Komparasi Algoritma Klasifikasi untuk Prediksi Minat Sekolah Tinggi Pelajar pada *Students Alcohol Consumption*

M. Rangga Ramadhan Saelan[1)], Deni Anugrah Sahputra[2)], Widiastuti[3)], Windu Gata[4)]

[1)2)3)4)] Program Studi Magister Ilmu Komputer, STMIK Nusa Mandiri
Jl. Kramat Raya No. 18, Jakarta Pusat, Jakarta
[1)] rangga.mgg@nusamandiri.ac.id
[2)] 14002345@nusamandiri.ac.id
[3)] widiastuti.wtu@nusamandiri.ac.id
[4)] windu@nusamandiri.ac.id

Abstrak

Terdapat banyak faktor yang menjadi kriteria penentu kinerja pelajar salah satu diantaranya adalah konsumsi alkohol oleh pelajar, hal ini dapat mempengaruhi pengambilan keputusan negatif yang menjadi faktor keberhasilan kinerja pelajar. Pada penelitian ini digunakan teknik klasifikasi untuk memprediksi minat pelajar dalam mengambil Langkah untuk melanjutkan pendidikan ke jenjang yang lebih tinggi yang dipengaruhi oleh berbagai faktor, salah satunya adalah tingkat konsumsi alkohol oleh pelajar. Dengan membuat model menggunakan algoritma klasifikasi Naïve Bayes dan Decision Tree yang diujikan pada data konsumsi alkohol oleh pelajar menggunakan *tools* Rapid Miner. Kemudian model yang dihasilkan dikomparasi untuk menentukan algoritma terbaik dalam mengidentifikasi kinerja pelajar. Dengan menggunakan Teknik Cross Validation didapatkan statistik yang menunjukan bahwa Algoritma Decision Tree memiliki kinerja lebih baik jika diabandingkan dengan Naïve Bayes. Algoritma Decision Tree memiliki tingkat akurasi sebesar 86.44% sedangkan Naïve Bayes hanya memiliki tingkat akurasi sebesar 82.60%. Dan berdasarkan statistic ROC, bisa dikatakan bahwa Naïve Bayes memiliki kinerja yang cukup buruk dengan tingkat Equal Error Rate (EER) sebesar 65%, sedangkan Decision Tree memiliki tingkat EER lebih rendah yaitu sebesar 55%. Dengan begitu algoritma Decision Tree memiliki kinerja lebih baik dalam mengidentifikasi kinerja pelajar dengan pengaruh berbagai faktor salah satunya alkohol.

Kata kunci: Pelajar, alkohol, algoritma.

Abstract

There are many factors that are determinants of student performance, one of which is the consumption of alcohol by students, this can affect negative decision making which is a factor in the success of student performance. In this study classification techniques are used to predict students' interest in taking steps to continue their education to a higher level which is influenced by various factors, one of which is the level of alcohol consumption by students. By creating a model using the Naïve Bayes classification algorithm and Decision Tree which are tested on alcohol consumption data by students using Rapid Miner tools. Then the resulting model is compared to determine the best algorithm in identifying student performance. By using the Cross Validation Technique, statistics are obtained that show that the Decision Tree Algorithm has better performance when compared to Naïve Bayes. Decision Tree algorithm has an accuracy rate of 86.44% while Naïve Bayes only has an accuracy rate of 82.60%. And based on ROC statistics, it can be said that Naïve Bayes has quite poor performance with an Equal Error Rate (EER) of 65%, while Decision Tree has a lower EER level of 55%. That way the Decision Tree algorithm has a better performance in identifying student performance with the influence of various factors one of which is alcohol.

Keywords: *Students, alcohol, algorithms.*

DOI: 10.34128/jsi.v6i2.236 *Received: July 25th, 2020*
Accepted: September 15th, 2020

1. PENDAHULUAN

Meskipun tidak sepenuhnya menjadi penentu seorang pelajar dikatakan memiliki kinerja yang baik atau buruk di sekolah menengah atas, tetapi minat lanjut sekolah tinggi demi memiliki pengalaman belajar setinggi mungkin adalah salah satu bahan yang menjadi penilaian bahwa seorang pelajar tersebut bisa dikatakan memiliki kinerja yang cukup baik. Dalam dunia Pendidikan saat ini, banyak faktor yang mempengaruhi kinerja pelajar di sekolah, salah satunya dengan melihat sisi keaktifan dan keberminatannya dalam melanjutkan sekolah dengan jenjang yang lebih tinggi. Dengan minat dan keinginan yang tinggi seperti itu, akan menjadi pembuka dari faktor keberhasilan pelajar tersebut. Tingkat keberhasilan siswa mencerminkan keberhasilan organisasi pendidikan, maka tren peningkatan keberhasilan siswa menjadi tujuan semua organisasi Pendidikan [1]. Penelitian mengenai dunia Pendidikan sangat penting bagi setiap Lembaga Pendidikan, bahan evaluasi yang dihasilkan dari penelitian sangat dibutuhkan untuk membuat mengambil keputusan demi menciptakan sistem Pendidikan yang baik sehingga Lembaga dapat menciptakan lulusan berkualitas, tidak hanya dari sisi akademik (*hard skill*) tetapi juga dari sisi non akademiknya (*soft skill*).

Tingkat konsumsi alkohol menjadi salah satu hal yang mempengaruhi faktor tingkat kinerja pelajar. Hal ini dapat mempengaruhi pengambilan keputusan negatif yang menjadi faktor keberhasilan kinerja pelajar. Tentunya hal ini harus di cegah dengan mengambil Langkah awal, prediksi awal terhadap seorang pelajar yang mengkonsumsi alkohol dapat mencegah resiko kegagalan yang dialami oleh pelajar dengan sistem Pendidikan dari hasil evaluasi yang didapatkan. Pada tahun 2016, National Institute of Health melaporkan bahwa 26% siswa kelas 8, 47% siswa kelas 10, dan 64% siswa kelas 12 semuanya memiliki pengalaman dalam mengonsumsi minuman beralkohol. Temuan ini menunjukkan tren percepatan penggunaan alkohol di kalangan siswa sekolah, karenanya kekhawatiran yang berkembang di kalangan masyarakat [2]. Hal ini menunjukan bahwa konsumis alkohol dikalangan pelajar khususnya tingkata menengah atas sangat memprihatinkan, Tentunya ini adalah sebuah masalah yang harus diatasi.

Metode Data Mining diterapkan pada data pendidikan dengan tujuan meningkatkan metode pengajaran, meningkatkan kualitas pengajaran, mengidentifikasi siswa yang lemah, mengidentifikasi faktor-faktor yang mempengaruhi kinerja akademik siswa. Pemanfaatan metode Data Mining ini untuk meningkatkan kualitas pendidikan, mengidentifikasi siswa yang membutuhkan peningkatan disebut sebagai Data Mining Pendidikan [3]. EDM telah menjadi minat penelitian utama bagi banyak peneliti. Fungsi utama dari Data Mining pendidikan adalah prediksi kinerja akademik siswa. Memprediksi kinerja akademik siswa membantu dalam mengidentifikasi sejumlah hal seperti siswa yang cenderung putus sekolah, siswa yang lemah dan membutuhkan peningkatan, siswa yang baik dalam bidang akademik tetapi belakangan memburuk [3].

Model klasifikasi dapat digunakan untuk mengidentifikasi kinerja pelajar berdasarkan keaktifan dan minat pelajar dalam memiliki jenjang sekolah yang lebih tinggi yang dipengaruhi oleh berbagai faktor salah satunya adalah tingkat konsumsi alkohol pada pelajar sekolah menengah atas . Pada penelitian ini menggunakan model algoritma klasifikasi naïve bayes dan Decision Tree, model yang dihasilkan akan dibandingkan berdasarkan kinerja kedua model algoritma tersebut, untuk mendapatkan algoritma terbaik pada klasifikasi untuk prediksi data konsumsi alkohol pelajar negara portugal.

2. TINJAUAN PUSTAKA

2.1 Klasifikasi

Tujuan dari *learning supervised* adalah untuk memasukkan klasifikasi untuk nilai respon kategori untuk memisahkan data menjadi kelas-kelas tertentu. Klasifikasi adalah proses dua langkah. Model dibuat pada langkah pertama menggunakan algoritma klasifikasi. Pada langkah kedua model dilatih, dan kinerja serta akurasi diukur. Tujuan utama klasifikasi adalah untuk mengklasifikasikan data ke dalam kelas yang berbeda sesuai dengan kendala. Ini digunakan untuk memprediksi kelas target dengan menganalisis dataset pelatihan. Untuk menentukan setiap kelas target, klasifikasi menggunakan data pelatihan yang ditetapkan untuk menemukan batas yang tepat. Setelah tugas selesai, ia membuat prediksi nilai respons. Proses ini dikenal sebagai klasifikasi [4].

2.2 Decision Tree

Decision Tree adalah algoritma penambangan data populer untuk klasifikasi. C4.5 adalah versi diperpanjang dari algoritma ID3 tradisional yang merupakan model prediksi berbasis struktur pohon. Simpul root berisi atribut paling efektif dari dataset yang tidak memiliki tepi masuk dan setiap tepi diberi label oleh nilai atau rentang tertentu. Node keputusan adalah simpul perantara yang melakukan beberapa pengujian pada atribut dan simpul daun berisi label kelas. Decision Tree membagi data menjadi set kereta dan set tes, dalam aturan fase pelatihan dihasilkan dari data kereta masuk yang digunakan dalam fase Pengujian untuk memprediksi label kelas dengan menganalisis data uji [5].

Atribut efektif dari rangkaian pelatihan dipilih dengan menggunakan rasio perolehan informasi dalam algoritma C4.5 di mana dalam algoritma Dec3 Tree Id3 hanya menggunakan penguatan informasi. Empat hal dimasukkan ke dalam Pohon Keputusan C4.5 yang tidak ada dalam Id3 yaitu,

1) Pemilihan atribut menggunakan rasio perolehan informasi
2) Penanganan nilai yang hilang
3) Dukungan untuk kedua atribut diskrit dan lanjutan
4) Pemangkasan pohon sambil membangun Pohon Keputusan

Entropy Decision Tree:

$$\text{Entropy (S)} \sum_{j=1}^{c} (p(S_1)\log_2(S_1)) \quad (1)$$

Dimana, C adalah kelas dan p adalah proporsi instance dalam S yang ditugaskan ke kelas J.

2.3 Naïve Bayes

Naïve Bayes adalah teknik klasifikasi yang sederhana dan mudah: Naïve Bayes mengklasifikasikan probabilitas posterior dan sebelumnya untuk menemukan kelas tertentu [6].

Teorema Bayes adalah:

$$P(M \mid N) = P(N \mid M) .P(N) P(M) \quad (2)$$

Di mana,
M- Beberapa hipotesis, sehingga data tuple N termasuk kelas yang ditentukan C
N- Beberapa bukti, menggambarkan dengan mengukur pada set atribut
P (M | N) - probabilitas posterior bahwa hipotesis M memegang memberikan bukti N
P (M) - probabilitas sebelum M, independen pada N
P (N | M) - probabilitas posterior bahwa dari N dikondisikan pada M.
Keuntungan:

1) Teknik ini bekerja dengan baik pada data numerik dan juga tekstual.
2) Penggolong ini mudah diimplementasikan dan perhitungannya sederhana dibandingkan dengan algoritma lainnya.
3) Karena dapat diterapkan pada kumpulan data besar, tidak diperlukan skema estimasi parameter iteratif yang rumit.
4) Mudah interpretasi representasi pengetahuan.
5) Berkinerja baik dan kuat.

Kekurangan:

1) Tidak mempertimbangkan frekuensi kemunculan kata.
2) Secara teoritis, classifier naif Bayes memiliki tingkat kesalahan minimum jika dibandingkan dengan classifier lain, tetapi secara praktis hal itu tidak selalu benar, karena asumsi independensi kelas bersyarat.

2.4 Penelitian Terkait

Penelitian ini tidak terlepas dari penelitian-penelitian sebelumnya yang menjadi acuan atas penulisan ini, diantaranya:

2.4.1 Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi Alkohol Siswa

Penelitian ini bertujuan untuk menerapkan dan melakukan analisis kinerja algoritma data mining untuk memprediksi konsumsi alkohol dan menganalisis faktor-faktor yang terkait pada siswa tingkat menengah. Adapun tahapan yang dilakukan ialah praprocess data, seleksi fitur, klasifikasi, dan evaluasi model. Pada tahap preprocess, beberapa fitur diubah menjadi bentuk yang sesuai untuk memudahkan proses klasifikasi. Selanjutnya, algoritma Gain Ratio dan Fast Correlation Based Filter (FCBF) digunakan untuk memilih fitur-fitur yang relevan dan penting untuk digunakan dalam tahapan klasifikasi. Decision Tree C5.0, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), dan Naive Bayes (NB) dieksekusi pada kelompok fitur yang terpilih. Akurasi model yang dibangun dievaluasi menggunakan 10-fold Cross-Validation (CV). Hasil penelitian menunjukkan bahwa model klasifikasi yang dibangun menggunakan Naïve Bayes memiliki nilai akurasi tertinggi dengan menggunakan 5 fitur terbaik dari Gain Ratio. Selain itu, penggunaan metode pemilihan fitur mampu meningkatkan performa dari seluruh klasifier secara umum. Pengujian lebih lanjut pada data yang sama maupun berbeda perlu dilakukan untuk mendapatkan gambaran lebih mendalam mengenai kinerja algoritma-algoritma yang digunakan [7].

2.4.2 Predicting Alcohol Consumption Behaviours of the Secondary Level Students

Data mining digunakan untuk prediksi motif minum. Kelemahan dari model penambangan data yang ada yang menggunakan sistem pra-pemrosesan adalah bahwa ia tidak mengidentifikasi atribut yang relevan yang secara efektif berkontribusi pada prediksi intensitas konsumsi alkohol siswa sekolah menengah. Untuk mengatasi keterbatasan ini, kami menyelidiki atribut yang paling relevan dari kinerja siswa sekolah menengah dengan menggunakan sistem yang diusulkan dan mendapatkan fakta berharga tentang perilaku siswa. Sistem yang diusulkan tergantung pada dua kali lipat dari pra-pemrosesan menggunakan diskritisasi dan pemilihan fitur, yang disebut sebagai Multistage Pra-pemrosesan (MSP) untuk penyelidikan pilihan subset fitur yang optimal dan klasifikasi yang baik. Sistem yang diusulkan mampu memprediksi apakah siswa kecanduan alkohol atau tidak dengan intensitas. Kami melakukan eksperimen komprehensif dengan bantuan pemilih fitur, yaitu pemilihan fitur berbasis korelasi (CFS), Information Gain (IG), Chi-Square (CS), dan Relief-F menggunakan pengklasifikasi yang berbeda. Hasil percobaan menunjukkan bahwa metode pemilihan fitur masing-masing meningkatkan kinerja klasifikasi berdasarkan akurasi, sensitivitas, presisi, ukuran-f, dan ROC. Seperti yang ditunjukkan dalam hasil yang diperoleh, kinerja tertinggi dicapai dalam hal akurasi 71,39%, sensitivitas 71,34%, presisi 66,86%, F-mengukur 68,43%, dan area ROC masing-masing 85,89%. Kata kunci: akurasi, konsumsi alkohol, pemilihan fitur [8].

2.4.3 Prediction of Alcohol Consumption among Portuguese Secondary School Students: A Data Mining Approach

Makalah ini diatur untuk melakukan percobaan perbandingan pada prediksi konsumsi alkohol di kalangan siswa sekolah menengah. Kumpulan data yang digunakan dalam proyek ini berisi 34 atribut yang dikumpulkan dari dua sekolah menengah Portugis pada tahun 2005-2006. Empat algoritma klasifikasi diusulkan dan diimplementasikan, yang meliputi Decision Tree, kNearest Neighbor (k-NN), Random Forest dan Naïve Bayes. Metode ini dilatih dan diuji menggunakan validasi silang 10 kali lipat. Hasil penelitian menunjukkan bahwa algoritma Pohon Keputusan menghasilkan nilai tertinggi untuk akurasi, penarikan dan presisi dibandingkan dengan algoritma klasifikasi lainnya. Selain itu, diamati bahwa metode Naïve Bayes dikombinasikan dengan normalisasi Interkuartil memberikan teknik klasifikasi alternatif yang menjanjikan di daerah tersebut [9].

3. METODE

3.1 Desain Penelitian

Penelitian ini menggunakan metode komparatif yaitu dengan melakukan perbandingan dua model algoritma yang berbeda, berdasarkan kinerja yang dihasilkan dari proses identifikasi model terhadap data penelitian dengan melihat statistik dari masing-masing model yang dujikan. Penelitian ini bertujuan membandingkan dan mengevaluasi kedua metode algoritma klasifikasi

p-ISSN: 2460-173X
e-ISSN: 2598-5841

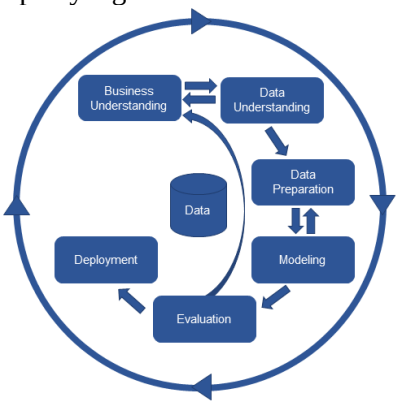
Naïve Bayes dan Decision Tree dalam mengidentifikasi minat pelajar menengah atas untuk melanjutkannya pendidikan ke jenjang yang lebih tinggi yang di pengaruhi oleh beberapa faktor khususnya pengaruh konsumsi alkohol pada pelajar.

3.2 Pengumpulan Data

Data penelitia ini menggunakan *data public* yang diambil dari situs *Kaggle.com.* Sumber data ini berasal dari negara Portugal dari data P. Cortez and A. Silva. *Using Data Mining to Predict Secondary School Student Performance*. *Data public* ini memiliki 33 atribut dengan jumlah *record* sebanyak 649 responden. 69 pelajar memilih untuk tidak akan melanjutkan nya keperguruan tinggi, dan 580 pelajar memilih untuk melanjutkan sekolah ke perguruan tinggi.

3.3 Metode Penelitian

Pada penelitian terdpat 6 tahapan yang dilakukan berdasarkan model eksperimen CRISP-DM



Gambar 1. *CRISP-DM Approach* [10]

3.3.1 Tahap Business Understanding

Pada tahap ini ditujukan untuk menentukan tujuan penelitian yaitu dengan mengimplementasikan beberapa metode terhadap data (*Student Consumption Alcohol*) dalam upaya meningkatkan nilai akurasi, yang kemudian hasil dari akurasi tersebut akan di komparasikan untuk melihat perbandingan dari kedua algoritma tersebut.

3.3.2 Tahap Data Understanding

Data yang digunakan adalah data sekunder/*public* yang diperoleh dari situs *public* yaitu kaggle (*www.kaggle.com*), data tersebut terdiri dari 33 atribut dan 649 record.

Tabel.1. Describe Atribute

Atribute	Record
School	*student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)*
Sex	*student's sex (binary: 'F' - female or 'M' - male)*
Age	*student's age (numeric: from 15 to 22)*
Address	*student's home address type (binary: 'U' - urban or 'R' - rural)*
Famsize	*family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)*
Pstatus	*parent's cohabitation status (binary: 'T' - living together or 'A' - apart)*
Medu	*mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)*
Fedu	*father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)*
Mjob	*mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')*
Fjob	*father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')*

Reason	*reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')*
Guardian	*student's guardian (nominal: 'mother', 'father' or 'other')*
Traveltime	*home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)*
Studytime	*weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)*
Failures	*number of past class failures (numeric: n if 1<=n<3, else 4)*
Schoolsup	*extra educational support (binary: yes or no)*
Famsup	*family educational support (binary: yes or no)*
Paid	*extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)*
Activities	*extra-curricular activities (binary: yes or no)*
Nursery	*attended nursery school (binary: yes or no)*
Higher	*wants to take higher education (binary: yes or no)*
Internet	*Internet access at home (binary: yes or no)*
Romantic	*with a romantic relationship (binary: yes or no)*
Famrel	*quality of family relationships (numeric: from 1 - very bad to 5 - excellent)*
Freetime	*free time after school (numeric: from 1 - very low to 5 - very high)*
Gout	*going out with friends (numeric: from 1 - very low to 5 - very high)*
Dalc	*workday alcohol consumption (numeric: from 1 - very low to 5 - very high)*
Walc	*weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)*
Health	*current health status (numeric: from 1 - very bad to 5 - very good)*
Absences	*number of school absences (numeric: from 0 to 93)*
G1	*first period grade (numeric: from 0 to 20)*
G2	*second period grade (numeric: from 0 to 20)*
G3	*final grade (numeric: from 0 to 20, output target)*

3.3.3 Tahap Data Preparation

Pada tahap ini meliputi pengolahan data yang didapat dari situs kaggle. Pengolahan data ini bertujuan untuk membangun dataset akhir yang akan di proses pada tahap pemodelan, mencakup pemilihan tabel, atribut-atribut data dan transformasi data. Kemudian dilakukan proses pembersihan data (data cleaning). Hal yang dilakukan pada proses ini, diantaranya: mencoba menghilangkan missing value, melancarkan noise, dan membenarkan inconsistencies dalam dataset yang dilakukan secara manual.

3.3.4 Tahap Modelling

Tahap modelling dilakukan setelah tahap preparation, yaitu melakukan pemodelan dengan menggunakan metode Naïve Bayes dan Random Forest.

3.3.5 Tahap Evaluation

Pada tahap ini melakukan evaluasi terhadap model-model yang telah terbentuk sehingga mendapatkan informasi model yang akurat.

3.3.6 Tahap Deployment

Dari model yang telah dihasilkan maka perlu diuji dengan menggunakan data baru dan dilakukan kembali evaluasi untuk keakuratan data.

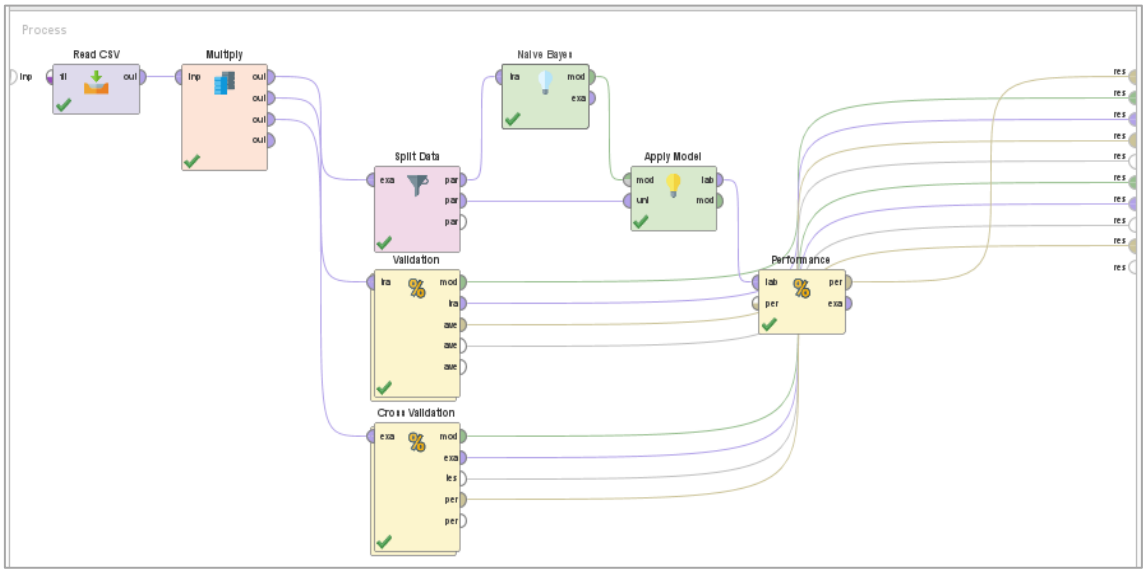
4. PEMBAHASAN

Hasil dari penelitian ini yaitu menguji keakuratan dari Analisa identifikasi pengaruh konsumsi alkohol terhadap kinerja pelajar berdasarkan prediksi minat untuk melanjutkan sekolah ke jenjang yang lebih tinggi dengan menggunakan model algoritma Naïve Bayes dan Decision Tree yang kemudian hasil dari pengujian model akan di bandingkan untuk mengetahui algoritma terbaik pada klasifikasi untuk prediksi data konsumsi alkohol pelajar negara Portugal.

p-ISSN: 2460-173X
e-ISSN: 2598-5841

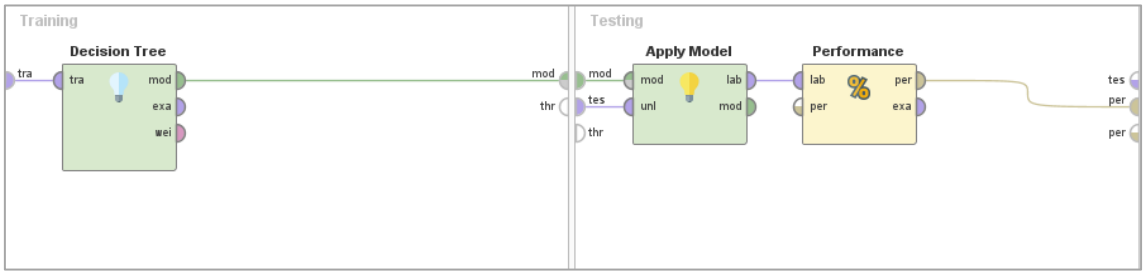
4.1 Pengujian Model dengan *Software*

Pada penelitian ini digunakan *software* Rapid Miner sebagai tools untuk menguji model algoritma yang di pakai. Model akan digunakan untuk mengolah data kemudian membandingkan kinerja dari hasil setiap algoritma yang telah di uji terhadap *data public*.



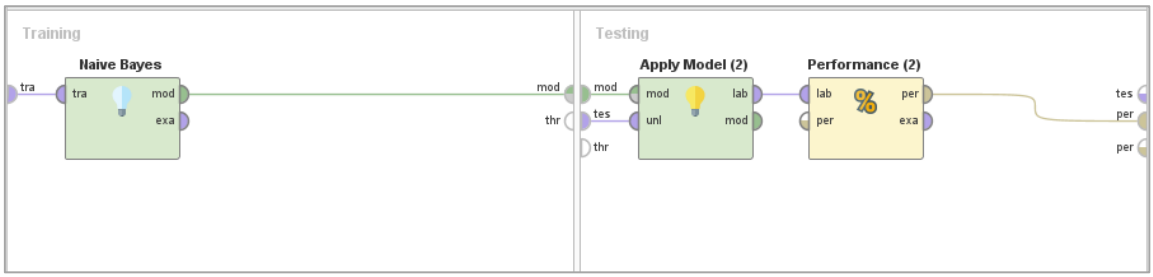
Gambar 2. Model pegujian algoritma

Model ini memiliki fitur multiply yang berguna untuk membuat cabang validation sehingga pada pengujian ini dapat menguji lebih dari 1 model algoritma.



Gambar 3. Model pengujian algoritma decision tree

Didalam x-validation terdapat proses pemodelan menggunakan algoritma Decision Tree, sehingga nantinya akan menghasilkan akurasi dari pengujian terhadap data public.



Gambar 4. Model pengujian algoritma naïve bayes

Didalam x-validation(2) terdapat proses pemodelan menggunakan Naïve Bayes, sehingga nantinya akan menghasilkan akurasi dari pengujian terhadap data public.

4.2 Hasil Pengujian Model

Berdasarkan model yang yang telah di bentuk kemudian dilakukan pengujian untuk mengukur tingkat akurasi dari dan membandingkan hasil pengujian dari kedua algoritma (Naïve Bayes dan Decision Tree), dengan menggunakan metode cross validation.

4.2.1 Confusion Matrix Naïve Bayes

Berdasarkan evaluasi yang di dapat dari table confusion matrix pada pengujian algoritma Naïve Baye, memiliki tingkat akurasi sebesar 82.60%, dengan 490 data yes di klasifikasi kan sebagai true yes, dan sebanyak 23 data yes sebagai false yes, kemudian 90 data no di klasifikasikan sebagai false no dan 46 data no sebagai true no.

Tabel 2. Confusion matrix naïve bayes

	true yes	true no	class precision
pred. yes	490	23	95.52%
pred. no	90	46	33.82%
class recall	84.48%	66.67%	

4.2.2 Confusion Matrix Decisioin Tree

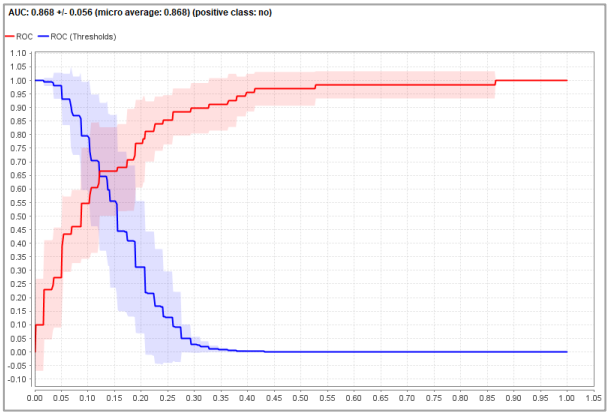
Berdasarkan evaluasi yang di dapat dari table confusion matrix pada pengujian algoritma Decision Tree, memiliki tingkat akurasi sebesar 86.44%, dengan 559 data yes di klasifikasi kan sebagai true yes, dan sebanyak 67 data yes sebagai false yes, kemudian 21 data no di klasifikasikan sebagai false no dan 2 data no sebagai true no.

Tabel 3. Confusion matrix decision tree

	true yes	true no	class precision
pred. yes	559	67	89.30%
pred. no	21	2	8.70%
class recall	96.38%	2.90%	

4.2.3 Kurva ROC Naïve Bayes

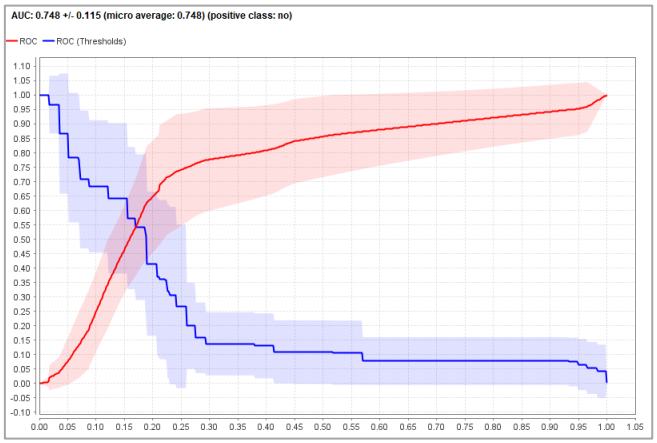
Kurva ROC terbentuk berdasarkan nilai yang telah dihasilkan pada perhitungan Tabel Confusion Matrix.



Gambar 5. Kurva ROC naïve bayes

Pada kurva di atas menunjukan ROC berada pada posisi true no sebesar 0.868 dengan tingkat EER (Equal Error Rate) sebesar 65%.

4.2.4 Kurva ROC Decision Tree



Gambar 6. Kurva ROC decision tree

Sedangkan pada kurva di atas menunjukan ROC berada pada posisi true no sebesar 0.748 dengan tingkat EER (Equal Error Rate) sebesar 55%.

4.3 Komparasi Hasil identifikasi algoritma

Berdasarkan model-model yang telah diuji menggunakan algoritma Naïve Bayes dan Decision Tree dengan metode Cross Validation terdapat perbedaan hasil statistic dari masing-masing model algoritma.

Tabel 4. Komparasi confusion matrix

	Naïve Bayes	Decision Tree
Accuracy	82.60%	86.44%
Precison	33.82%	8.70%
Recall	66.67%	2.90%

Berdasarkan tabel diatas terlihat algoritma Decisson Tree memiliki tingkat akurasi yang paling tinggi sebesar 86.44% dibandingkan dengan algoritma Naïve Bayes yang memiliki tingkat akurasi sebesar 82.60%. tingkat precision dan recall dari masing-masing algoritma berdasarkan class true no memiliki hasil sebagaimana yang tertera pada tabel 4.

Tabel 5. Komparasi kurva ROC

	Naïve Bayes	Decision Tree
EER	65%	55%

Melihat dari tingkat Equal Error Rate (EER) pada masing-masing algoritma, Naïve Bayes memiliki EER sebesar 65% dan Decision Tree sebesara 55%. Dari statistic ini bisa dikatakan bahwa berdasarkan model yang diujikan Naïve Bayes memiliki kinerja yang cukup buruk dikarenakan tingkat EER yang lebih tinggi jika dibandingkan dengan algoritma Decision Tree yang memiliki EER dibawah algoritma Naïve Bayes.

5. KESIMPULAN

Algoritma Naïve Bayes memiliki akurasi sebesar 82.60% sedangkan pada pengujian algoritma Decision Tree memiliki akurasi sebesar 86.44%. Dengan melihat statistik dari kurva ROC terdapat *Equal Error Rate* (EER) dari masing-masing algoritma, Naïve Bayes memiliki tingkat EER sebesar 65%, sedangkan Decision Tree memiliki tingkat EER sebesar 55%, berdasarkan

tingkat EER dapat diartikan bahwa algoritma Naïve Bayes memiliki kinerja lebih buruk jika dibandingkan dengan algoritma Decision Tree yang memiliki tingkat EER lebih rendah.

Berdasarkan statistic yang dihasilkan, dapat disimpulkan bahwa identifikasi klasifikasi untuk prediksi minat lanjut sekolah pada data *Student Consumption Alcohol*, dalam hal ini kinerja algoritma Decision Tree lebih baik jika dibandingkan dengan algoritma Naïve bayes.

DAFTAR PUSTAKA

[1] S. Pal and V. Chaurasia, "Is Alcohol Affect Higher Education Students Performance: Searching and Predicting Pattern Using Data Mining Algorithms," *SSRN Electron. J.*, vol. 6, no. 4, pp. 8–17, 2017, doi: 10.2139/ssrn.2991214.

[2] I. Print, I. Online, I. Cd-rom, and A. K. Hamoud, "Research in Science , Technology , Engineering & Mathematics Selection of Best Decision Tree Algorithm for Prediction and Classification of Students ' Action," no. October, pp. 26–32, 2016.

[3] S. Roy and A. Garg, "Predicting academic performance of student using classification techniques," *2017 4th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Comput. Electron. UPCON 2017*, vol. 2018-Janua, pp. 568–572, 2017, doi: 10.1109/UPCON.2017.8251112.

[4] T. Bhardwaj and P. Somvanshi, *Machine Intelligence and Signal Analysis*, vol. 748. Springer Singapore, 2019.

[5] A. Koli and S. Shinde, "Parallel decision tree with map reduce model for big data analytics," *Proc. - Int. Conf. Trends Electron. Informatics, ICEI 2017*, vol. 2018-Janua, pp. 735–739, 2018, doi: 10.1109/ICOEI.2017.8300800.

[6] . O. A., "Comparative Study of Classification Algorithm for Text Based Categorization," *Int. J. Res. Eng. Technol.*, vol. 05, no. 02, pp. 217–220, 2016, doi: 10.15623/ijret.2016.0502037.

[7] N. Sagala and H. Tampubolon, "Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi Alkohol Siswa," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 2, p. 98, 2018, doi: 10.23917/khif.v4i2.7061.

[8] A. K. Shukla, P. Singh, and M. Vardhan, "Predicting Alcohol Consumption Behaviours of the Secondary Level Students," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3170173.

[9] S. Ismail, N. I. A. N. Azlan, and A. Mustapha, "Prediction of alcohol consumption among Portuguese secondary school students: A data mining approach," *ISCAIE 2018 - 2018 IEEE Symp. Comput. Appl. Ind. Electron.*, pp. 383–387, 2018, doi: 10.1109/ISCAIE.2018.8405503.

[10] B. A. Esen, "How to manage Machine Learning/Deep Learning project?," 2018. https://medium.com/@akanesen/how-to-manage-machine-learning-deep-learning-project-51f9b0fe164f (accessed May 14, 2020).