khazanah informatika

Jurnal Ilmu Komputer dan Informatika

Komparasi Kinerja Algoritma Data Mining pada *Dataset* Konsumsi Alkohol Siswa

Noviyanti Sagala*, Hendrik Tampubolon
Departemen Sistem Informasi
Universitas Kristen Krida Wacana
Jakarta, Indonesia
*noviyanti.sagala@ukrida.ac.id

Abstrak-Data mining melakukan proses ekstraksi pengetahuan yang diperoleh dari sekumpulan data dalam jumlah besar. 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 *pra-process* 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.

Kata kunci: *Data Mining;* Konsumsi Alkohol Siswa*, Naïve Bayes,* KNN*, Decision Tree.*

1. Pendahuluan

Data mining adalah proses mengekstrasi dan mengidentifikasi pengetahuan yang didapatkan dari sekumpulan data yang cukup besar. Salah satu teknik data mining adalah klasifikasi yang digunakan untuk memprediksi kelas pada suatu label tertentu. Model klasifikasi yang dibangun berdasarkan data latih dan menggunakan model tersebut untuk mengklasifikasian data uji.

Data mining telah menjadi area penelitian yang esensial dikarenakan potensinya pada institusi pendidikan[1]. Adapun beberapa penelitian sebelumnya menggunakan data mining berfokus pada prestasi akademik siswa, misalnya kelulusan atau nilai akhir siswa namun mengetahui pola perilaku siswa juga bermanfaat untuk meningkatkan evaluasi prestasi akademik siswa [2] [3]. Pola perilaku siswa mencakup perilaku di dalam dan di luar lingkungan sekolah. Salah satunya perilaku konsumsi alkohol.

Konsumsi alkohol merupakan masalah kesehatan yang umum, terutama di kalangan generasi muda. Data yang didapat dari World Health Organization (WHO) menunjukan tingkat peminum usia dini dan pola konsumsi alkohol meningkat 71% di 73 negara [4]. Kecanduan alkohol dapat dipengaruhi beberapa faktor seperti genetika, lingkungan sosial, dan kesehatan mental. Konsumsi alkohol pada usia muda (pelajar) memberikan efek jangka panjang pada otak dan kinerja akademik siswa. Secara khusus, siswa yang pernah mengkonsumsi alkohol yang berlebihan cenderung mengalami kesulitan berkaitan dengan memori otak dan kemampuan untuk fokus. Efek yang paling berbahaya adalah dapat memicu penggunaan narkoba, seperti marijuana, kokain, atau heroin. Di sisi lain, ditemukan bahwa sekolah berperan besar dalam memprediksi perilaku konsumsi alkohol di kalangan siswa [5].

Penelitian tentang konsumsi alkohol pada siswa telah dilakukan dengan beberapa algoritma data mining di antaranya penelitian yang dilakukan oleh Palaniappan [6], model klasifikasi yang dikembangkan dengan menggunakan algoritma AutoMLP mencapai akurasi lebih tinggi yakni 64,54% jika dibandingkan dengan Artificial Neural Networks (ANN) hanya 61,78%. Fabio *et al.,* menerapkan teknik klasifikasi dan *clustering* untuk prediksi konsumsi alkohol dengan membuat segmentasi data menggunakan K-Means dan teknik data mining seperti Decision Tree, SVM, Bayesian Network, dan KNN. Hasil penelitian menunjukkan SVM lebih efisien daripada model lainnya [7]. Hariharan *et al*, secara khusus menyebutkan bahwa faktor banyaknya waktu luang setelah sekolah adalah faktor yang paling mempengaruhi seorang siswa menjadi pecandu alkohol. Model prediksi dibangun menggunakan algoritma *Random Forest* [8]. Tetapi belum ada penelitian yang melakukan komparasi kinerja algoritma Naïve Bayes, K-Nearest Neighbor (KNN), Decision Tree C5.0, dan SVM sehingga belum diketahui metode yang paling akurat dalam membangun model klasifikasi pada *dataset* konsumsi alkohol pada siswa.

Kinerja algoritma data mining juga dipengaruhi oleh atribut-atribut yang digunakan pada tahap klasifikasi. Semakin banyak atribut yang digunakan boleh jadi semakin menurunkan performa metode klasifikasi jika fitur yang dipilih memiliki kekuatan diskriminasi kelas yang buruk [9]. Oleh karena itu, pemilihan atribut-atribut yang paling relevan dan informatif adalah hal yang penting dalam meningkatkan informasi tentang proses, mengurangi biaya dan penyimpanan, serta meningkatkan performa algoritma klasifikasi yang digunakan [10]. Pada penelitian ini, 2 algoritma pemilihan fitur Filter-based yaitu Gain Ratio dan Fast Correlation Based Feature (FCBF) diimplementasikan.

Dalam penelitian ini dilakukan komparasi kinerja algoritma data mining untuk mengetahui metode dengan performa prediksi konsumsi alkohol pada siswa yang terbaik dan faktor-faktor yang mempengaruhi siswa kecanduan terhadap alkohol berdasarkan model data yang ada.

2. Landasan Teori

a. *Decision Tree* C5.0

Algoritma decision tree merupakan algoritma yang paling sering digunakan untuk klasifikasi. Decision tree adalah sebuah diagram alir yang terdiri atas 3 *node* yaitu *root node, internal node, dan leaf node*. Algoritma C5.0 merupakan pengembangan dari algoritma C4.5 di mana lebih unggul dalam kecepatan, memori, dan efisiensi. Selain itu, algoritma C5.0 mampu menangani berbagai macam tipe data (kontinu, kategorikal, dan *timestamps*) dan *missing values* [11].

Decision tree dibentuk menggunakan nilai *entropy* dan *information gain*. *Entropy* menyatakan *impurity* suatu kumpulan objek dengan n kelas ditunjukkan oleh persamaan 1 dan persamaan 2 [12].

$$Info(D) = \sum_{i=0}^{n} p_i \, Log(p_i) \quad (1)$$

$$Info_A(D) = \sum_{j=0}^{v} \frac{|D_j|}{|D|} x \, Info(D) \quad (2)$$

Information gain digunakan untuk mengukur ketidakpastian dalam teori informasi ditunjukkan oleh persamaan 3.

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

b. *Naïve Bayes*

Naïve Bayes adalah metode data mining yang sederhana dan mudah untuk diimplementasikan dibandingkan metode yang lain dalam konteks klasifikasi. Metode ini juga mampu mengolah data numeric dan teks. *Theorema Bayesian* dijelaskan seperti pada persamaan 4 [13].

$$P(H|E) = \frac{P(H) \prod_{i=1}^{q} P(E_i|H)}{P(E)} \quad (4)$$

$P(H|E)$ adalah probabilitas data dengan vector E pada kelas H. $P(H)$ = probabilitas awal kelas H. $\prod_{i=1}^{q} P(E_i|H)$ = probabilitas independen kelas H dari semua fitur dalam vektor E.

Tabel 1. Deskripsi data yang digunakan

Atribut	Deskripsi
Sex	Jenis kelamin siswa (angka biner: perempuan{0} atau laki-laki{1})
Age	Usia siswa (numerik: 15-22)
School	Asal sekolah (binary: Gabriel Pereira or Mounsinho)
Address	Daerah tempat tinggal siswa (kota atau desa)
PStatus	Status orang tua (hidup bersama atau terpisah)
Medu	Pendidikan ibu (numerik: 0 ke 4.a)
Fedu	Pendidikan ayah (numerik: 0 ke 4.a)
Mjob	Pekerjaan ibu (nominal b.)
Fjob	Pekerjaan ayah (nominal b.)
Guardian	Wali (nominal: ibu, ayah atau lainnya)
Famsize	Jumlah keluarga(biner: <=3 atau > 3)
Famrel	Kualitas hubungan keluarga (numerik: 1sampai 5)
Reason	Alasan memilih sekolah (nominal: dekat dengan rumah, reputasi sekolah, preferensi kursus atau lainnya)
Traveltime	Waktu perjalanan pulang ke rumah (numerik: 1sampai 5)
Studytime	Waktu belajar mingguan (numerik: 1 sampai 5)
Failures	Jumlah kelas sebelumnya yang tidak lulus (numerik: n jika 1 <=n<3, else 4)
Schoolsup	Dukungan pendidikan ekstra dari sekolah (ya atau tidak)
Famsup	Dukungan pendidikan ekstra dari keluarga (ya atau tidak)
acitivites	Aktivitas ekstra-kurikuler (ya atau tidak)
Paidclass	Kelas ekstra berbayar (ya atau tidak)
Internet	Akses Internet di rumah (ya atau tidak)
Nursety	Nursery school (ya atau tidak)
Higher	Keinginan melanjutkan studi ke jenjang lebih tinggi (ya atau tidak)
Romantic	Memiliki hubungan romantis (ya atau tidak)
Freetime	Waktu luang setelah (numerik: 1 sampai 5)
Goout	Bergaul dan pergi bermain dengan teman-teman (numerik: 1 sampai 5)
Walc	Komsumsi alkohol di akhir pekan (numerik: 1 sampai 5)
Dalc	Komsumsi alkohol di hari kerja(numerik: 1 sampai 5)
Health	Kondisi kesehatan (numerik: 1- sangat buruk sampai 5-sangat baik)
Absences	Jumlah ketidakhadiran disekolah (numerik: 0 sampai 93)
G1	Nilai periode pertama (numerik: 0 sampai 20)
G2	Nilai periode kedua (numerik: 0 sampai 20)
G3	Nilai periode ketiga(numerik: 0 sampai 20)

Tabel 2. Konversi nilai berdasarkan sistem erasmus (*europe system*)

	Range Score	Conversion
I (Very good)	16-20	A
II (Good)	14-15	B
III (Satisfactory)	12-13	C
IV (Sufficient)	10-11	D
V (Fail)	0-9	E

c. *K-Nearest Neighbor* (KNN)

KNN bekerja secara sederhana dimana menyimpan seluruh kasus yang ada dan mengklasifikasikan kasus-kasus baru berdasarkan kemiripan (fungsi jarak). Kasus baru diklasifikasikan berdasarkan jarak data baru ke beberapa data/tetangga terdekat. Menghitung jarak menggunakan fungsi *Euclidean Distance* [14].

$$Euclidean = \sqrt{\sum_{i=1}^{k} (y_{2i} - y_{1i})^2} \quad (5)$$

y_2=data latih, y_1=data uji, i= variable data, n = dimensi data

d. *Support Vector Machine* (SVM)

SVM merupakan salah satu algoritma *machine learning* yang paling popular dan efektif digunakan pada klasifikasi dan pengenalan citra. Prinsip kerja SVM adalah mencari ruang pemisah yang paling optimal dari suatu data dalam kelas yang berbeda. Performa SVM sangat dipengaruhi oleh fungsi kernel dan parameter yang digunakan. Pada penelitian ini digunakan kernel *RBF (Radial Basis Function)* karena mempunyai performa yang baik dengan kesalahan pelatihan yang minimum [15].

3. Metode

a. *Dataset*

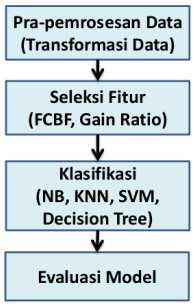
Data yang digunakan dalam penelitian ini diambil dari UCI *Machine Learning* dengan 1024 data dan 33 atribut [16].

Tabel 1.

b. Metodologi

Pada penelitian ini, model prediksi komsumsi alkohol siswa dibangun dalam beberapa tahapan, yakni pra-proses data, seleksi fitur, membangun model klasifikasi, serta evaluasi model. Keseluruhan tahapan dilakukan menggunakan R Programming dan RStudio (IDE).

Metode yang diusulkan dalam studi ini dapat dilihat pada Gambar 1.



Gambar 1. Metodologi yang diusulkan

Pra-pemrosesan data adalah tahapan yang membutuhkan waktu yang lebih lama dibandingkan tahapan yang lainnya, dikarenakan data diolah agar dapat memenuhi kebutuhan dari inputan data pada setiap algoritma *data mining* yang diusulkan.

Tahapan pertama ialah *data-cleaning* yang dilakukan untuk memeriksa dan mengisi data yang hilang serta menangani data yang tidak konsisten. Setelah tahapan *data-cleaning* selesai dan diasumsikan bahwa atribut sudah dapat diakses sebagai kolom, tranformasi data dilakukan, di antaranya atribut konsumsi alkohol di hari kerja (Dalc), konsumsi alkohol di akhir pekan (Walc), jumlah ketidakhadiran (Absence), dan nilai periode pertama sampai periode ketiga(G1-G3). Kemudian, target atribut (Alc) dirumuskan menggunakan persamaan (6) [6]. Selanjutnya, atribut "*Absence*" dimodifikasikan ke dalam nilai biner, di mana *absence* dirumuskan dengan persamaan (7) dan dengan mengasumsikan bahwa semakin tinggi jumlah ketidakhadiran maka semakin tinggi obsesi untuk meminum alkohol. Selain itu, atribut G1, G2, G3 diklasifikasikan ke dalam 5 level sesuai dengan sistem konversi nilai dari Erasmus sebagaimana dapat dilihat pada tabel 2. Selanjutnya, atribut numerik ditransformasikan menjadi kategorikal, di mana transformasi ini dilakukan secara otomatis dengan pemrograman R.

$$Alc = \frac{Dalc \times 5 + Walc \times 2}{7} \quad (6)$$

$$Absence = \begin{cases} 1, & absence \geq 10\ hari \\ 0, & selainnya \end{cases} \quad (7)$$

Seleksi fitur terdiri dari dua tahap. Pertama, seleksi fitur yang digunakan untuk pelatihan dan pengujian. Kedua, penilaian kemampuan metode pemilihan fitur menggunakan pengklasifikasian yang berbeda. Metode berbasis filter kemudian didemonstrasikan di mana metode ini tidak memerlukan eksekusi ulang (*re-execution*) untuk algoritma penambangan data yang berbeda, serta memerlukan waktu yang lebih sedikit. Metode berbasis filter ini menerapkan beberapa pemeringkatan (*ranking*) atas atribut-atribut yang ada, yang artinya seberapa penting setiap fitur tersebut untuk proses klasifikasi.

Pada penelitian ini, dua algoritma berbasis *filter* diaplikasikan, yakni FCBF menggunakan pendekatan *multivariate* dan Gain Ratio menggunakan pendekatan *univariate*. Fitur-fitur diberi peringkat sesuai dengan relevansinya terhadap target variabel. Pada algoritma Gain Ratio, batas ambang (*threshold*) didefinisikan dengan memilih fitur-fitur terbaik dengan urutan 5,10,15,20, 25,30, dan 31 [17]. Kumpulan fitur tersebut kemudian digunakan sebagai data masukan (input) pada fase klasifikasi. Fase terakhir adalah klasifikasi dengan tujuan untuk memprediksi kelas dari target secara akurat untuk setiap kasus dalam data.

Tahapan klasifikasi terdiri dari dua langkah yakni membangun model dan menggunakan model yang sudah dibangun. Pada proses membangun model, kuantitas dari proses pengujian dan pelatihan diatur menggunakan *cross-validation*. Setelah itu, Naive Bayes, Decision Tree C5.0, KNN, dan SVM diaplikasikan dan dibandingkan pada sekumpulan data yang berisi fitur-fitur terpilih. Akan tetapi, parameter-parameter terkait tidak dirinci secara spesifik sebelumnya melainkan menjalankan lansung model yang ada di *library caret* yang tersedia pada R *Programming*.

Tabel 3. Hasil algoritma FCBF

Atribut	Information gain
Sex	0.03
Goout	0.018
G1	0.014
Absences	0.0076

Tabel 4. Hasil akurasi tertinggi dari masing-masing algoritma

Model	Akurasi
Naive Bayes	1
KNN	0.883
C5.0	0.876
SVM	0.893

Tabel 5. Akurasi dari model klasifikasi dengan gain ratio

Jumlah Atribut	Model	Akurasi
5	Naive Bayes	1
15	KNN	0.9126
30	C5.0	0.9047
31	SVM	0.893

Pada model KNN yang mana memiliki keunggulan pada kepraktisannya dan sederhana, nilai *k* yang optimum ditentukan menggunakan *k-fold cross validation* kemudian menggunakan nilai tersebut untuk membangun model prediksi. Metrik jarak yang digunakan pada algoritma ini adalah *Euclidean Distance*. Model SVM dibangun menggunakan kernel *Radial Basic Function* (RBF). Nilai parameter Cost (C) dan gamma (γ) dipilih dengan kombinasi terbaik dari kedua parameter tersebut di mana nilainya ditentukan menggunakan *cross validation*. Oleh karena itu, nilai dari C dan γ perlu dirinci dengan tepat, yakni $0 \leq C \leq 10000 \leq C \leq 1000$ dan $0.01 \leq \gamma \leq 10$ $0.01 \leq \gamma \leq 10$. Semakin besar rentang nilai pada *Cost* dan Gamma, semakin lama waktu komputasi yang diperlukan, namun memberikan rentang nilai yang terlalu kecil juga berakibat pada hasil yang tidak bisa diterima [18].

Setelah model klasifikasi prediktif selesai dibangun, dilakukan evaluasi kinerja dari model tersebut. Keakuratan dari model prediksi ini kemudian diestimasi menggunakan data uji. Dalam tahap ini estimasi tingkat akurasi model dilakukan dengan menggunakan metode *10-fold cross validation*. Metode ini membagi data ke dalam sepuluh partisi (*fold*) dari total data kemudian diambil satu *fold* untuk digunakan sebagai data uji dan kumpulan dari *fold* lainnya sebagai data latih, proses ini berlangsung selama sepuluh kali dan *fold* data uji yang digunakan berbeda setiap kalinya. Lalu rata rata dari pengujian ini menjadi nilai akurasi dari model.

4. Hasil dan Pembahasan

Hasil dari tahapan *data cleaning* adalah tidak ditemukan data yang hilang, kosong, serta data yang tidak konsisten. Kemudian hasil dari tahapan lainnya akan didiskusikan pada bagian sub bab berikut, yakni: hasil dari teknik seleksi fitur dan penerapan algoritma data mining.

a. Hasil Teknik Seleksi Fitur

Hasil seleksi fitur pada algoritma FCBF dengan menggunakan kumpulan fitur yang lengkap dapat dilihat pada Tabel 3.

Tabel 3 mendeskripsikan atribut dengan skor *information gain* lebih tinggi memiliki pengaruh yang lebih kuat dalam pengklasifikasian data. Atribut jenis kelamin ternyata merupakan atribut yang sangat berpengaruh dalam menentukan apakah seorang siswa terindikasi menjadi peminum atau tidak. Hal ini diperkuat dari laporan tentang status global alkohol 2014 yang diterbitkan oleh WHO yang merepresentasikan bahwa peminum laki-laki lebih besar dari peminum perempuan. Perbedaan *information gain* dari atribut *Goout* dan *G1* ialah 0.004 yang berarti frekuensi bepergian (*going out*) dan nilai periode pertama (G1) tidak terlalu jauh berbeda. Selain itu, tingkat ketidakhadiran (*absence*) siswa yang semakin tinggi akan berakibat semakin besar pula peluang untuk siswa menjadi pecandu alkohol.

Selanjutnya, Gain Ratio di implementasikan untuk memberi peringkat pada fitur sesuai dengan relevansinya terhadap label. Hasil dari Gain Ratio diurutkan sesuai dengan tingkatan pengaruh terbesar dari atribut terhadap label, yakni; *sex, goout, studytime, absences, freetime, G1, G2, G3, higher, reason, school, address, Fjob, guardian, famsize, schoolsup, nuersery, famsup, Mjob, Pstatus, paid, romantic, internet, activities, age, Medu, Fedu, traveltime, failures, famrel, health*. Dari hasil tersebut kemudian digunakan 5, 10, 15, 20, 25, 30 atribut tertinggi dalam proses klasifikasi.

Hasil Gain Ratio dengan menggunakan dua atribut pertama yang paling penting yakni *sex, goout,* memberikan hasil yang sama pada FCBF. Hal ini menunjukkan bahwa jenis kelamin dan frekuensi bepergian memiliki pengaruh yang lebih besar untuk menentukan siswa tersebut peminum atau tidak daripada atribut yang lainnya. Namun berbeda dengan hasil Gain Ratio dan FCBF pada atribut G1 di mana Gain Ratio memiliki pengaruh yang lebih besar pada siswa dibandingkan atribut *Studytime*.

Kemudian seluruh algoritma klasifikasi yang diimplentasikan pada sekumpulan fitur terpilih hasil dari algoritma pemilihan fitur dan kinerja algoritma dihitung pada data uji, hasilnya kemudian digunakan untuk menganalisis kemampuan dari masing-masing model.

b. Hasil Algoritma Klasifikasi

Hasil algoritma klasifikasi pada seluruh fitur (31 fitur dapat dilihat pada tabel 4. Model NB mengungguli model lainnya dengan rata-rata akurasi 100% diikuti model SVM, sedangkan model yang dibangun menggunakan algoritma Decision Tree C5.0 menghasilkan akurasi terendah.

Hasil akurasi keempat *classifier* yang diimplementasikan pada fitur terpilih hasil algoritma FCBF menunjukkan algoritma C5.0, NB dan SVM mencapai akurasi tertinggi pada 89.23%, tetapi algoritma KNN mencapai akurasi 91.30%. KNN berhasil mengungguli model lainnya dengan 4 fitur terbaik hasil FCBF pada nilai k=7.

Performa algoritma C5.0, SVM, KNN, dan NB pada hasil fitur terpilih Gain Ratio sangat bervariasi. Akurasi tertinggi dari setiap model dapat dilihat pada tabel 5 pada setiap jumlah atribut yang digunakan. Algoritma C5.0 menghasilkan akurasi tertinggi (90.47%) ketika menggunakan top-30 fitur. Algoritma SVM mempertahankan hasil akurasi yang sama dengan hasil saat menggunakan fitur-fitur hasil FCBF yakni 89.32% dengan nilai parameter $C = 1 C = 1$ dan $\gamma = 1 \gamma = 1$. SVM juga memberikan hasil yang konsisten ketika menggunakan fitur top-5. Pada model KNN, akurasi meningkat setelah menggunakan fitur top-15. Di sisi lain, model NB berhasil mencapai akurasi 100% pada saat menggunakan *top-5* fitur, yang berarti model NB mengungguli model lainnya hanya dengan menggunakan 5 fitur.

Beberapa penelitian terdahulu menggunakan dataset yang sama menunjukkan hasil yang berbeda dengan penelitian ini dikarenakan penelitian ini menggunakan pra-pemrosesan data dan algoritma seleksi fitur yang berbeda.

Eksperimen oleh Fabio Mat dan M. Amran [19] menggunakan Decision Tree di KNIME Tool memberikan akurasi 92%. Selain itu, atribut siswa laki-laki dan frekuensi bepergian merupakan 2 atribut yang paling berpengaruh dalam memprediksi siswa yang kecanduan terhadap alkohol. Hasil ini serupa dengan hasil yang didapatkan dari penelitian yang dilakukan penulis. Improvisasi dilakukan di mana metode yang diajukan menggunakan kombinasi antara algoritma K-Means Clustering dan Decision Tree, Bayesian Network, KNN, serta SVM. Algoritma SVM mengungguli performa algoritma yang lain dengan *precision* dan *recall* yang sama yaitu 98%. Pada penelitian ini tidak diimplementasikan metode pemilihan atribut. Kinerja algoritma AutoMLP dan standar MLP dibandingkan dan model klasifikasi yang dibangun menggunakan AutoMLP mencapai akurasi yang lebih baik dengan 64.54%. Eksperimen dilakukan pada RapidMiner Tool dan tidak menggunakan metode pemilihan fitur [6].

Performa Naïve Bayes juga mengungguli algoritma KNN, J48, ANN, dan ZeroR pada *dataset* diabetes [20]. Algoritma Naïve Bayes juga mencapai nilai akurasi, presisi, dan *recall* terbaik dibandingkan Decision Tree dan k-Nearest neighbor untuk mencari *design alternative* pada alat simulasi energi [21]. Penelitian lainnya pada referensi [22] – [24] juga menunjukkan hasil yang sama ketika performa Naïve Bayes dan Decision Tree dibandingkan.

Naïve Bayes menghasilkan performa yang baik dikarenakan kemampuan algoritma tidak hanya menangani atribut-atribut yang saling memiliki keterkaitan atau tidak memiliki keterkaitan apapun saat mengolah data dari domain yang bervariasi. Saat diimplementasikan pada kasus klasifikasi, kombinasi dari keseluruhan relasi atribut-atribut dalam kelas tertentu akan mempengaruhi bahkan menghilangkan relasi antar dua atribut dan tidak akan mempengaruhi klasifikasi. Distribusi keseluruhan relasi atribut-atribut pada kelas tertentu yang mempengaruhi klasifikasi Naïve Bayes tidak hanya relasi atau ketergantungan antar dua atribut yang berbeda [25]-[26].

Dalam penelitian ini, sama halnya dengan penelitian sebelumnya, ditemukan bahwa model klasifikasi yang dibangun menggunakan algoritma Naïve Bayes (NB) adalah model yang terbaik dibandingkan dengan model lainnya walaupun menggunakan teknik seleksi fitur yang berbeda.

5. Kesimpulan

Pada penelitian ini, beberapa teknik klasifikasi diterapkan pada data konsumsi alkohol siswa. Variasi algoritma klasifikasi yang diterapkan adalah Naïve Bayes, Decision Tree C5.0, KNN, dan SVM. Penentuan fitur-fitur relevan dan penting dicapai dengan algoritma Gain Ratio dan FCBF. Pemilihan fitur ini bertujuan untuk mengetahui faktor-faktor yang paling berpengaruh terhadap konsumsi alkohol pada siswa. Model klasifikasi yang paling efisien dikelola dengan membandingkan rata-rata akurasi menggunakan *10-fold cross validation*. Secara umum, kinerja *classifier* menggunakan pemilihan fitur lebih baik daripada menggunakan seluruh fitur. Namun, Naive Bayes dengan 5 fitur terbaik hasil Gain Ratio menghasilkan kinerja terbaik dengan akurasi 100%. Sementara, Decision Tree C5.0 dengan fitur lengkap menghasilkan akurasi model prediksi terendah di antara model lainnya. Namun, untuk memberikan lebih banyak wawasan atau gambaran terhadap kinerja algoritma klasifikasi yang digunakan, pengujian lebih lanjut pada *dataset* yang sama maupun menggunakan *dataset* yang berbeda perlu dilakukan.

6. Daftar Pustaka

[1] R. Sumitha, E. S. Vinothkumar, and P. Scholar, "Prediction of Students Outcome Using Data Mining Techniques," *Int. J. Sci. Eng. Appl. Sci.*, vol. 2, no. 6, pp. 132–139, 2016.

[2] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015.

[3] R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, "Data Mining Models for Student Careers," *Expert Sys. App.*, vol. 42, no.13, pp. 5508–5521, 2015.

[4] W. H. Organisation, "Global status report on alcohol and health," *World Heal. Organ.*, pp. 1–100, 2014.

[5] S. Kairouz and E. M. Adlaf, "Schools, Students and Heavy Drinking: a Multilevel Analysis," *Addict. Res. Theory*, vol. 11, no. 6, pp. 427–439, 2003.

[6] S. Palaniappan, N. A. Hameed, A. Mustapha, and N. A. Samsudin, "Classification of Alcohol Consumption among Secondary School Students," vol. 1, no. 4, pp. 224–226, 2017.

[7] M.-P. Fabio, D. la Hoz-Manotas Alexis, M.-O. Roberto, M.-P. Ubaldo, D.-M. Jorge, and C.-N. Harold, "Designing A Method for Alcohol Consumption Prediction Based on Clustering and Support Vector Machines," *Res. J. Appl. Sci. Eng. Technol.*, vol. 14, no. 4, pp. 146–154, 2017.

[8] B. Hariharan, R. Krithivasan, and A. Deborah, "Prediction of Secondary School Students' Alcohol Addiction using Random Forest," *Int. J. Comput. Appl.*, vol. 149, no. 6, pp. 975–8887, 2016.

[9] Syaiful and Harianto, "Pemilihan Fitur untuk Monitoring dan Klasifikasi Kondisi Pahat," vol. 37, no. 1, pp. 32–40, 2016.

[10] M. M. Abdul Jalil, F. Mohd, and N. M. Mohamad Noor, "A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection," *Procedia Comput. Sci.*, vol. 116, pp. 113–120, 2017.

[11] R. Revathy and R. Lawrance, "Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 1, pp. 50–58, 2017.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Technique*. 2011.

[13] O. Ardhapure, G. Patil, D. Udani, and K. Jetha, "Comparative Study of Classification Algorithm for Text Based Categorization," *Int. J. Res. Eng. Technol.*, vol. 5, no. 2, pp. 217–220, 2016.

[14] Y. Kustiyahningsih, D. R. Anamisa, and N. Syafa'ah, "Sistem Pendukung Keputusan untuk Menentukan Jurusan pada Siswa SMA Menggunakan Metode KNN dan SMART," Skripsi, Universitas Trunojoyo, Madura, 2013.

[15] A. M. Puspitasari, D. E. Ratnawati, and A. W. Widodo, "Klasifikasi Penyakit Gigi Dan Mulut Menggunakan Metode Support Vector Machine," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 2, pp. 802–810, 2018.

[16] D. Dheeru and E. K. Taniskidou, "UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2017.

[17] N. Sagala and J. Wang, "A Comparative Study for Classification on Different Domain," *10th Intl. Conf. on Mach. Learn. and Comp.*, pp. 1–5, 2018.

[18] S. W. Lin, K. C. Ying, S. C. Chen, and Z. J. Lee, "Particle Swarm Optimization for Parameter Determination and Feature Selection of Support Vector Machines," *Exp. Syst. Appl.*, vol. 35, no. 4, pp. 1817–1824, 2008.

[19] F. Pagnotta and M. A. Hossain, "Using Data Mining to Predict Secondary School Student Alcohol Consumption," *Dep. Comput. Sci. Univ. Camerino.*, pp. 1–9, 2016.

[20] A. S. Rani and S. Jyothi, "Performance Analysis of Classification Algorithms under Different Datasets," 3rd Intl. Conf. on Comp. for Sustainable Global Dev. (INDIACom), pp. 1584-1589, 2016.

[21] A. Ashari, I. Paryudi, and A. M. Tjoa, "Performance Comparison between Naïve bayes, Decision Tree and k-Nearest neighbor in Searching Alternative Design in an Energy Simulation Tool," *Intl. J. of Adv. Comp. Science and App.*, vol 4, pp 33-39, 2013.

[22] R. M. Rahman and F. Afroz, "Comparison of Various Classification Techniques using Different Data Mining Tools for Diabetes Diagnosis," *J. Softw. Eng. Appl.*, vol. 6, no. 1, pp. 85–97, 2013.

[23] L. Dan, L. Lihua, Z. Zhaoxin, "Research of Text Categorization on WEKA," *3rd Intl. Conf. on Intelligent Sys. Design and Engi. App.*, 2013.

[24] J. Huang, J. Lu, C. X. Ling, "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy," *3rd IEEE Int. Conf. on Data Mining*, 2003.

[25] E. Frank, L. Trigg, G. Holmes, and I. H. Witten, "Technical note: Naive Bayes for Regression," *Mach. Learn.*, vol. 41, no. 1, pp. 5–25, 2000.

[26] H. Zhang, "The Optimality of Naive Bayes," *Florida Artif. Intell. Res. Soc. Conf.*, no. 2, pp. 1–6, 2004.