3, menggunkan teknik pemisahan persentase, algoritma ID3 menghasilkan rata-rata nilai presisi sebesar 79% dan sensitivitas sebesar 78%. Algoritma XGBoost mencapai rata-rata nilai presisi sebesar 87% dan sensitivitas sebesar 86%. Penerapan teknik validasi silang k-fold, algoritma ID3 mencatatkan rata-rata presisi dan sensitivitas masing-masing sebesar 82% dan 81%, sedangkan XGBoost memperoleh rata-rata nilai presisi sebesar 88% dan recall sebesar 87%. Dari sisi akurasi, algoritma ID3 mencatatkan akurasi sebesar 83% dengan teknik pemisahan persentase dan 84% dengan teknik validasi silang k-fold. Algoritma XGBoost mencatatkan akurasi sebesar 88% dengan teknik pemisahan persentase dan 89% dengan teknik validasi silang k-fold. Berdasarkan nilai presisi, recall, dan akurasi dari algoritma XGBoost terbukti memberikan peforma yang lebih unggul khususnya ketika menggunakan teknik validasi silang k-fold dalam penelitian klasifikasi penyakit PCOS.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilaksanakan disimpulkan pengujian menggunakan teknik validasi silang k-fold mampu meningkatkan kinerja algoritma ID3 dan XGBoost dalam mengklasifikasikan PCOS. Hal ini terlihat dari hasil perbandingan yang menujukkan teknik teknik validasi silang k-fold memberikan peforma yang lebih unggul dibandingkan teknik pemisahan persentase. Berdasarkan hasil evaluasi kedua

algoritma, disimpulkan algoritma XGBoost menunjukkan hasil yang lebih optimal dibandingkan algoritma ID3, dengan kinerja yang lebih unggul pada pengujian menggunakan teknik validasi silang k-fold maupun teknik pemisahan persentase. Pengujian dengan XGBoost, nilai akurasi berada pada rentang 86%-89%, sementara ID3 menghasilkan akurasi antara 83%-84%. Disimpulkan penggunaan algoritma XGBoost dengan teknik validasi silang k-fold adalah salah satu metode yang efektif dalam klasifikasi penyakit PCOS. Model yang dikembangkan dalam penelitian ini diharapkan dapat digunakan sebagai alat deteksi dini otomatis untuk penyakit PCOS, yang akan sangat bermanfaat bagi tenaga medis. Penelitian lanjutan bisa mengeksplorasi kombinasi algoritma lain untuk melihat apakah ada metode yang bisa memberikan hasil yang lebih baik.

DAFTAR PUSTAKA

[1] L. Xing *et al.*, "Depression in polycystic ovary syndrome: Focusing on pathogenesis and treatment," *Front. Psychiatry*, vol. 13, p. 1001484, 2022, doi: 10.3389/fpsyt.2022.1001484.

[2] D. S. Hanani, A. Ardiyanti, and N. V. Ika P, "Hubungan Dukungan Sosial Terhadap Kecemasan Pasien Polycycstic Ovary Syndrome (PCOS)," *Detect. J. Inov. Ris. Ilmu Kesehat.*, vol. 1, no. 3, pp. 197–211, 2023.

[3] D. Himawan, "Aplikasi Data Mining Menggunakan Algoritma ID3 Untuk Mengklasifikasi Kelulusan Mahasiswa Pada Universitas Dian Nuswantoro Semarang," pp. 1–10, 2014.

[4] A. C. Fauzan, "Implementasi Algoritma Decision Tree Iterative Dichotomiser 3 (ID3) Untuk Prediksi Keberhasilan Pengobatan Penyakit Kutil Menggunakan Cryotherapy Implementation of Decision Tree Iterative Dichotomiser 3 (ID3) Algorithm for Predicting the Success of Wa," vol. 4, no. 1, pp. 73–82, 2022, doi: 10.30812/bite.v4i1.1949.

[5] O. Kristanto, "Penerapan Algoritma Klasifikasi Data Mining ID3 Untuk Menentukan Penjurusan Siswa SMAN 6 Semarang," 2014.

[6] R. S. Wahono, C. Supriyanto, F. I. Komputer, and U. D. Nuswantoro, "Penanganan Fitur Kontinyu dengan Feature Discretization Berbasis Expectation Maximization Clustering untuk Klasifikasi Spam Email Menggunakan Algoritma ID3," *J. Intell. Syst.*, vol. 1, no. 2, pp. 148–155, 2015.

[7] I. Srimenganti, I. Taufik, and E. Mulyana, "Implementasi Algoritma Decision Tree (ID3) Untuk Penyakit Campak," *Semin. Nas. Tek. Elektro*, pp. 235–242, 2018.

[8] F. Ferdina, N. Satyahadewi, and D. Kusnandar, "Penerapan Algoritma Iterative Dichotomiser 3 (Id3) Dalam Klasifikasi Faktor Risiko Penyakit Diabetes Melitus," *Var. J. Stat. Its Appl.*, vol. 5, no. 2, pp. 139–146, 2023, doi: 10.30598/variancevol5iss2page139-146.

[9] R. Ridho and H. Hendra, "Klasifikasi Diagnosis Penyakit Covid-19 Menggunakan Metode Decision Tree," *JUST IT J. Sist. Informasi, Teknol. Inf. dan Komput.*, vol. 11, no. 3, pp. 69–75, 2022.

[10] G. Abdurrahman, H. Oktavianto, and M. Sintawati, "Optimasi Algoritma XGBoost Classifier Menggunakan Hyperparameter Gridesearch dan Random Search Pada

Klasifikasi Penyakit Diabetes," *INFORMAL Informatics J.*, vol. 7, no. 3, p. 193, 2022, doi: 10.19184/isj.v7i3.35441.

[11] R. Punnoose and P. Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms," *Int. J. Adv. Res. Artif. Intell.*, vol. 5, no. 9, pp. 22–26, 2016, doi: 10.14569/ijarai.2016.050904.

[12] P. Septiana Rizky, R. Haiban Hirzi, and U. Hidayaturrohman, "Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 15, no. 2, pp. 228–236, 2022, doi: 10.36456/jstat.vol15.no2.a5548.

[13] M. K. Nasution, R. R. Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma Naïve Bayes Dan Algoritma Xgboost Pada Klasifikasi Penyakit Diabetes," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 9765–9772, 2021.

[14] D. Kurnia, M. Itqan Mazdadi, D. Kartini, R. Adi Nugroho, and F. Abadi, "Seleksi Fitur dengan Particle Swarm Optimization pada Klasifikasi Penyakit Parkinson Menggunakan XGBoost," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 5, pp. 1083–1094, 2023, doi: 10.25126/jtiik.20231057252.

[15] Y. Purbolingga, D. Marta, A. Rahmawatia, and B. Wajhi, "Perbandingan Algoritma CatBoost dan XGBoost dalam Klasifikasi Penyakit Jantung," *J. APTEK Vol. 15 No 2 126-133*, vol. 15, no. 2, pp. 126–133, 2023.

[16] P. Kottarathil, "Polycystic ovary syndrome (PCOS)." Accessed: Jun. 03, 2024. Available: https://www.kaggle.com/datasets/prasoonkottarathil/polycystic-ovary-syndrome-pcos

[17] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[18] F. Adams, R. A. Dwi Anggoro, M. B. Satria, A. W. Oktavia, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma Naïve Bayes, Decision Tree, dan Support Vector Machine," *Semin. Nas. Mhs. Ilmu Komput. dan Apl.*, vol. 2, no. 2, pp. 260–268, 2021.

[19] A. A. Mortara, M. Permatasari, A. Desiani, Y. Andriani, and M. Arhami, "Perbandingan Algoritma C4.5 dan Adaptive Boosting dalam Klasifikasi Penyakit Alzheimer," *J. Teknol. dan Inf.*, vol. 13, no. 2, pp. 196–207, 2023, doi: 10.34010/jati.v13i2.10525.

[20] J. A. Sidette, E. Eko, and O. D. Nurhayati, "Pendekatan Metode Pohon Keputusan Menggunakan Algoritma ID3 Untuk Sistem Informasi Pengukuran Kinerja PNS," *J. Sist. Inf. Bisnis*, vol. 4, no. 2, pp. 75–86, 2014, doi: 10.21456/vol4iss2pp75-86.

[21] R. Forest, C. Dan, and X. Classifier, "Penentuan kelayakan promosi pegawai menggunakan algoritma random forest classifier dan xgboost classifier," vol. 6, pp. 773–783, 2023, doi: 10.37600/tekinkom.v6i2.949.

[22] M. W. Dwinanda, N. Satyahadewi, and W. Andani, "Classification of Student Graduation Status Using Xgboost Algorithm," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 3, pp. 1785–1794, 2023, doi: 10.30598/barekengvol17iss3pp1785-1794.

[23] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform.*, vol. 5, no. September, pp. 697–711, 2021.

[24] M. Christianto, J. Andjarwirawan, and A. Tjondrowiguno, "Aplikasi analisa sentimen pada komentar berbahasa Indonesia dalam objek video di website YouTube menggunakan metode Naïve Bayes classifier," *J. Infra*, vol. 8.1, pp. 255–259, 2020.

[25] D. Putra and A. Wibowo, "Prediksi Keputusan Minat Penjurusan Siswa SMA Yadika 5 Menggunakan Algoritma Naïve Bayes," *Pros. Semin. Nas. Ris. Dan Inf. Sci.*, vol. 2, pp. 84–92, 2020.

[26] L. Farokhah, "Implementasi K-Nearest Neighbor untuk Klasifikasi Bunga Dengan Ekstraksi Fitur Warna RGB," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 6, pp. 1129–1136, 2020, doi: 10.25126/jtiik.2020722608.