JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER

VOL. 5. NO. 2 FEBRUARI 2020
P-ISSN: 2685-8223 | E-ISSN: 2527-4864

ANALISIS KINERJA ALGORITMA C4.5 DAN NAÏVE BAYES DALAM MEMPREDIKSI KEBERHASILAN SEKOLAH MENGHADAPI UN

Yeni Angraini[1]; Siti Fauziah[2]; Jordi Lasmana Putra[3]

Ilmu Komputer[1]
STMIK Nusa Mandiri[1]
www.nusamandiri.ac.id[1]
yeniangraini3@gmail.com[1], sitifauziah478@gmail.com[2], jordy.jlp@bsi.ac.id[3]

Abstract— *The national exam (UN) is one of the determinants of student graduation, both elementary school, junior high school and even high school. There are many businesses that are carried out by schools to prepare their students to face national examinations. In fact almost all schools provide material deepening to their students for subjects tested at the national examination. Therefore, this study was conducted to determine the level of success of the school in preparing students in facing national examinations. The method used is a decision tree with C4.5 algorithm and naïve Bayes algorithm. From the results of the study, the results of the accuracy of the naïve bayes algorithm were as big as 95,50% , while accuracy using the c4.5 algorithm is equal to 78,50%. Then it can be concluded that the predictions generated from the naïve bayes algorithm are better compared to the c4.5 algorithm .*

Keywords*: Prediction of UN success, C4.5 Algorithm, Naive Bayes Algorithm.*

Intisari— Ujian nasional (UN) merupakan salah satu penentu kelulusan siswa, baik siswa sekolah dasar, sekolah menengah pertama, bahkan sekolah menengah atas. Banyak usaha yang dilakukan sekolah untuk mempersiapkan siswanya untuk menghadapi ujian nasional. Bahkan hampr semua sekolah memberikan pendalaman materi kepada siswanya untuk mata pelajaran yang di ujikan di ujian nasional. Karena itu penelitian ini dilakukan untuk mengetahui tingkat keberhasilan sekolah dalam mempersiapkan siswanya dalam menghadapi ujian nasional. Adapun metode yang digunakan adalah decision tree dengan algoritma C4.5 dan algoritma naïve bayes. Dari hasil penelitian didapatkan hasil akurasi algoritma naïve bayes sebesar 95,50% , sedangkan akurasi menggunakan algoritma c4.5 sebesar 78,50%. Maka dapat disimpulkan bahwa prediksi yang di hasilkan dari algoritma naïve bayes lebih baik di bandingkan algoritma c4.5.

Kata Kunci*: Prediksi keberhasilan UN, Algoritma C4.5, Algoritma Naive Bayes.*

PENDAHULUAN

Ujian Nasional (UN) merupakan salah satu usaha pemerintah untuk melakukan penjaminan persamaan mutu pendidikan antar daerah yang dilakukan oleh pusat penilaian pendidikan (Hartanto, 2015). Sebagaimana tercantum dalam peraturan pemerintah Nomor 19 tahun 2005. Dalam menghadapi ujian nasional banyak usaha yang dilakukan oleh siswa untuk mempersiapkan diri mereka. Dan bukan hanya para siswa, bahkan pihak sekolahpun ikut berusaha untuk mempersiapkan siswa-siswinya dalam menghadapi ujian nasional.

Data mining merupakan proses atau kegiatan untuk mengumpulkan data yang berukuran besar(Saleh, 2015), dan kemudian mengekstraksi data tersebut menjadi informasi-informasi yang nantinya dapat digunakan. Penerapan data mining dalam sebuah penelitian dapat mengenali suatu pola pengetahuan yang menawarkan solusi untuk masalah pendidikan(Rahman, Muhammad, & Firdaus, 2016).

Dalam penelitian ini peneliti mencoba untuk memproses data rata-rata nilai ujian nasional siswa per matapelajaran untuk mengetahui tingkat sekolah dalam mempersiapkan siswanya untuk menghadapi ujian nasional. Disini tiap sekolah di kelompokan kedalam tiga kelompok yaitu tinggi, sedang, dan rendah berdasarkan rata-rata nilai ujian setiap mata pelajaran yang di ujikan dalam ujian nasional. Adapun algoritma yang digunakan adalah C4.5 (Ridwan, 2017) dan juga Naïve Bayes (Sabilla & Putri, 2017).

Sebelum membahas algoritma C4.5 perlu dijelaskan terlebih dahulu algoritma ID3 karena C4.5 adalah ekstensi dari algoritma decision-tree ID3. Algoritma ID3/C4.5 ini secara rekursif membuat sebuah decision tree berdasarkan training data yang telah disiapkan(Deteksi et al., 2019).

Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistic yang dikemukakan oleh ilmuan Inggris Thomas Bayes(Nawawi et al., 2019), yaitu memprediksi

peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema Bayes. Teorema tersebut dikombinasikan dengan Naïve dimana diasumsikan kondisi antar atribut saling bebas. Persamaan teorema Bayes adalah:

$$P(H|X = \frac{P(X|H).P(H)}{P(X)} \quad \text{..... (1)}$$

Keterangan :
Data dengan class yang belum diketahui
Hipotesis data X merupakan suatu class spesifik
$(H|X)$: Probabilitas hipotesis H berdasar kondisi$X$ (posteriori probability)
(H) : Probabilitas hipotesis H (prior probability)
$(X|H)$: Probabilitas X berdasarkan kondisi pada hipotesis H
$P(X)$: Probabilitas X(Bustami, 2014)

Adapun aplikasi yang digunakan dalam penelitian ini adalah Rapid Miner. Rapid Miner adalah salah satu aplikasi opensource yang dapat digunakan untuk melakukan proses data mining. Salah satu metode data mining adalah menggunakan regresi linier. Regresi linier merupakan metode sattistik yang digunakan untuk untuk melakukan estimasi atau perkiraan berdasarkan data yang ada(Imelda A.Muis & Muhammad Affandes, 2015). RapidMiner menyediakan prosedur data mining dan mechine learning termasuk: ETL (Extraction, Transformation, Loading), data preprocessing, visualisasi, modelling, dan evaluasi.

Bagian – bagian pada tampilan Rapid Miner

A. Tipe Nilai

Pada Tools RapidMiner ada beberapa tipe nilai yang digunakan yaitu:

1. Nominal
   Nominal adalah tipe nilai yang digunakan berdasarkan nilai secara kategori.
2. Numeric
   Nilai numerik secara umum
3. Integer
   Tipe nilai yang digunakan untuk bilangan bulat
4. Real
   Tipe nilai yang digunakan untuk bilangan yang nyata
5. Text
   Tipe nilai yang digunakan untuk teks bebas tanpa struktur.
6. Binomial
   Tipe nilai yang digunakan untuk nilai yang terdiri dari dua nilai
7. Polynomial
   Digunakan untuk nominal lebih dari dua nilai.
8. Date_Time
   Digunakan untuk tanggal dan waktu

B. Prespektif Dan View

Sebuah prespektif berisikan pilihan elemen-elemen GUI , yang disebut dengan View, yang dapat dikonfigurasi secara bebas. Berikut perspective yang terdapat pada tools RapidMiner:

1. Perspekttif Selamat Datang (Welcome prespective)
2. Perspektif desain (Desaign perspective)
3. Perspektif hasil (Result Prespective)(Imelda A.Muis & Muhammad Affandes, 2015)

BAHAN DAN METODE

Data yang di gunakan merupakan data hasil ujian nasional sekolah menengah atas di daerah banda aceh.

Langkah-langkah dalam penelitian ini mengikuti tahapan penambangan data yang dikenal dengan proses Knowledge Discovery in Databases(Pang-Ning, Steinbach, & Kumar, 2006) sebagai berikut:

1. Proses seleksi yaitu proses melakukan pemilihan terhadap sekumpulan data yang dimiliki menjadi data target. Dalam penelitian ini dimiliki data rata-rata hasil Ujian Nasional (UN) siswa setiap sekolah di Banda Aceh tahun 2012. Keseluruhan data tersedia dari seluruh sekolah, seluruh provinsi di Indonesia dan dapat diakses melalui portal Open Data Banda Aceh (OBDA) di link data.bandaacehkota.go.id yang merupakan layanan portal katalog Open Data Pemerintah Kota Banda Aceh yang bertujuan memudahkan masyarakat dalam mendapatkan data dengan format data terbuka. Untuk kepentingan penelitian ini, data target yang akan digunakan adalah data untuk sekolah tingkat Sekolah Menengah Atas provinsi Banda Aceh yang berisi rata-rata hasil nilai ujian tiap sekolah menengah atas.
2. Proses preprocessing. Pada tahap ini dilakukan proses pembersihan data untuk data yang mengandung noise.
3. Proses transformasi. Data yang berasal dari portal Open Data Banda Aceh (OBDA)tersebut akan diolah ke dalam bentuk file yang siap untuk ditambang, dimana isi dari beberapa atribut di ubah enjadi bentuk numerik. Untuk atribut yang di ubah adalah status sekolah yang sebelumnya bernilai negeri atau swasta di ubah menjadi 1 dan 2, dan juga untuk data jurusan yang sebelumnya IPA atau IPS menjadi nilai 1 dan 2. Format data yang digunakan adalah spreadsheet.

4. Proses data mining (penambangan data). Pada tahap ini, data yang diperoleh pada tahap ke 3 dilakukan penambangan data menggunakan dua jenis algoritma yaitu algoritma C4.5 dan juga algoritma Naïve Bayes, dengan tujuan untuk mendapatkan hasil akurasi terbaik di antara kedua algoritma trsebut.
5. Evaluasi dan intepretasi. Pada tahap ini, hasil penambangan data pada tahap ke 4 dievaluasi dan diintepretasikan untuk menghasilkan kesimpulan. Elausi sendiri dilakukan dengan melakukan Confusion Matrix

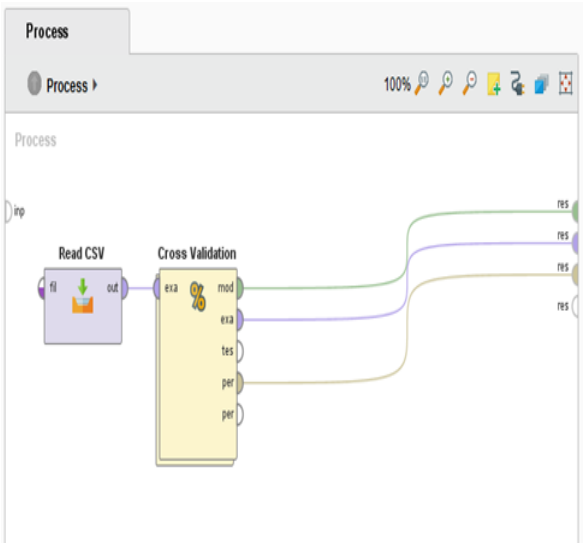
HASIL DAN PEMBAHASAN

Eksperimen dan pengujian Model C4.5

Pembuatan model C4.5 dilakukan pada dataset yang terdiri dari 8 atribut yang merupakan atribut dari prediksi prestasi siswa dan class yang merupakan hasil akhir prediksi. Data kemudian di validasi agar proses pelatihan dapat berjalan dengan cepat dan mampu digunakan untuk melakukan pelatihan.

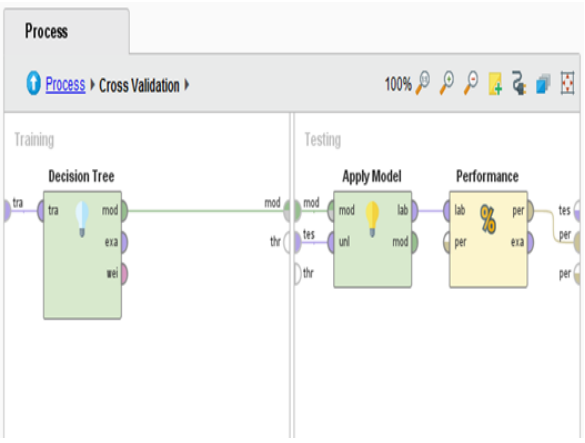
Tahap ini dibuatkan model pengolahan dengan menggunakan perangkat lunak aplikasi Rapidminer seperti gambar 1 di bawah ini:

Jika anda menyajikan tabel harus menyebutkan nama table dan sebutkan sumber tabel, seperti contoh dibawah ini:



Sumber: (Angraini Yeni, Fauziah Siti, 2020)
Gambar1. Model Proses DT

Setelah data dimasukkan ke dalam aplikasi, kemudian pilih algoritma Decision Tree dan tambahkan Apply Model dan Performance untuk menampilkan hasil dari pengolahan data.



Sumber: (Angraini Yeni, Fauziah Siti, 2020)
Gambar2. Model validasi DT

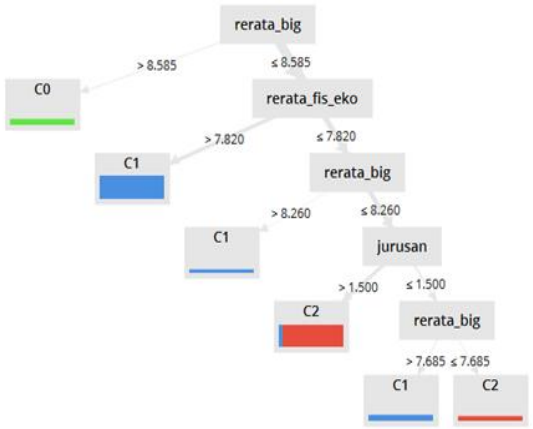
Model dari algoritma C4.5 yaitu berupa pohon keputusan, untuk dapat membuat pohon keputusan, langkah pertama adalah menghitung jumlah class yang tinggi dan yang tidak berprestasi dari masing-masing class berdasarkan atribut yang telah ditentukan dengan menggunakan data training. Kemudian menghitung Entropy (Total) dengan menggunakan persamaan:

$$Entropy(S) = \sum_{i=1}^{n} -pi * log_2\, pi \quad \text{.....(2)}$$

Setelah entropy dari atribut sudah didapat maka langkah berikutnya menghitung gain dengan menggunakan persamaan sebagai berikut:

$$Gain(S, A) = S - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * S_i \quad \text{.....(3)}$$

Setelah didapatkan hasil perhitungan entropy dan gain, maka pohon keputusan yang terbentuk dapat dilihat seperti gambar 3 di bawah ini:



Sumber: (Angraini Yeni, Fauziah Siti, 2020)
Gambar3. Pohon keputusan menggunakan algoritma C4.5

Dari gambar 3 kita dapat menyimpulkan bahwa, sekolah dengan nilai rata-rata bahasa inggris yang tinggi berpotensi untuk mendapatkan nilai rata-rata keseluruhan yang tinggi pula, sehingga tingkat kelulusan siswa di sekolah tersebut semakin besar.

Evaluasi Model Dengan Confusion Matrix

Model confusion matrix akan membentuk matrix yang terdiri dari true C0 atau tupel C0, true C1 atau tupel C1, dan true C2 atau tupel C2, kemudian masukan data testing yang sudah disiapkan ke dalam confusion matrix sehingga didapatkan hasil pada tabel dibawah ini:

Tabel 1. *Confusion Matrix* Algoritma Klasifikasi C4.5

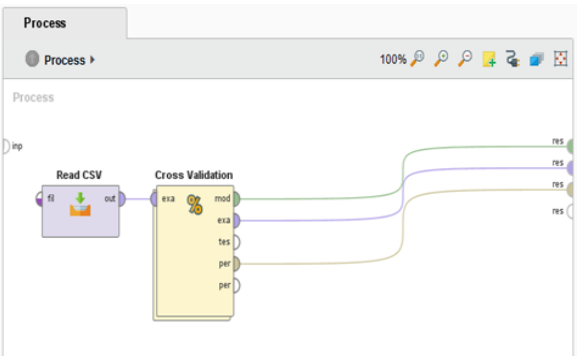
	True C0	*True C1*	*True C2*	Class Precision
Pred. C0	21	1	5	77,78%
Pred. C1	1	3	0	75.00%
Pred. C2	3	0	14	82,35%
Class Recall	84.00%	75.00%	73,68%	

Sumber: (Angraini Yeni, Fauziah Siti, 2020)

Berdasarkan tabel 1 di atas tersebut maka dapat dihitung nilai accuracy dengan menggunakan algoritma C4.5 sebesar 78,50%.

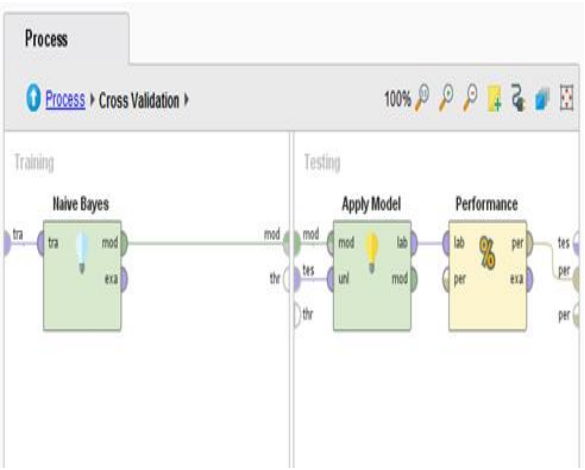
Eksperimen dan Pengujian Model Naïve Bayes

Pembuatan model Naïve Bayes dilakukan pada dataset yang terdiri dari 8 atribut yang merupakan atribut dari prediksi keberhasilan sekolah dalam mempersiapkan siswanya untuk menghadapi ujian nasional dan class yang meupakan hasil akhir prediksi. Data kemudian di validasi agar proses pelatihan dapat berjalan dengan cepat dan mampu digunakan untuk melakukan pelatihan. Tahap ini dibuatkan model pengolahan dengan menggunakan perangkat lunak aplikasi Rapidminer seperti gambar 4 di bawah ini:



Sumber: (Angraini Yeni, Fauziah Siti, 2020)
Gambar4. Model Proses Naïve Bayes

Setelah data dimasukkan ke dalam aplikasi, kemudian pilih algoritma Naïve Bayes dan tambahkan Apply Model dan Performance untuk menampilkan hasil dari pengolahan data.



Sumber: (Angraini Yeni, Fauziah Siti, 2020)
Gambar5. Model validasi Naïve Bayes

Evaluasi Model Dengan Confusion Matrix

Model confusion matrix akan membentuk matrix yang terdiri dari true C0 atau tupel C0, true C1 atau tupel C1, dan true C2 atau tupel C2, kemudian masukan data testing yang sudah disiapkan ke dalam confusion matrix sehingga didapatkan hasil pada tabel 2 dibawah ini:

Tabel 2. *Confusion Matrix* Algoritma Naïve Bayes

	True C0	*True C1*	*True C2*	Class Precision
Pred. C0	24	0	1	96.00%
Pred. C1	1	4	0	80.00%
Pred. C2	0	0	18	100.00%
Class Recall	96.00%	100.00%	794,74%	

Sumber: (Angraini Yeni, Fauziah Siti, 2020)

Berdasarkan tabel 2 di atas tersebut maka dapat dihitung nilai *accuracy* dengan menggunakan algoritma C4.5 sebesar 95,50%.

Komparasi Model Algoritma C4.5 dengan Algoritma *Naïve Bayes*

Hasil pengujian model C4.5 dibandingkan dengan model *Naïve Bayes* dapat dilihat pada tabel berikut:

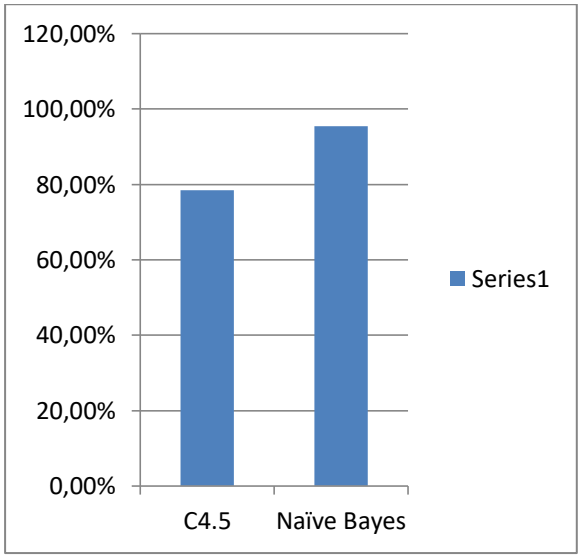
Tabel 3. Pengujian Algoritma C4.5 dan *Naïve Bayes*

	Accuracy
C4.5	78.50%
Naïve Bayes	95,50%

Sumber: (Angraini Yeni, Fauziah Siti, 2020)

Terakreditasi Sinta Peringkat 3 (Sinta 3) berdasarkan SK Dirjen Penguatan RisBang Kemenristekdikti Nomor 28/E/KPT/2019, Tanggal 26 September 2019

Berikut perbedaan Akurasi Algoritma C4.5 dan Algoritma *Naïve Bayes,* dapat dilihat pada gambar di bawah ini:



Sumber: (Angraini Yeni, Fauziah Siti, 2020)
Gambar6. Perbandingan Accuracy Algoritma C4.5 dan Naïve Bayes

Dari hasil pengujian di atas, dengan dilakukan evaluasi baik secara confusion matrix terbukti bahwa maupun ROC terbukti bahwa pengujian yang dilakukan oleh algoritma Naïve Bayes memilki nilai akurasi yang lebih tinggi dibanding C4.5. Nilai akurasi untuk model algoritma Naïve Bayes sebesar 95,50% dan nilai akurasi model algoritma C4.5 sebesar 78,50% . Berdasarkan nilai tersebut diperoleh selisih akurasi sebesar 17%.

KESIMPULAN

Dalam penelitian ini dilakukan analisa dan komparasi dua metode klasifikasi data mining yang memiliki karakteristik yang berbeda yaitu, algoritma C4.5 atau Decision Tree yaitu metode yang mengubah data menjadi pohon keputusan dan menghasilkan rule, dan Naïve Bayes yaitu metode dengan menghitung probabilitas kemunculan data antara satu data dengan yang lain. Dari hasil pengujian dengan mengukur kinerja kedua metode tersebut menggunakan confusion matrix, kurva ROC dan t-Test pada dataset diketahui bahwa Naïve Bayes memiliki nilai akurasi 95,50% dan signifikan terhadap algoritma C4.5 memiliki nilai akurasi 78,50%. Adapun model yang telah terbentuk selanjutnya dapat dikembangkan dan dapat diimplementasikan ke dalam sebuah aplikasi sehingga dapat membantu dan memudahkan bagi para pemegang kepentingan dalam pengambilan sebuah keputusan untuk memprediksi prestasi siswa.

REFERENSI

Angraini Yeni, Fauziah Siti, L. P. J. (2020). *Laporan Akhir Penelitian Mandiri:Analisis Kinerja Algoritma C4.5 Dan Naïve Bayes Dalam Memprediksi Keberhasilan Sekolah Menghadapi UN*. Jakarta.

Bustami. (2014). Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi. *Jurnal Informatika (Yogyakarta)*, *8*(1), 884–898. https://doi.org/10.26555/jifo.v8i1.a2086

Imelda A.Muis & Muhammad Affandes, M. . (2015). Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet. *Sains, Teknologi Dan Industri.UIN Sultan Syarif Kasim Riau*, *12*(2), 189–197. Retrieved from http://103.193.19.206/index.php/sitekin/article/view/1010

Nawawi, H. M., Purnama, J. J., Hikmah, A. B., Komputer, S. I., Informasi, S. S., Bina, U., & Informatika, S. (2019). KOMPARASI ALGORITMA NEURAL NETWORK DAN NAÏVE BAYES, *15*(2), 189–194. https://doi.org/10.33480/pilar.v15i2.669

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). Introduction to data mining: Solution Manual. *Library of Congress*, 796. https://doi.org/10.1016/0022-4405(81)90007-8

Rahman, F., Muhammad, D., & Firdaus, I. (2016). Penerapan Data Mining Metode Naïve Bayes Untuk Prediksi Hasil Belajar Siswa Sekolah Menengah Pertama (Smp). *Al Ulum Sains Dan Teknologi*, *1*(2), 76–78. Retrieved from http://ojs.uniska-bjm.ac.id/index.php/JST/article/view/436

Ridwan, M. (2017). SISTEM REKOMENDASI PROSES KELULUSAN MAHASISWA BERBASIS ALGORITMA KLASIFIKASI C4.5. *Jurnal Ilmiah Informatika*, *2*(1), 105–111. https://doi.org/10.5281/JIMI.V2I1.27

Sabilla, W. I., & Putri, T. E. (2017). Prediksi Ketepatan Waktu Lulus Mahasiswa dengan k-Nearest Neighbor dan Naïve Bayes Classifier. *Jurnal Komputer Terapan*, *3*(2), 233–240.

Retrieved from https://jurnal.pcr.ac.id/index.php/jkt/article/view/1544

Saleh, A. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Citec Journal*, *2*(3), 207–217. https://doi.org/doi.org/10.24076/citec.2015v2i3.49