JINTEKS (Jurnal Informatika Teknologi dan Sains) ISSN 2686-3359 (Online)
Vol. 4 No. 3, Agustus 2022, hlm. 185 – 192

KOMPARASI ALGORITMA KLASIFIKASI UNTUK MENENTUKAN EVALUASI KINERJA TERBAIK PADA STATUS AKREDITASI SEKOLAH/MADRASAH KALIMANTAN TIMUR BERDASARKAN IASP 2020

Taghfirul Azhima Yoga Siswa[1*], Naufal Azmi Verdikha[2]
[1,2]Teknik Informatika, Universitas Muhammadiyah Kalimantan Timur
email: tay758@umkt.ac.id

Abstrak: Indonesia mulai beradaptasi pada era revolusi industri 4.0 ke era society 5.0 dengan penerapan teknologi modern dan menciptakan peluang baru pada semua aspek kehidupan. Selain pengembangan infrastruktur, rencana pemindahan Ibu Kota Negara (IKN) ke Provinsi Kalimantan Timur (KALTIM) juga menjadi catatan penting dalam kesiapan sumber daya manusia berkualitas yang dapat dilihat dari mutu pendidikan dengan status akreditasi sekolah. Penelitian ini bertujuan untuk melakukan komparasi terhadap beberapa algoritma klasifikasi seperti C4.5, Naïve Bayes, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM) dan Logistic Regression untuk mencari kinerja terbaik dalam klasifikasi status akreditasi sekolah/madrasah provinsi Kalimantan Timur berdasarkan IASP 2020. Tahap preprocessing membagi data dilakukan menggunakan metode cross validation yang bersumber pada data BAN S/M KALTIM tahun 2020-2021 berjumlah 295 record. Kemudian dilakukan evaluasi kinerja algoritma untuk mencari nilai *Accuracy*, *Precision* dan *Recall* menggunakan *confusion matrix*. Hasil komparasi algoritma klasifikasi didapatkan bahwa algoritma SVM memiliki kinerja terbaik dibandingkan dengan algoritma klasifikasi lainnya dengan nilai *accuracy* 88.8%, *precision* 91%, *recall* 81%, dan *f1-score* 84%, sehingga didapatkan kesimpulan bahwa algoritma SVM memiliki kinerja terbaik dibanding algoritma lainnya dalam klasifikasi data status akreditasi sekolah/madrasah KALTIM berdasarkan IASP 2020.

Kata Kunci : C4.5, Naive Bayes, KNN, SVM, Logistic Regression, Kinerja Algoritma Klasifikasi, Akreditasi Sekolah, IASP 2020

Abstract: *Indonesia began to adapt in the era of the industrial revolution 4.0 to the era of society 5.0 with the application of modern technology and creating new opportunities in all aspects of life. In addition to infrastructure development, the plan to relocate the State Capital (IKN) to East Kalimantan Province (KALTIM) is also an important note in the readiness of quality human resources which can be seen from the quality of education with school accreditation status. This study aims to compare several classification algorithms such as C4.5, Naïve Bayes, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM) and Logistic Regression to find the best performance in the classification of school/madrasah accreditation status in Kalimantan province. East based on IASP 2020. The preprocessing stage for dividing data is carried out using the cross validation method sourced from BAN S/M KALTIM data for 2020-2021 totaling 295 records. Then evaluate the performance of the algorithm to find the value of Accuracy, Precision and Recall using the confusion matrix. The results of the classification algorithm comparison show that the SVM algorithm has the best performance compared to other classification algorithms with 88.8% accuracy, 91% precision, 81% recall, and 84% f1-score, so it can be concluded that the SVM algorithm has the best performance compared to other algorithms in classification of data on the accreditation status of KALTIM schools/madrasahs based on IASP 2020.*

Keywords **:** C4.5, Naive Bayes, KNN, SVM, Logistic Regression, Classification Algorithm Performance, School Accreditation, IASP 2020

PENDAHULUAN

Negara Indonesia mulai beradaptasi pada era revolusi industri 4.0 ke era society 5.0 dengan penerapan teknologi modern dan menciptakan peluang baru pada semua aspek kehidupan. Selain pengembangan infrastruktur, wacana pemindahan Ibu Kota Negara (IKN) ke Provinsi Kalimantan Timur (KALTIM) menjadi catatan penting dalam kesiapan sumber daya manusia berkualitas yang dapat dilihat dari mutu pendidikan dengan status akreditasi sekolahnya. Sehingga kegiatan akreditasi diharapkan mampu mendorong untuk dapat menciptakan suasana kondusif bagi perkembangan pendidikan serta memberikan arahan dalam melakukan penjaminan mutu sekolah/madrasah yang berkelanjutan di Kaltim.

Pada proses kegiatan akreditasi sekolah, sesuai pada peraturan Mendiknas Nomor 29 Tahun 2005, pemerintah telah membentuk Badan Akreditasi Nasional Sekolah/Madrasah (BAN S/M) yang bertujuan sebagai badan yang melakukan evaluasi dan menetapkan kelayakan program satuan pendidikan berdasarkan pada standar nasional pendidikan. Peringkat akreditasi predikat A (Unggul), predikat akreditasi B (Baik), predikat akreditasi C (Cukup), dan peringkat tidak terakreditasi (TT) merupakan hasil penilaian masing-masing komponen dalam menentukan peringkat akreditasi pada sekolah/madrasah yang dilaksanakan oleh BAN-SM [1].

Peraturan Pemerintah No. 19 Tahun 2005 pada dasarnya merupakan proses penilaian akreditasi mengacu pada standar nasional pendidikan sesuai tentang Standar Nasional Pendidikan pasal 2 ayat (1) bahwa cakupanya antara lain (1) standar isi (2) standar proses (3) standar kompetensi lulusan (4) standar pendidik dan tenaga kependidikan (5) standar sarana dan prasarana (6) standar pengelolaan (7) standar pembiayaan dan (8) standar penilaian pendidikan. Pada tahun 2018 kemudian Badan

 ISSN 2686-3359 (Online)

Akreditasi Nasional Sekolah/Madrasah (BAN-S/M) telah melakukan rancangan dan menerapkan perubahan sistem akreditas, dari sebuah paradigma berbasis kepatuhan administratif (*compliance*) menjadi berbasis pada kinerja (*performance*) yang diberi istilah IASP 2020. Adapun komponen utama yang dinilai dari instrumen IASP ini adalah mutu lulusan, proses pembelajaran, mutu guru, dan manajemen sekolah/madrasah [2].

Penelitian sebelumnya yang pernah dilakukan terkait *data mining* akreditasi sekolah diantaranya yaitu dengan menerapkan algoritma klasifikasi Support Vector Machine (SVM) untuk data akreditasi SD Magelang dengan hasil akurasi berdasarkan fungsi kernel Gaussian RBF sebesar 93,90%, berdasarkan 77 data dari 82 SD yang diklasifikasikan dengan benar sesuai kelas aslinya [3]. Kemudian penelitian lainya membandingkan klasifikasi sekolah dasar negeri akreditasi di kota Semarang dengan metode K-NN dan MARS yang menghasilkan klasifikasi terbaik dari metode K-NN menggunakan K=5 dengan tingkat kesalahan terkecil dan diperoleh informasi yang benar. Data diolah sebanyak 159 dan data kesalahan klasifikasi sebanyak 9. Klasifikasi terbaik hasil dari metode MARS adalah ketika menggunakan kombinasi BF=32, MI=2, MO=1 karena menghasilkan Generalized Cross Validation (GCV) terkecil dan diperoleh informasi data klasifikasi yang benar sebanyak 164 dan klasifikasi data yang salah sebanyak 4 [4].

Penelitian data mining klasifikasi akreditasi sekolah lainnya juga pernah dilakukan dengan implementasi metode jaringan syaraf tiruan untuk menganalisis data akreditasi sekolah menengah atas atau madrasah Aliyah, di mana hasil pengolahan dengan matlab diakukan pembagian masing-masing 21 data untuk data training dan 21 data untuk data testing dengan pola arsitektur terbaik 8-5-1 dengan persentase kebenaran 100% [5]. Penelitian serupa yaitu klasifikasi akreditasi sekolah menengah pertama di sulawesi dengan algoritma jaringan syaraf tiruan backpropagation, jumlah data yang dianalsis 1872 record dan dilakukan proses pembagian data menggunakan 3-fold cross validation. Hasil akurasi yang didapatkan sebesar 93,42% melalui struktur yang paling optimal pada dua hidden layer dengan variasi dari 10 dan 20 *neuron* [6].

Beberapa penerapan algoritma data mining klasifikasi lainnya juga digunakan untuk membangun satu model prediksi pada akreditasi sekolah di tingkat SMP menggunakan algotitma Regresi Logistik, Bagging Random Forest dan Naive Bayes. Disimpulkan bahwa model algoritma Random Forest menghasilkan nilai prediksi lebih baik dari 3 model lainnya. Namun, hasil akurasi ini tidak semata-mata memberikan gambaran bahwa model Random Forest merupakan model yang lebih baik daripada model model lainnya dalam semua kasus pemodelan data klasifikasi [7]. Dilanjutkan penerapan klasifikasi akreditasi dengan algoritma k-nearest neighbor (KNN) pada data sekolah menengah atas, yang menghasilkan nilai tertinggi diperoleh pada penggunaan atribut sebanyak 8 dan nilai k = 3, sehingga menghasilkan nilai akurasi sebesar 80,7051%. Penelitian lainya melakukan Klasifikasi Akreditasi SMA di Sumatera dengan algoritma / metode naive bayes menghasilkan Akurasi tertinggi dengan penggunaan jumlah attribut sebanyak 8 dan 9 yakni 94,165% [8].

Adapun penelitian penelitian terdahulu yang telah diuraikan sebelumnya terkait klasifikasi akreditasi sekolah hanya mengacu pada penerapan model algoritma dan komparasi metode klasifikasi yang terbatas pada beberapa metode saja. Kemudian aspek penilaian akreditasi yang digunakan pada penelitian sebelumnya juga masih menggunakan acuan standar instrumen penilaian yang lama berdasarkan kriteria 1) standar isi, 2) standar proses, 3) standar kompetensi lulusan, 4) standar pendidik dan kependidikan, 5) standar sarana dan prasarana, 6) standar pengelolaan, dan 7) standar pembiayaan dan standar penilaian. Perbedaan dari penelitian sebelumnya pada dasarnya penelitian ini dilakukan bertujuan untuk melakukan komparasi terhadap beberapa algoritma klasifikasi seperti C4.5, Naïve Bayes, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM) dan Logistic Regression untuk mencari kinerja terbaik dalam klasifikasi status akreditasi sekolah/madrasah Kalimantan Timur berdasarkan acuan standar instrumen penilaian terbaru yaitu IASP 2020 yang terdikir dari komponen 1) mutu lulusan, 2) proses pembelajaran, 3) mutu guru dan 4) manajemen sekolah/madrasah. Hal-hal tersebut merupakan salah satu cara menciptakan gagasan untuk mengadakan penelitian ini.

TINJAUAN PUSTAKA

Standar Nasional Pendidikan

Definisi akreditasi menurut Kamus Besar Bahasa Indonesia merupakan sebuah istilah pengakuan terhadap lembaga pendidikan yang diberikan oleh badan yang berwenang setelah dinilai bahwa lembaga itu memenuhi syarat kebakuan atau kriteria tertentu [1]. Standar Nasional Pendidikan merupakan kriteria minimal tentang sistem pendidikan di seluruh wilayah hukum Negara Kesatuan Republik Indonesia. Pemerintah telah menetapkan Badan Akreditasi Nasional Sekolah/Madrasah (BAN-S/M) [9].

Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan berdasarkan kriteria pembentuk keputusan [10]. Algoritma C4.5 mirip seperti pohon dengan memiliki node internal (bukan daun) untuk deskripsi atribut pada tiap-tiap cabang menggambarkan hasil atribut yang diuji dan tiap daun menggambarkan kelas [11].

 ISSN 2686-3359 (Online)

Tahapan-tahapan dalam membuat sebuah pohon keputusan dengan metode algoritma C4.5 antara lain [12] :

a. Menyiapkan data training.
b. Menentukan akar dari pohon. Sebelum menghitung nilai *Gain* dari atribut, hitung dahulu nilai *entropy*. Kemudian menghitung nilai *Gain* dari masing-masing atribut untuk menentukan akar, nilai *Gain* yang paling tinggi yang akan menjadi akar pertama.

$$Entropy(s) = \sum_{i=1}^{n} \quad - pi * pi \quad (1)$$

Keterangan :
S : Himpunan kasus
A : Atribut
n : Jumlah partisipasi S
pi : Proporsi dari Si terhadap S

c. Menghitung nilai Gain dengan metode information gain.

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \quad \frac{|Si|}{|S|} * Entropy(S) \quad (2)$$

Keterangan :
S : Himpunan kasus
A : Atribut
n : Jumlah partisipasi atribut A
|Si| : Jumlah kasus pada partisi ke-i
|S| : Jumlah kasus dalam S

d. Ulangi langkah ke-2 hingga semua tupel telah terpartisi.

e. Proses partisipasi pohon keputusan berhenti apabila memenuhi status antara lain :
   1) Semua tupel dalam node N mendapat kelas yang sama.
   2) Tidak ada atribut di dalam tupel yang dipartisi lagi.
   3) Tidak ada tupel di dalam cabang yang kosong.

Algoritma Naive Bayes

Algoritma Naïve Bayes merupakan salah satu algoritma yang ada pada teknik data mining pada model klasifikasi [13]. Dalam melakukan teknik klasifikasi, algoritma naïve bayes menggunakan metode probabilitas dan statistik yang dikenalkan oleh ilmuwan Inggris bernama Thomas Bayes. Naïve bayes digunakan untuk memprediksi kemungkinan atau peluang di masa yang akan datang berdasarkan pengalaman di masa sebelumnya, istilah tersebut sering dikenal dengan Teorema Bayes. Persamaan dalam Teorema Bayes pada umumnya adalah sebagai berikut [14]:

$$P(A \mid B) = \frac{P(A)P(B|A)}{P(B)} \quad (3)$$

Keterangan:
P(A | B) : Peluang terjadinya A dengan syarat B telah terjadi
P(B | A) : Peluang terjadinya B dengan syarat A telah terjadi
P (A) : Peluang terjadinya A
P (B) : Peluang terjadinya B

Algoritma K-Nearest Neighbor (K-NN)

Algoritma *K- Nearest Neighbor* bekerja dengan cara mencari nilai *k* objek atau pola data (dari semua pola training yang tersedia) yang paling terdekat dengan pola masukan dan memilih kelas dengan jumlah pola terbanyak di antara nilai *k* pola tersebut [15].

Rumus *Euclidien Distance* :

$$Euclidien\ distance = \sqrt{\sum_{i=1}^{P} \quad (a_{k-} b_k)^2} \quad (2.1)$$

Sumber : [16]
Keterangan :
a_k = Sampel data
b_k = Data uji atau *testing*
p = Dimensi data
i = Variabel data

Algoritma *K- Nearest Neighbor* adalah salah satu teknik klasifikasi yang digunakan untuk penyelesaian masalah pada bidang data mining dengan pendekatan untuk memilih kasus baru kemudian menghitung kedekatan dengan kasus lama [17].

Pada algoritma K-NN ada beberapa tahapan yang harus dilakukan yang dijelaskan pada uraian berikut :

a. Menentukan nilai k (jumlah tetangga terdekat).
b. Melakukan perhitungan jarak antar data *testing* dan seluruh data *training* menggunakan rumus jarak *Euclidean*
c. Mengurutkan jarak (*ranking*).
d. Menggunakan pemilihan kelas sebagai prediksi dari data testing tersebut.

Algoritma Support Vector Machine (SVM)

Algoritma Support Vector Machine (SVM) merupakan metode yang bekerja menggunakan pemetaan secara nonlinear untuk mengubah data pelatihan dengan dimensi yang lebih tinggi [11]. Dengan demkikan dimensi baru akan mencari *hyperplane* untuk memisahkan secara linier dengan pemetaan nonlinear yang tepat ke dimensi lebih tinggi, sehingga data dari dua kelas tersebut selalu dapat dipisahkan dengan *hyperplane*.

$$f(x) = w^t \phi(x) + b \ (5)$$

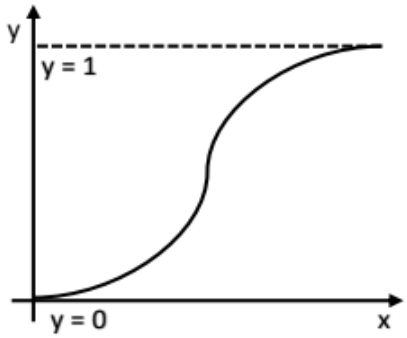
Keterangan:
$\boldsymbol{b}$: Bias
$\boldsymbol{x}$: ($\boldsymbol{x1}$, $\boldsymbol{x2}$,……..$\boldsymbol{xD}$)
$\boldsymbol{T}$: Variabel Input

 ISSN 2686-3359 (Online)

$\boldsymbol{w}$: $(\boldsymbol{w0}, \boldsymbol{w1}, \ldots\ldots \boldsymbol{wD})$
$\boldsymbol{T}$: Parameter Bobot
$\phi(\boldsymbol{x})$: Fungsi Transformasi fitur

Algoritma Logistic Regression

Logistic regression merupakan algoritma yang dapat memisahkan dataset menjadi dua bagian yang disebut dengan *binary classification* menggunakan metode prediksi probabilitas. *Logistic regression* menghasilkan *output* yang bersifat kualitatif dan kategori [18].



Gambar 1. Logistic Regression

Grafik ini membagi dataset menjadi dua *class* = 1 dan *class* = 0 tepat di tengah, yaitu saat Y = 0.5. *Class* merupakan prediksi probabilitas (p atau P) yang dirumuskan:

$$p \geq 0.5, class = 1 \ (6)$$

$$p < 0.5, class = 0 \ (7)$$

Probabilitas regresi logistik sebagai berikut:

$$p = \frac{e^{\beta_0+\beta_1+\varepsilon_i}}{1+e^{-(\beta_0+\beta_1 X+\varepsilon_i)}} \ (8)$$

Atau

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X+\varepsilon_i)}} (9)$$

Keterangan:
P : Probabilitas
e : Fungsi Eksponen

Cross Validation

Istilah *cross validation* merupakan bentuk sederhana dari sebuah metode statistik yang dapat digunakan dalam melakukan evaluasi kinerja algoritma. Biasa disebut juga dengan *k-fold validation* yaitu dapat melakukan percobaan sebanyak k-kali untuk parameter model yang sama. Jumlah fold standar untuk memprediksi tingkat *error* dari data penelitian ini menggunakan 5-*fold cross validation*.

Gambar 2. Ilustrasi pembagian data menggunakan 10-*Fold Cross Validation*

Confusion Matrix

Confusion Matrix merupakan suatu metode yang digunakan dalam melakukan pengukuran permorma/kinerja konsep model data mining klasifikasi. Biasanya metode ini dapat melakukan perhitungan dengan 4 keluaran, antara lain : *precision, accuracy, recall* dan *error rate*. Evaluasi klasifikasi ini didasarkan pada pengujian untuk memperkirakan model objek yang benar dan salah [19].

Tabel 1. *Confusion Matrix*

Correct Classification	*Classified as*	
	Predicted (+)	Predicted (-)
Actual (+)	*True Positive*	*False Negative*
Actual (-)	*False Positive*	*True Negative*

Berdasarkan tabel confusion matrix diatas [20] :

a. *True Positive* (TP) adalah jumlah memprediksi hasil positif dan itu benar.
b. *False Positive* (FP) adalah jumlah memprediksi hasil negatif dan itu benar.
c. *False Negative* (FN) adalah jumlah memprediksi hasil negative dan itu salah.
d. *True Negative* (TN) adalah jumlah memprediksi hasil negative dan itu banar.

Berdasarkan tabel *confusion matrix* diatas dapat digunakan untuk menghitung nilai *recall*, *precision*, *accuracy* dan *error rate*. Berikut rumus nya :

1. *Accuracy* merupakan gambaran seberapa akurat model klasifikasi yang benar dengan ketentuan sebagai berikut:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \ (10)$$

2. *Recall* atau *true positive* (TP) adalah gambaran keberhasilan sebuah model untuk mendapatkan kembali sebuah informasi, denan ketentuan sebagai berikut.

$$Recall = \frac{TP}{TP+FP} \ (11)$$

3. *Precision* (P) adalah gambaran akurasi antara data yang diminta dengan hasil prediksi model klasifikasi dengan ketentuan sebagai berikut:

 ISSN 2686-3359 (Online)

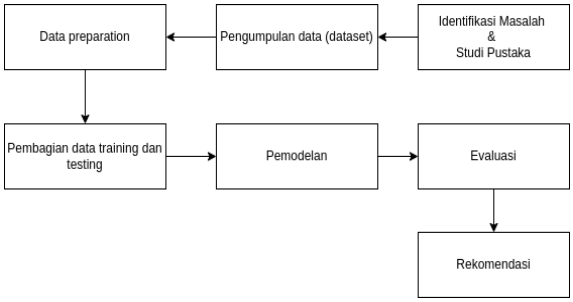
$$Precision = \frac{TP}{TP + FN} \quad (12)$$

4. *F1-Score* adalah gambaran komparasi rata rata antara *precision* dan *recall* yang dibobotkan dengan ketentuan sebagai berikut:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (13)$$

METODE

Metode penelitian yang digunakan dalam penelitian ini adalah merupakan metode eksperimen, dimana meteode penelitian ini merupakan salah satu penelitian kuantitatif dengan memanipulasi satu atau lebih variabel bebas (*independent variable*), mengontrol variabel lain yang relevan, dan mengamati efek dari manipulasi pada variabel terikat (*dependent variable*) [21]. Dalam penelitian ini dilakukan beberapa eksperimen terhadap algoritma klasifikasi pada data Status Akreditasi Sekolah/Madrasah Kalimantan Timur untuk menemukan evaluasi kinerja terbaik.



Gambar 3. Alur Tahapan Penelitian

Tahapan pertama yakni identifikasi masalah dan studi pustaka, Masalah yang diangkat dalam penelitian ini adalah mencari kinerja terbaik dalam klasifikasi status akreditasi sekolah/madrasah provinsi KALTIM. Studi pustaka juga dilakukan dalam mencari *gap* penelitian pada penelitian-penelitian sebelumnya yang berkaitan dengan klasifikasi status akreditasi sekolah. Tahap kedua, pengumpulan data (dataset). Pengumpulan data adalah langkah awal dalam melakukan analisis data dalam penelitian. Data yang digunakan pada penelitian ini adalah data sekunder yang bersumber dari BAN S/M Kaltim. Data yang digunakan adalah data BAN S/M Kaltim tahun 2020-2021. Data-data tersebut meliputi NPSN, nama sekolah, jenjang, kota kabupaten, mutu lulusan, proses pembelajaran, mutu guru, manajemen S/M, nilai akhir, peringkat, status akreditasi dan tahun akreditasi. Data awal yang didapatkan dari BAN S/M sejumlah 3.451 record. Atribut awal meliputi NPSN, Nama Sekolah, Kota/Kabupaten, Jenjang, Mutu Lulusan, Proses Pembelajaran, Mutu Guru, Manajemen S/M, Nilai Akhir, Peringkat, Status Akreditasi, Tahun Akreditasi. Tahap ketiga adalah data *preparation*. Pada tahap ini ada beberapa tahapan yang dilakukan seperti data *selection*, data *transformation*, dan data *cleaning*. Beberapa tahapan tersebut dilakukan agar data yang digunakan saat proses analisis data memiliki kualitas yang baik.

a. Data *Selection*: Proses data selection atau pemilihan data yang dilakukan pada dataset akreditasi sekolah dilakukan untuk memilih data dan atribut yang sesuai. Proses pemilihan data dibatasi pada rentang tahun 2020 hingga 2021 saja, menyesuaikan instrumen akreditasi sekolah terbaru IASP 2020. Pemilihan atribut yang akan digunakan adalah pada atribut nama sekolah, jenjang, kota kabupaten, mutu lulusan, proses pembelajaran, mutu guru, manajemen s/m, dan peringkat untuk komparasi pemodelan nantinya. Sedangkan penghapusan atribut yang tidak digunakan adalah pada atribut NPSN, nilai akhir, status akreditasi, dan tahun akreditasi.
b. Data *Transformation*: proses data transformation atau transformasi data yang dilakukan dalam penelitian ini adalah mengubah nilai dari atribut-atribut yang bersifat kategorik menjadi numerik, hal ini dilakukan karena pada penerapan menggunakan library sklearn hanya bisa menerima nilai atribut numerik. Beberapa atribut yang ditransformasi datanya meliputi Jenjang dan Kota/Kabupaten. Contoh data yang telah dilakukan transformasi dapat dilihat pada gambar 4 berikut :

	Nama_Sekolah	Jenjang	Kota_Kabupaten	Mutu_Lulusan	Proses_Pembelajaran	Mutu_Guru	Manajemen_S/M	Peringkat
0	SD NEGERI 001 GUNUNG SARI	3	0	32.61	27.96	17.1	16.62	A
1	SD NEGERI 001 LESAN DAYAK	3	0	25.45	20.71	12.6	14.88	C
2	SD NEGERI 001 LONG LAMCIN	3	0	27.05	23.82	14.4	14.19	C
3	SD NEGERI 001 SAMBALIUNG	3	0	30.00	28.00	16.0	17.00	B
4	SD NEGERI 002 BUYUNG BUYUNG	3	0	23.07	21.75	14.4	13.15	C

Gambar 4. Contoh data setelah ditransformasi

c. Data *Cleaning*: proses terakhir pada tahapan persiapan data adalah data cleaning atau pembersihan data. Pembersihan data yang dilakukan pada dataset akreditasi sekolah adalah menghapus nilai data yang bernilai 0 untuk atribut yang bernilai numerik dan memiliki nilai #N/A (*no value is available*) atau tidak memiliki nilai. Proses pembersihan data menghasilkan data sebanyak 295 record data.

Tahap keempat pembagian data *training* dan data *testing* yang digunakan untuk meguji algoritma. Dataset akan dibagi menjadi dua bagian yakni data training dan data testing. Pada pembagian dataset digunakan teknik *cross validation* dengan menggunakan 5-*Fold*. Tahap kelima pemodelan. Tahap ini dilakukan setelah dataset dibagi menjadi 2 bagian, pada tahapan ini data training digunakan untuk melakukan pemodelan. Tahap keenam evaluasi. Dalam tahap evaluasi dalam penelitian ini menggunakan teknik *confusion matrix*, teknik ini

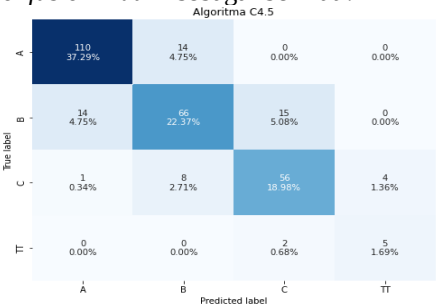
 ISSN 2686-3359 (Online)

digunakan untuk mengetahui bagaimana kinerja dari beberapa algoritma klasifikasi yang dikomparasikan dalam penelitian ini, sehingga didapatkan algoritma klasifikasi terbaik dalam melakukan klasifikasi status sekolah/madrasah provinsi KALTIM. Tahap terakhir rekomendasi. Rekomendasi diberikan dari terkait algoritma klasifikasi apa yang memiliki kinerja paling baik, sehingga jika BAN S/M ingin membuat aplikasi klasifikasi akreditasi sekolah/madrasah provinsi KALTIM dapat menerapkan algoritma-algoritma yang diterapkan pada penelitian ini.

HASIL DAN PEMBAHASAN

A. Algoritma C4.5

Berdasarkan pengujian dari penerapan algoritma C4.5 menggunakan teknik *confusion matrix* dalam melakukan klasifikasi pada data status akreditasi sekolah/madrasah di KALTIM didapatkan tabel *confusion matrix* sebagai berikut :

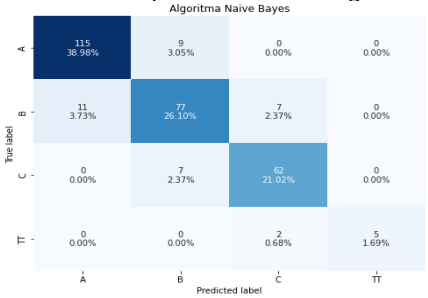


Gambar 5. *Confusion Matrix* pada Algoritma C4.5

Dari hasil tabel *confusion matrix* pada algoritma C4.5 didapatkan nilai *accuracy* 80.3%, nilai precision 74%, nilai *recall* 78%, dan nilai *f1-score* 75%. Dari tabel *confusion matrix* juga ditemukan bahwa prediksi status akreditasi yang tidak tepat paling banyak terdapat pada status akreditasi yang seharusnya terakreditasi B tetapi diprediksi oleh C4.5 dengan akreditasi A, total ada 15 data atau 5.08% prediksi yang tidak tepat pada kelas tersebut.

B. Algoritma Naive Bayes

Berdasarkan pengujian dari penerapan algoritma Naive Bayes menggunakan teknik *confusion matrix* dalam melakukan klasifikasi didapatkan tabel *confusion matrix* sebagai berikut :



Gambar 6. *Confusion Matrix* Algoritma Naive Bayes

Berdasarkan tabel *confusion matrix* pada algoritma Naive Bayes didapatkan nilai accuracy sebesar 88%, nilai *precision* 90%, nilai *recall* 84% dan nilai *f1-score* 86%. Pada gambar 6 juga ditemukan bahwa prediksi status akreditasi yang tidak tepat paling banyak, ada pada status akreditasi yang seharusnya terakreditasi B tetapi diprediksi oleh Naive Bayes dengan akreditasi A, total ada 11 data atau 3.73% prediksi yang tidak tepat pada kelas tersebut.

C. Algoritma K-Nearest Neighbor (K-NN)

Pada algoritma K-NN beberapa pengujian dilakukan, dengan melakukan *trial and error* terhadap nilai k. Nilai k yang dicoba pada penelitian ini k=1, 3, 5 dan 7. Hasil percobaan nilai k yang dilakukan pada algoritma K-NN dapat dilihat pada tabel berikut :

Tabel 1. Hasil Percobaan nilai k pada algoritma K-NN

Nilai k	*Accuracy*	*Precision*	*Recall*	*F1-Score*
k=1	81.4%	80%	73%	75%
k=3	83.1%	87%	73%	77%
k=5	82%	87%	72%	76%
k=7	83.4%	88%	74%	78%

Berdasarkan hasil beberapa percobaan nilai k yang terlihat pada tabel diatas ditemukan bahwa nilai k=7 mendapatkan akurasi tertinggi sebesar 83.4%, sedangkan nilai k=1 memiliki akurasi terendah sebesar 81.4%. Dari percobaan nilai k juga dapat diketahui bahwa nilai k pada K-NN juga berpengaruh dengan kinerja algoritma, walaupun hanya memiliki selisih kurang dari 5%. Algoritma K-NN dengan nilai k=7 digunakan dalam komparasi dengan algoritma klasifikasi lainnya untuk mencari kinerja terbaik.

D. Algoritma Support Vector Machine (SVM)

Pada algoritma SVM beberapa pengujian juga dilakukan, dengan melakukan percobaan beberapa kernel yang ada pada algoritma SVM seperti kernel *Radial Basic Function* (RBF), linear dan *polynomial*. Hasil dari beberapa percobaan kernel tersebut dapat dilihat pada tabel berikut :

Tabel 2. Hasil Percobaan kernel pada algoritma SVM

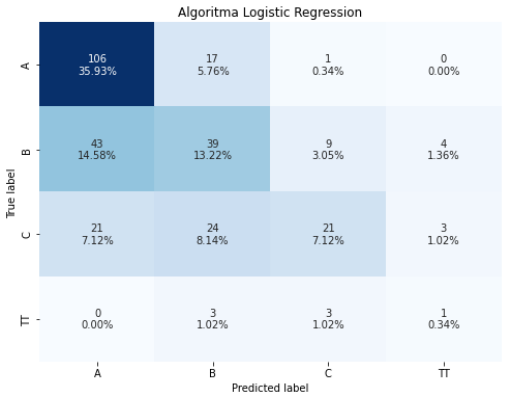
Kernel	*Accuracy*	*Precision*	*Recall*	*F1-Score*
RBF	87.8%	91%	77%	81%
Linear	88.5%	87%	84%	85%
Polynomial	88.8%	91%	81%	84%

Pada tabel 2 terlihat kinerja dari percobaan beberapa kernel SVM ditemukan bahwa kernel Polynomial memiliki kinerja terbaik dengan *accuracy* sebesar 88.8% dari beberapa kernel yang ada, sedangkan kernel RBF memiliki kinerja terburuk dari beberapa kernel yang dicobakan walaupun perbedaan *accuracy* pada kedua kernel tersebut hanya 1%. Algoritma SVM dengan kernel RBF digunakan dalam komparasi dengan algoritma klasifikasi lainnya untuk mencari kinerja terbaik.

E. Algoritma Logistic Regression

Berdasarkan pengujian dari penerapan algoritma Logistic Regression menggunakan teknik *confusion matrix* dalam melakukan klasifikasi didapatkan tabel *confusion matrix* sebagai berikut :

 ISSN 2686-3359 (Online)



Gambar 4. Hasil Tabel *Confusion Matrix* pada Algoritma Logistic Regression

Berdasarkan tabel *confusion matrix* pada algoritma *Logistic Regression* didapatkan nilai accuracy sebesar 56.6%, nilai *precision* 46%, nilai *recall* 43% dan nilai *f1-score* 43%. Pada gambar 4 juga ditemukan bahwa prediksi status akreditasi yang tidak tepat paling banyak, ada pada status akreditasi yang seharusnya terakreditasi B tetapi diprediksi oleh *Logistic Regression* dengan akreditasi A, total ada 43 data atau 14.58% prediksi yang tidak tepat pada kelas tersebut.

F. Komparasi Algoritma Klasifikasi

Berdasarkan pengujian dari komparasi algoritma klasifikasi untuk melihat hasil kinerja masing masing algoritma dalam melakukan klasifikasi data status akreditasi sekolah/madrasah di Provinsi Kalimantan Timur didapatkan hasil sebagai berikut :

Tabel 3. Hasil Komparasi Algoritma Klasifikasi

	Accuracy	*Precision*	*Recall*	*F1-Score*
C4.5	80.3%	74%	78%	75%
Naive Bayes	87.8%	90%	84%	86%
K-NN	83.4%	88%	74%	78%
SVM	88.8%	91%	81%	84%
Logistic Regression	56.6%	46%	43%	43%

Dari hasil komparasi algoritma klasifikasi yang terlihat pada tabel 3 ditemukan bahwa Algoritma SVM dengan kernel Polynomial memiliki *accuracy* tertinggi dari beberapa algoritma klasifikasi lainnya. Hal ini sejalan dengan penelitian oleh [22] algoritma SVM lebih baik dari pada C4.5 dan Naive Bayes. Begitu pula penelitian [23] yang menunjukan SVM juga lebih baik dalam memprediksi *cardiovascular diseases* dibandingkan dengan K-NN dan *Logistic Regression*. Algoritma dengan akurasi terendah ditemukan pada algoritma *Logistic Regression* dengan selisih *accuracy* dengan SVM sebesar 32.2%. Dari hasil tersebut juga ditemukan bahwa algoritma Naive Bayes memiliki nilai tertinggi pada nilai *recall* sebesar 84% dan *f1-score* sebesar 86%.

KESIMPULAN DAN SARAN

Hasil dari pengujian telah ditunjukan bahwa penggunaan algoritma C4.5, Naive Bayes, K-NN, SVM dan *Logistic Regression* dapat digunakan dalam melakukan klasifikasi data status akreditasi sekolah/madrasah di provinsi KALTIM. Dari hasil pengujian juga ditemukan bahwa algoritma SVM memiliki kinerja terbaik dibandingkan algoritma klasifikasi lainnya, hal ini dibuktikan dengan *accuracy* yang didapatkan algoritma SVM sebesar 88.8%, sedangkan algoritma *Logistic Regression* memiliki akurasi terendah yang dibuktikan dengan *accuracy* sebesar 56.6%. Dari hasil komparasi juga terlihat algoritma K-NN, Naive Bayes dan C4.5 memiliki *accuracy* yang baik dalam melakukan klasifikasi dengan rata-rata *accuracy* lebih dari 80%.

Saran untuk peneliti-peneliti lainnya yang ingin melakukan penelitian dalam melakukan klasifikasi status akreditasi sekolah/madrasah dapat menambahkan beberapa atribut lain yang belum digunakan pada penelitian ini seperti nilai Indikator Pemenuhan Relatif (IPR). Kedepannya juga peneliti lain dapat melakukan optimasi terhadap beberapa algoritma yang digunakan pada penelitian ini agar meningkatkan kinerja algoritma seperti menambah tingkat akurasi dan mempercepat waktu proses algoritma klasifikasi. Terakhir jika BAN S/M Provinsi Kalimantan Timur ingin membuat aplikasi cerdas yang yang dapat digunakan untuk memprediksi status akreditasi sekolah/madrasah dapat menggunakan algoritma *Support Vector Machine* (SVM) dengan kernel *Polynomial* karena memiliki kinerja terbaik dari algoritma klasifikasi lainnya.

 ISSN 2686-3359 (Online)

DAFTAR PUSTAKA

[1] B. – S. BAN – S/M, "Badan Akreditasi Nasional Sekolah/Madrasah," 2021. http://ban-sm.or.id.

[2] B.-S. BAN-S/M, "Instrumen Akreditasi Satuan Pendidikan 2020 Jenjang Sekolah Menengah Atas/ Madrasah Aliyah," 2020. https://bansm.kemdikbud.go.id/ unduh/get/93 (accessed Jun. 05, 2022).

[3] P. A. Octaviani, Yuciana Wilandari, and D. Ispriyanti, "Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang," *J. Gaussian*, vol. 3, no. 8, pp. 811–820, 2014.

[4] B. Merluarini, D. Safitri, and A. Hoyyi, "Perbandingan Analisis Klasifikasi Menggunakan Metode K-Nearest Neighbor (K-NN) dan Multivariate Adapative Regression Spline(MARS) pada Data Akreditasi Sekolah Dasar Negeri di Kota Semarang," *J. Gaussian*, vol. 3, no. 3, pp. 313–322, 2014.

[5] D. Dariyo and R. Dasmita, "Jaringan Syaraf Tiruan Untuk Akreditasi Sekolah Menengah Atas/Madrasah Aliyah," *J. Gaussian*, vol. 3, no. 4, pp. 77–90, 2014.

[6] R. E. Utama, "Klasifikasi Akreditasi Sekolah Menengah Pertama di Pulau Sulawesi Menggunakan Jaringan Syaraf Tiruan Backpropagation," Universtias Sanata Dharma, 2019.

[7] M. Irvan, Y. Purnama, and R. Vhalery, "Model Prediktif Untuk Akreditasi Sekolah Tingkat Sekolah Menengah Pertama (Smp)," *Res. Dev. J. Educ.*, vol. 5, no. 2, p. 03, 2019, doi: 10.30998/rdje.v5i2.3747.

[8] D. F. Tambunan, "Klasifikasi Akreditasi SMA di Pulau Sumatera Menggunakan Metode Naive Bayes," Universitas Sanata Dharma, 2020.

[9] *Peraturan Mendiknas Nomor 29 Tahun 2005*. Indonesia, 2005.

[10] D. A. Adeniyi, Z. Wei, and Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method," *Appl. Comput. Informatics*, vol. 12, no. 1, pp. 90–108, 2016, doi: 10.1016/j.aci.2014.10.001.

[11] P. P. Widodo, R. T. Handayanto, and H. Herlawati, *Penerapan data mining dengan Matlab*. Bandung: Rekayasa Sains, 2013.

[12] K. Kusrini and L. Lutfhi, *Algoritma Data Mining*. Yogyakarta: CV Andi Offset, 2009.

[13] B. Bustami, "Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi," *J. Inform. Ahmad Dahlan*, vol. 8, no. 1, p. 102632, 2014, doi: 10.26555/jifo.v8i1.a2086.

[14] P. Eko, *Data mining : konsep dan aplikasi menggunakan MATLAB*. Andi Offset, 2012.

[15] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika, 2017.

[16] J. Indriyanto, *Algoritma K-Nearest Neighbor Untuk Prediksi Nasabah Asuransi*. Penerbit NEM, 2021.

[17] D. Nofriansyah, *Konsep Data Mining Vs Sistem Pendukung Keputusan*. Yogyakarta: Deepublish, 2015.

[18] R. Primartha, *Algoritma Machine Learning*. Informatika, 2021.

[19] S. Dewi, "Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan," *J. Techno Nusa Mandiri*, vol. 13, no. 1, pp. 60–66, 2016.

[20] M. F. Rahman, D. Alamsah, M. I. Darmawidjadja, and I. Nurma, "Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN)," *J. Inform.*, vol. 11, no. 1, p. 36, 2017, doi: 10.26555/jifo.v11i1.a5452.

[21] R. Rukminingsih, G. Adnan, and M. A. Latief, *Metode Penelitian Pendidikan (Kuantitatif, Kualitatif & Penelitian Tindakan Kelas)*. Yogyakarta: Erhaka Utama, 2020.

[22] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.

[23] M. M. Saim and H. Ammor, "Comparative study of machine learning algorithms (SVM, Logistic Regression and KNN) to predict cardiovascular diseases," in *E3S Web of Conferences*, 2022, vol. 351, p. 5, doi: 10.1051/e3sconf/202235101037.