Media Aplikom
Volume 13 Nomer 2| Desember, 2021| https://doi.org/10.33488/1.ma.2.1.305

Pengujian Model Prediksi Menggunakan Metode Data Mining Classification Decision Tree Untuk Penentuan Peminatan Peserta Didik

Siska Narulita[1], Andreas Tigor Oktaga[2], Ika Susanti[3]
Sistem dan Teknologi Informasi, Institut Teknologi dan Bisnis Semarang
Email: siska.itbs@gmail.com

Abstrak

Ketidaktahuan akan kemampuan peserta didik dapat membawa dampak pada perkembangan potensi karirnya. Penelitian terkait analisis kemampuan siswa telah banyak dilakukan, berupa nilai dan profil. *Data mining* merupakan proses analisis kumpulan untuk menemukan hubungan tak terduga serta meringkas data tersebut menggunakan cara baru yang bisa dipahami dan memberi manfaat bagi pemiliknya (Larose, 2006). Penelitian ini dilakukan untuk menguji akurasi model prediksi menggunakan metode *data mining classification decision tree (Algoritma C4.5).* Penelitian sebelumnya dilakukan oleh Setiawati (2016) berjudul *Model Hybrid Metode SAW* dan *TOPSIS* untuk Menentukan Peminatan Peserta Didik SMA. Dengan menggunakan metode ini, diperoleh hasil bahwa tingkat akurasi dari model prediksi sebesar 86,84% dan nilai AUC sebesar 0,752 termasuk kategori *fair classification* (Gorunescu, 2011). Sedangkan penelitian sebelumnya dengan model *hybrid metode SAW* dan *TOPSIS* menghasilkan nilai akurasi sebesar 80,3%. Sehingga, model prediksi *data mining* metode *classification decision tree (Algoritma C4.5)* menghasilkan nilai akurasi lebih baik, karena itu dapat dibuat acuan dalam prediksi penentuan peminatan peserta didik.

Kata kunci: Kemampuan; Potensi Karir; Akurasi; Peserta Didik; Algoritma C4.5

Testing the Prediction Model Using the Data Mining Classification Decision Tree Method for Determining Student Interests

Abstract

Ignorance of students' abilities can have an impact on the development of their career potential. Many studies related to the analysis of student abilities have been carried out, in the form of grades and profiles. Data mining is a collection analysis process to find unexpected relationships and summarize the data using new ways that can be understood and benefit the owner (Larose, 2006). This research was conducted to test the accuracy of the prediction model using the data mining classification decision tree method (algorithm C4.5). A previous study conducted by Setiawati (2016) entitled Hybrid Model of SAW and TOPSIS Methods to Determine the Interests of High School Students. By using this method, the results obtained that the accuracy of the prediction model is 86.84% and the AUC value of 0.752 is included in the fair classification category (Gorunescu, 2011). Meanwhile, previous research using a hybrid model of the SAW and TOPSIS methods resulted in an accuracy value of 80.3%. Thus, the data mining prediction model of the classification decision tree method (algorithm C4.5) produces

better accuracy values, because it can be used as a reference in predicting the determination of students' specialization.

Keywords: Ability; Career Potential; Accuracy; Learners; Algorithm C4.5

Pendahuluan

Mengetahui minat dan bakat peserta didik sangat penting. Dengan mengetahui bakat dan peminatan dari peserta didik, para pendidik dapat mengetahui kemampuan, bakat dan peminatan peserta didiknya. Pendidik dapat mengarahkan peserta didiknya sesuai kemampuan, bakat dan peminatan yang dimiliki. Peserta didik sendiri dapat mengembangkan kemampuan atau bakat yang dimiliki. Ketidaktahuan akan kemampuan, bakat atau peminatan dari peserta didik akan membawa dampak pada perkembangan potensi bahkan karirnya untuk masa depan. Penelitian tentang analisis minat dan bakat siswa juga dilakukan oleh (Hoffler, Kohler, & Parchmann, 2019). Dalam penelitiannya yang berjudul *Scientist of the Future*: *An Analysis of Talented Students Interests*, mengemukakan bahwa pentingnya mengetahui ketrampilan, sifat dan minat siswa ketika mencari dan mendidik ilmuwan untuk masa depan, dan itu sudah harus dimulai di tingkat sekolah (Hoffler, Kohler, & Parchmann, 2019). Penelitian lainnya tentang peminatan peserta didik dilakukan oleh (Ito & McPherson, 2018). Penelitiannya yang berjudul *Factors Influencing High School Students Interest in* pSTEM tersebut bertujuan untuk mengidentifikasi faktor-faktor yang mempengaruhi minat siswa wanita dalam mengejar karir di bidang ilmu fisika, teknologi, teknik dan matematika, karena menurutnya karis tersebut banyak didominasi siswa laki-laki (Ito & McPherson, 2018). Di dalam pedoman tentang peminatan yang diterbitkan oleh Kemendikbud, penentuan peminatan oleh siswa atau peserta didik yaitu suatu proses penentuan pilihan bakat dan minat terkait keahlian yang dimiliki berdasarkan peluang di lapangan dan potensi diri (Pendidikan, 2013). Dari data peserta didik berupa profil dan nilai, dapat digali lebih dalam sehingga akan diperoleh informasi yang berharga tentang peserta didik.

Dengan melakukan penggalian terhadap sekumpulan data tersebut dapat diperoleh pengetahuan baru yang tersembunyi dan tentunya mempunyai manfaat. (Larose, 2006), dalam bukunya yang berjudul *Data Mining Methods and Models*, mengemukakan definisi *data mining* adalah suatu proses analisis dari kumpulan data yang bertujuan untuk menemukan hubungan tak terduga serta meringkas data tersebut menggunakan cara baru yang bisa dipahami dan membawa manfaat bagi pemiliknya. (Larose, 2006). memprediksi bahwa *data mining* akan menjadi salah satu perkembangan yang paling maju pada dekade selanjutnya, selain itu MIT *Technology Review* menetapkan *data mining* menjadi salah satu dari sepuluh teknologi muktakhir yang akan mengubah dunia (Larose, 2006). Penelitian tentang *data mining* sampai dengan saat ini masih terus dilakukan, misalnya penelitian yang dilakukan oleh Głębockaa & Zdrodowskab, 2021) yang berjudul *Analysis Children with Disabilities Self Care Problems based on Selected Data Mining Techniques*. Dalam penelitiannya tersebut digunakan teknik *data mining* untuk klasifikasi permasalahan perawatan diri pada anak penyandang disabilitas, sehingga dapat membantu dan mendukung pekerjaan diagnostik dan terapis. Penelitian lainnya tentang *data mining* dilakukan oleh (Wood, 2021) dalam penelitiannya yang diberi judul *Prediction and Data Mining of Burned Areas of Forest Fires*: *Optimized Data Maching and Mining Algorithm Provides Valuable*

Insight, menggunakan *data mining* untuk prediksi dan pendataan area kebakaran hutan, sehingga dapat mengurangi terjadinya dan menyebarnya kebakaran hutan. Penggunaan *data mining* untuk prediksi hasil tanaman dilakukan oleh (Kamath, P. Patil, & S, 2021) dalam penelitiannya yang berjudul *Crop Yield Forecasting using Data Mining*. Dari penelitian yang dilakukannya tersebut, para petani dapat menentukan tanaman terbaik untuk pertanian mereka berdasarkan jenis tanah, pH dan pupuk.

Dalam sektor pendidikan, penelitian yang menggunakan *data mining* juga telah banyak dilakukan. (Viloria, Data Mining Techniques and Multivariate Analysis to Discover Patterns in University Final Researches, 2019) dalam penelitiannya *Data Mining Techniques and Multivariate Analysis to Discover Pattern in University Final Researches*, melakukan penelitian menggunakan metode *data mining* klasifikasi untuk menemukan pola dalam penelitian akhir mahasiswa di Fakultas Sains Universitas Mumbai. (Viloria, Determinating Student Interactions in a Virtual Learning Environment using Data Mining, 2019) juga melakukan penelitian untuk menentukan interaksi siswa dalam lingkungan pembelajaran virtual menggunakan *data mining*. Pada penelitian yang diberi judul *Determinating Student Interactions in a Virtual Learning Environment using Data Mining*, (Viloria, Determinating Student Interactions in a Virtual Learning Environment using Data Mining, 2019) juga menggunakan metode klasifikasi. Penelitian lainnya dalam sektor pendidikan dengan menggunakan teknik atau metode *data mining* yaitu penelitian oleh (Arcinas, Sajja, Asif, & dkk, 2021), *Role of Data Mining in Education for Improving Students Performance for Social Change*. Dalam penelitiannya menggunakan berbagai metode atau teknik *data mining* dalam prediksi karakteristik yang mempengaruhi pilihan mahasiswa terhadap suatu bidang studi di perguruan tinggi. (Matzavela & Alepis, 2021) melakukan penelitian yang diberi judul *Decision Tree Learning Through a Predictive Model for Student Academic Performance in Intelligent m-Learning Environment* untuk memprediksi model kinerja akademik siswa dalam mempelajari sistem cerdas m-*Learning*. (Matzavela & Alepis, 2021) memakai metode atau teknik *data mining* klasifikasi *decision tree* dalam penelitiannya tersebut. (Irawan, 2019) melakukan penelitian untuk penentuan jurusan metode *data mining* k-*Means*. Penelitiannya tersebut diberi judul *Implementation of Data Mining for Determining Majors using* k-*Means Algorithm in Students of* SMA Negeri 1 Pangkalan Kerinci.

Pada penelitian yang dilakukan ini juga menggunakan metode *data mining* klasifikasi *decision tree* untuk penentuan peminatan peserta didik. *Dataset* yang digunakan adalah data siswa Sekolah Menengah Atas Negeri (SMAN) 8 Semarang. *Dataset* ini sebelumnya juga telah digunakan pada penelitian yang dilakukan oleh (Setiawati, 2016) dalam penelitiannya yang berjudul Model Hybrid Metode SAW dan TOPSIS untuk Menentukan Peminatan Peserta Didik SMA. Kedua metode tersebut termasuk dalam metode *Multiple Criteria Decision Making* (MCDM). Hasil penelitiannya menunjukkan bahwa dalam menentukan minat dari peserta didik memakai kedua metode MCDM tersebut menghasilkan tingkat akurasi sebesar 80,3%. Hasil penelitian yang dilakukan dengan menggunakan metode *data mining* diharapkan menghasilkan nilai akurasi lebih baik dibandingkan dengan menggunakan metode MCDM yang diterapkan pada penelitian sebelumnya, sehingga dapat membantu pihak sekolah dalam proses penentuan peminatan peserta didik.

Metode Penelitian

Tinjauan Pustaka

1) Peminatan Peserta Didik

Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional pada Bab V tentang Peserta Didik Pasal 12 ayat (1) poin (b) menyebutkan bahwa tiap peserta didik pada tiap satuan pendidikan berhak memperoleh pelayanan pendidikan sesuai dengan minat, bakat dan kemampuan yang dimiliki (Pemerintah, 2003). Berdasarkan pasal tersebut, tiap peserta didik mempunyai hak untuk mendapatkan pelayanan pendidikan yang sesuai dengan kemampuan, bakat dan minat yang dimiliki dan pihak sekolah wajib mendukung dan mengembangkan peminatan dari peserta didik. Menurut Pedoman Peminatan Peserta Didik yang dikeluarkan oleh Badan Pengembangan Sumber Daya Manusia Pendidikan dan Kebudayaan dan Penjaminan Mutu Pendidikan, Kementerian Pendidikan dan Kebudayaan, peminatan peserta didik adalah suatu proses penentuan pilihan oleh peserta didik terkait bidang keahlian yang dimiliki berdasarkan potensi yang dimiliki dan prospek yang tersedia (Pemerintah, 2003).

Pada arahan Pasal 1 ayat (1) Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional, layanan peminatan dari peserta didik merupakan salah satu upaya pengarahan dan pemberian fasilitas untuk perkembangan peserta didik agar secara aktif dapat mengembangkan potensinya (Pemerintah, 2003). Pemberian layanan penentuan minat dari peserta didik itu merupakan tanggung jawab kepala sekolah yang melibatkan semua komponen yang ada di sekolah. Adapun tujuan peminatan peserta didik secara umum adalah membantu tiap peserta didik di segala jenjang pendidikan untuk menanamkan, memantapkan, memilih serta menetapkan minat pada mata pelajaran dan pendalamannya yang diikuti, pilihan karir atau pilihan studi lanjut sampai ke jenjang perguruan tinggi (Pemerintah, 2003).

2) *Data Preprocessing*

Basis data saat ini sangat rentan terhadap *noisy*, data yang hilang (*missing*) dan data yang tidak konsisten (*inconsistent*) karena ukurannya yang sangat besar dan berasal dari berbagai sumber yang heterogen (J. Han & Pei, 2012). Oleh karena itu, sebelum *dataset* diolah menggunakan *data mining*, perlu dilakukan *preprocessing data* untuk membantu meningkatkan kualitas hasil *data mining* serta meningkatkan efisiensi dan kemudahan metode *data mining*. Terdapat beberapa teknik dalam *preprocessing data*, yaitu (J. Han & Pei, 2012) :

a. *Data* Cleaning

Data cleaning diterapkan untuk menghilangkan *noise*, mengisi nilai yang hilang dan memperbaiki inkonsistensi data.

b. *Data Integration*

Data integration menggabungkan berbagai data dari sumber berbeda ke dalam *penyimpanan* yang koheren seperti *data warehouse*.

c. *Data Reduction*

Data reduction dilakukan untuk mengurangi ukuran data misalnya dengan cara *menggabungkan* data, menghilangkan fitur yang berlebihan atau dengan mengelompokkan data.

d. *Data Transformation*

Data transformation diterapkan agar skala data berada dalam rentang yang lebih kecil, hal ini dapat meningkatkan akurasi dan efisiensi algoritma *data mining* yang melibatkan pengukuran jarak, teknik transformasi data misalnya normalisasi. Pada penelitian ini digunakan metode *min–max normalization*, yaitu metode normalisasi dengan melakukan transformasi linear terhadap data asli dan nilai yang dinormalisasi berada pada kisaran atau *range* tertentu (Saranya, dkk, 2013). Adapun kelebihan dari metode normalisasi *min-max* adalah adanya keseimbangan nilai perbandingan antar data disaat sebelum maupun sesudah dilakukannya proses normalisasi. Metode normalisasi *min-max* banyak digunakan para peneliti untuk melakukan normalisasi karena metode ini tergolong mudah dan hasil yang diperoleh tidak bias, sehingga dapat mempermudah perhitungan normalisasi data dan lebih efisien (Martiana, 2003). Rumus perhitungan metode normalisasi *min-max* sebagai berikut:

1. Untuk data yang digolongkan ke dalam *low better*:

$$x_i^*(k) = \frac{x_i(k) - max\ x_i(k)}{min\ x_i(k) - max\ x_i(k)}$$

2. Untuk data yang digolongkan ke dalam *high better*:

$$x_i^*(k) = \frac{x_i(k) - min\ x_i(k)}{max\ x_i(k) - min\ x_i(k)}$$

Keterangan:

$x_i^*(k)$	:	Nilai data yang telah dinormalisasi
$x_i(k)$	:	Nilai data yang akan dinormalisasi
$min\ x_i(k)$	:	Nilai data yang akan dinormalisasi dan mempunyai nilai yang paling kecil
$max\ x_i(k)$	:	Nilai data yang akan dinormalisasi dan mempunyai nilai yang paling besar

3) *Data Mining*

Menurut (Larose, 2006), *data mining* merupakan proses analisa suatu kumpulan data untuk menemukan pengetahuan yang tidak diduga sebelumnya dan meringkas data dengan menggunakan cara yang berbeda dari sebelumnya, sehingga dapat dimengerti dan memberikan manfaat bagi pemilik data. Kata *mining* itu sendiri berasal dari kata *mine* dari bahasa Inggris yang artinya menambang sumber daya yang tersembunyi (Aprilla, Baskoro, & dkk, 2013). Metode atau teknik *data mining* dapat dibagi sesuai dengan tugasnya, yaitu (Aprilla, Baskoro, & dkk, 2013) :

a. *Classification*

Klasifikasi (*classification*) merupakan teknik *data mining* dengan cara mengekstrak model dengan menggambarkan kelas data yang penting (J. Han & Pei, 2012). Model tersebut disebut klasifikasi dengan melakukan prediksi label kelas berdasarkan kategori (J. Han & Pei, 2012). Sebagai contoh, membuat model

klasifikasi untuk mengkategorikan aplikasi pinjaman bank dengan label kelas aman atau beresiko. Terdapat banyak metode klasifikasi yang diusulkan oleh para peneliti dalam *machine learning*, *pattern recognition* (pengenalan pola) dan statistik. Pengaplikasian metode *data mining* klasifikasi sudah banyak dilakukan, seperti deteksi penipuan (*fraud detection*), penentuan target pemasaran, prediksi kinerja (*performance prediction*), manufaktur dan diagnosa medis (*medical diagnosis*).

b. *Association*

Metode *data mining* asosiasi (*association*) merupakan metode atau teknik yang digunakan untuk mengenali pola atau perilaku dari kejadian-kejadian khusus yang mana hubungan asosiasi terdapat pada setiap kejadian tersebut (Aprilla, *et al*, 2013). Contoh aplikasi metode asosiasi ialah *Market Basket Analysis* (MBA), yaitu memprediksi barang belanjaan yang dibeli pelanggan secara bersamaan.

c. *Clustering*

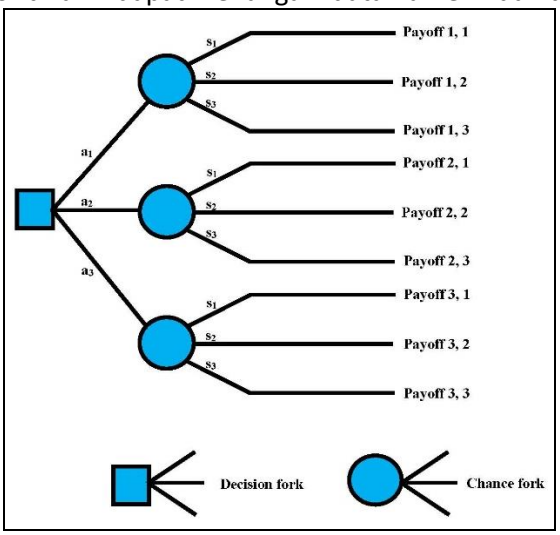
Menurut (Aprilla, Baskoro, & dkk, 2013), metode *data mining* klastering (*clustering*) merupakan metode atau teknik yang digunakan untuk analisa pengelompokan data yang belum didefinisikan (tidak ada label kelas).

4) *Decision Tree*

Decision Tree atau pohon keputusan merupakan model dari rangkaian keputusan atau ketetapan yang menuju pada solusi atau pemecahan masalah yang dihasilkan. Menurut (J. Han & Pei, 2012), *decision tree* adalah *flowchart* yang berbentuk seperti pohon (*tree*) yang mana setiap simpul yang ada merepresentasikan uji atribut, tiap cabang menggambarkan hasil uji dan simpul pada daun merepresentasikan kelas. Jadi, alur atau urutan pada *decision tree* yang memegang prediksi dimulai pada simpul akar menuju simpul daun (J. Han & Pei, 2012).

5) Algoritma C4.5

Salah satu algoritma yang digunakan untuk membuat *decision tree* adalah algoritma C4.5. Algoritma ini dapat menangani data numerik dan diskrit.



Gambar 1. *Decision Tree* Secara Umum

Menurut (Gorunescu, 2011), tahapan dari algoritma C4.5 adalah:

a. Persiapkan *dataset training*.
b. Menghitung nilai *gain* yang paling tinggi dari setiap atribut atau berdasarkan pada perhitungan *index entropy* dengan menggunakan rumus perhitungan sebagai berikut:

$$Entropy(i) = I_E(i) = -\sum_{j=1}^{m} f(i,j).log_2[f(i,j)]$$

Keterangan:

i	:	himpunan kasus
m	:	jumlah partisi i
$f(i,j)$	:	proporsi j terhadap i

c. Menghitung nilai *gain*. Adapun rumus perhitungannya sebagai berikut:

$$Entropy_{split} = \sum_{i=1}^{p} \frac{n_i}{n} I_E(i)$$

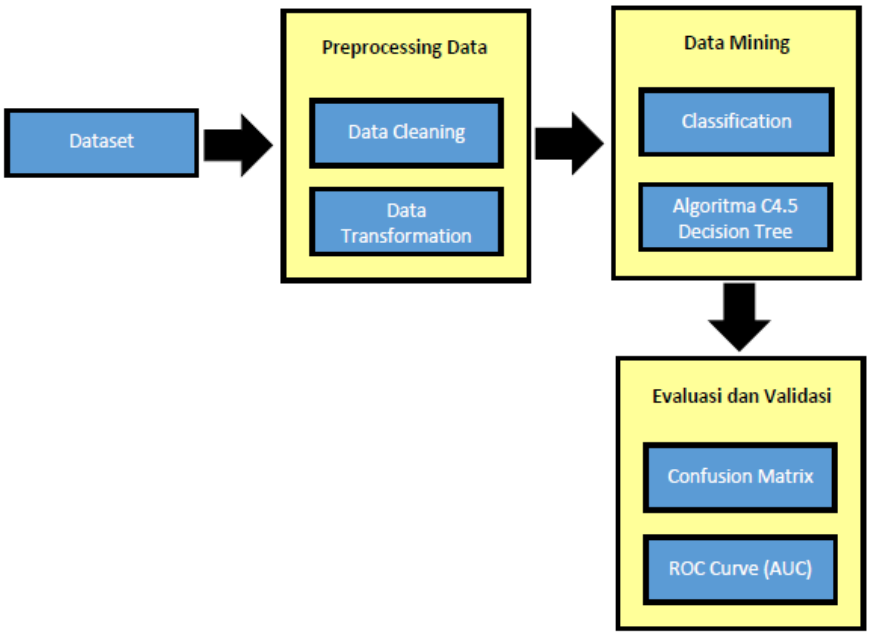
Keterangan:

p	:	jumlah partisi atribut
n_i	:	proporsi n_i terhadap i
n	:	jumlah kasus dalam n

d. Mengulang langkah nomor 2 sampai semua *record* terpartisi. Partisi *decision tree* akan berhenti ketika:
   1. Keseluruhan *tupel* dalam *record* simpul m memperolah kelas sama.
   2. Atribut dalam *record* tidak ada yang dipartisi lagi.
   3. Pada cabang yang kosong, tidak ada *record*.

Metode

Metode penelitian yang dipakai di penelitian ini ditunjukkan pada gambar 2.



Gambar 2. Metode Penelitian

1) *Dataset*

Dataset adalah suatu himpunan data yang diperoleh dari informasi-informasi pada masa sebelumnya yang dapat diproses menggunakan metode atau teknik *data mining* untuk menghasilkan pengetahuan atau informasi baru. *Dataset* yang digunakan pada penelitian ini adalah dataset peserta didik SMAN 8 Semarang yang juga digunakan dalam penelitiannya (Setiawati, 2016) menggunakan metode *Multiple Criteria Decision Making*. Di dalam *dataset* terdapat tiga belas atribut dimana satu atribut merupakan label (realitas) dan 254 *record* data.

2) *Preprocessing Data*

Preprocessing data adalah proses pengolahan data yang dilakukan sebelum dilakukan proses atau metode *data mining* dilakukan, untuk membantu peningkatan kualitas hasil *data mining* serta peningkatan efisiensi dan kemudahan dalam teknik atau metode *data mining* (J. Han & Pei, 2012). Dalam penelitian ini, teknik *preprocessing* data yang dilakukan adalah:

a. *Data Cleaning*

Pada *dataset* peserta didik SMAN 8 Semarang terdapat beberapa data yang masih kosong atau tidak lengkap dan data yang tidak konsisten (inconsistent), oleh karena itu perlu dilakukan pembersihan data dengan teknik *preprocessing data cleaning*.

b. *Data Transformation*

Di dalam *dataset* peserta didik SMAN 8 Semarang terdapat data dengan tipe berbeda, sehingga diperlukan transformasi data pada *dataset* tersebut agar bisa diolah dengan *data mining* dengan menggunakan *tool* RapidMiner. Karena terdapat data pada *atribut dataset* dengan rentang nilai yang berbeda, maka pada penelitian ini digunakan teknik normalisasi *min-max*. Adapun atribut pada *dataset* peserta didik SMAN 8 Semarang ditunjukkan tabel berikut:

Tabel 1. Atribut dan Tipe Data

Atribut	Tipe Data	Keterangan
No_PPD	Integer	
Raport_MTK	Integer	
Raport_IPA	Integer	
Raport_IPS	Integer	
UN_MTK	Integer	
UN_IPA	Integer	
Angket_Siswa	Binomial	
Angket_Ortu	Binomial	
IQ	Integer	
Score_IPA	Integer	
Score_IPS	Integer	
Saran	Binomial	
Realita	Binomial	Label/*Class*

3) Data Mining

Setelah *preprocessing data* selesai dilakukan, langkah selanjutnya adalah proses *data mining*. Pada penelitian ini digunakan metode *data mining classification* untuk penentuan peminatan peserta didik. Teknik atau metode *classification* yang dipakai ialah *decision tree* dengan algoritma C4.5.

4) Evaluasi dan Validasi

Model yang terbentuk dari proses *data mining* akan dilakukan evaluasi dan validasi. Untuk mengetahui tingkat akurasinya digunakan *Confusion Matrix*. Semakin tinggi nilai akurasi, maka akan semakin baik model yang dihasilkan dari proses *data mining* tersebut. Selain itu, teknik evaluasi dan validasi juga menggunakan aturan ROC *Curve* yang disebut *Area Under the* ROC *Curve* (AUC) untuk menilai atau mengukur performansi model *data mining*. Performansi algoritma dikatakan baik jika kurva mendekati titik 0,1 dan dikatakan buruk jika kurva yang dihasilkan mendekati garis melintang dari titik 0,0 atau garis *baseline*. Nilai AUC yang semakin mendekati 1 berarti semakin baik model prediksinya.

Hasil dan Pembahasan

Dataset

Sampel *dataset* peserta didik SMAN 8 Semarang ditunjukkan pada gambar berikut:

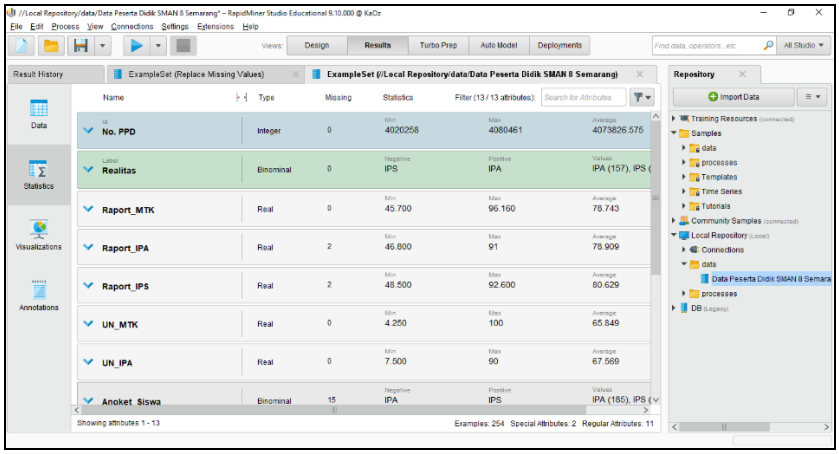
No_PPD	Raport_MTK	Raport_IPA	Raport_IPS	UN_MTK	UN_IPA	Angket_Siswa	Angket_Ortu	IQ	Score_IPA	Score_IPS	Saran	Realitas
4060268	83,60	80,30	82,00	67,50	75,00	IPA	IPA	125,00	123,00	121,00	IPA	IPA
4060011	74,30	76,50	80,50	77,50	70,00	IPA	IPA	123,00	121,00	119,00	IPA	IPA
4080152	78,16	77,00	75,83	65,00	72,50	IPA	IPA	95,00	91,00	93,00	IPS	IPS
4080055	80,83		82,83	60,00	70,00	IPS	IPS	104,00	102,00	103,00	IPS	IPS
4050302	84,50	81,50	83,70	75,00	77,50	IPA	IPA	114,00	113,00	110,00	IPA	IPA
4070532	76,00	77,00	77,00	45,00	62,50	IPS	IPS	117,00	114,00	113,00	IPA	IPS
4060251	81,67	79,00	80,67	72,50	62,50	IPA	IPA	108,00	107,00	106,00	IPA	IPA
4060212	83,40	80,60		77,50	62,50	IPA	IPA	94,00	92,00	93,00	IPS	IPA
4060766	76,67	78,83	83,00	60,00	7,50	IPA	IPA	97,00	95,00	96,00	IPS	IPS
4080125	83,33	83,50	86,50	92,50	76,25	IPA	I;PA	116,00	114,00	112,00	IPA	IPA
4070130	81,80	77,70	77,50	65,00	70,00	IPA	IPA	108,00	107,00	106,00	IPA	IPA
4080361	85,50	87,10		50,00	72,50	IPA	IPA	104,00	100,00	101,00	IPS	IPA
4080029	91,50	89,30	88,70	97,50	77,50	IPA	IPA	114,00	113,00	111,00	IPA	IPA
4070486	76,00	76,00	78,20	65,00	65,00	IPS	IPS	108,00	105,00	103,00	IPA	IPA
4070264	82,00	81,80	81,50	65,00	67,50	IPA	IPA	103,00	100,00	101,00	IPS	IPA
4070122	76,40	80,10	80,20	67,50	62,50	IPA	IPA	117,00	114,00	112,00	IPA	IPA
4070176	78,40	78,80	78,80	47,50	72,50	IPA	IPA	104,00	99,00	101,00	IPS	IPA
4070189	74,20		78,00	60,00	60,00	IPS	IPS	106,00	105,00	104,00	IPA	IPS
4070392	81,80	79,00	78,40	82,50	62,50	IPA	IPA	114,00	113,00	111,00	IPA	IPA

Gambar 3. *Dataset* Peserta Didik SMAN 8 Semarang
Sumber : Setiawati, (2016)

Data Preprocessing

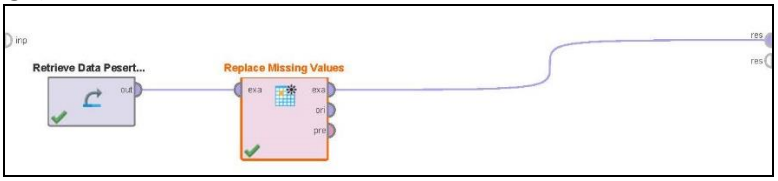
1) *Data Cleaning*

Pada hasil statistik RapidMiner, terdapat *missing value* pada *dataset* yang digunakan.



Gambar 4. *Statistics* RapidMiner *Dataset* Awal

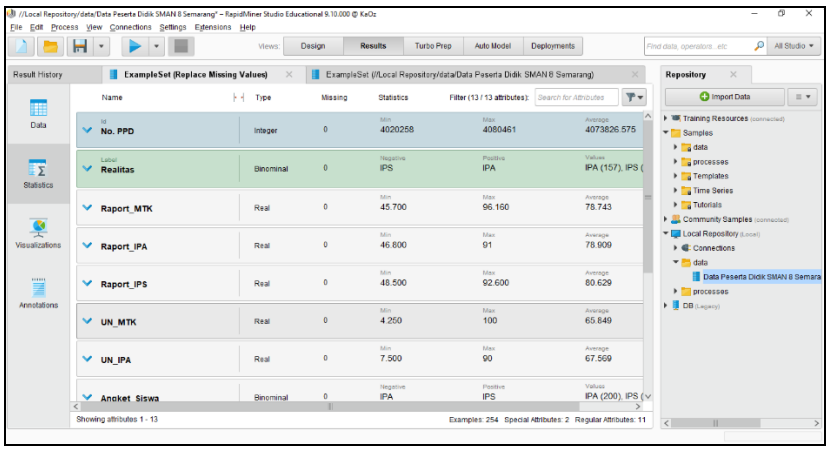
Untuk menghilangkan *missing value* pada *dataset* tersebut, dilakukan preprocessing *data* dengan teknik *data cleaning*. Di sini data *missing value* diisi dengan nilai rata-rata (*average*).



Gambar 5. Proses *Replace Missing Value*

(*Data Cleaning*) pada RapidMiner

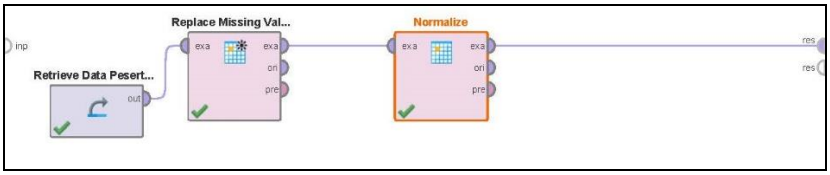
Hasil dari *data cleaning* untuk *missing value* data dilihat pada gambar *statistic* RapidMiner di bawah ini.



Gambar 6. *Statistics* RapidMiner *Dataset* Setelah Proses *Replace Missing Value* (*Data Cleaning*)

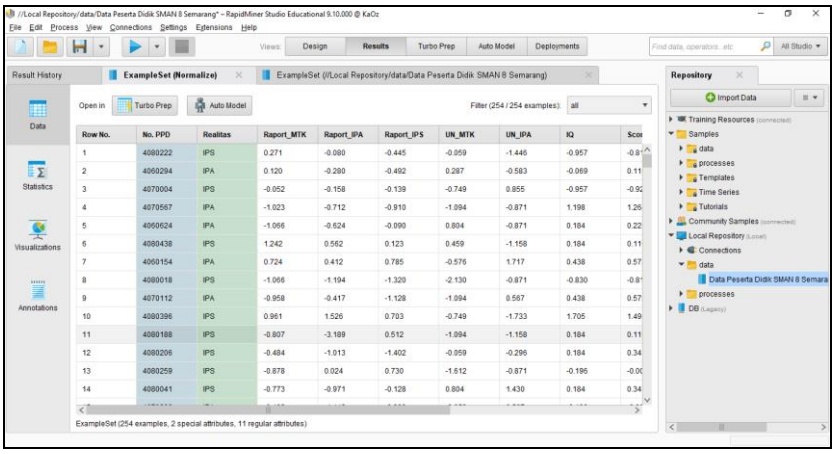
2) *Data Transformation*

Dataset yang telah melalui proses *data cleaning* masih perlu dilakukan proses *data transformation* hal ini dikarenakan pada *dataset* tersebut terdapat data dengan *range* nilai yang berbeda. Proses *data transformation* yang digunakan di sini adalah normalisasi (*normalize*). Prosesnya ditunjukkan gambar berikut:



Gambar 7. Proses *Normalize* (*Data Transformation*) pada RapidMiner

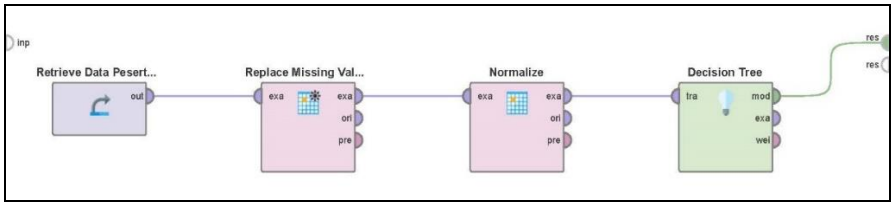
Hasil dari proses *data transformation* tersebut ditunjukkan pada tabel *dataset* RapidMiner di bawah ini:



Gambar 8. *Dataset* pada RapidMiner
Setelah Proses *Normalize* (*Data Transformation*)

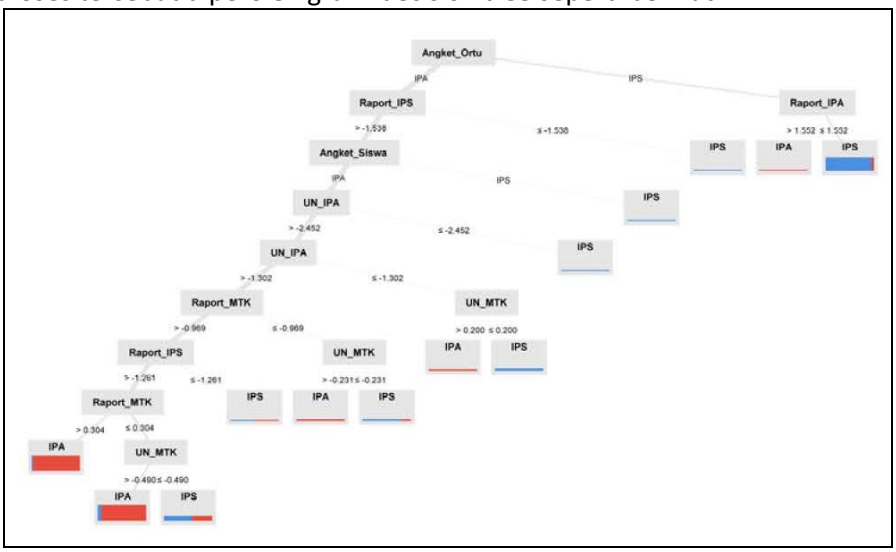
3) *Data Mining*

Setelah proses normalisasi, *dataset* siap untuk diproses dengan metode *data mining*. Teknik *data mining* yang dipakai ialah metode *classification decision tree* dengan algoritma C4.5.



Gambar 9. Proses *Data Mining* pada RapidMiner

Dari proses tersebut diperoleh grafik *decision tree* seperti berikut ini:

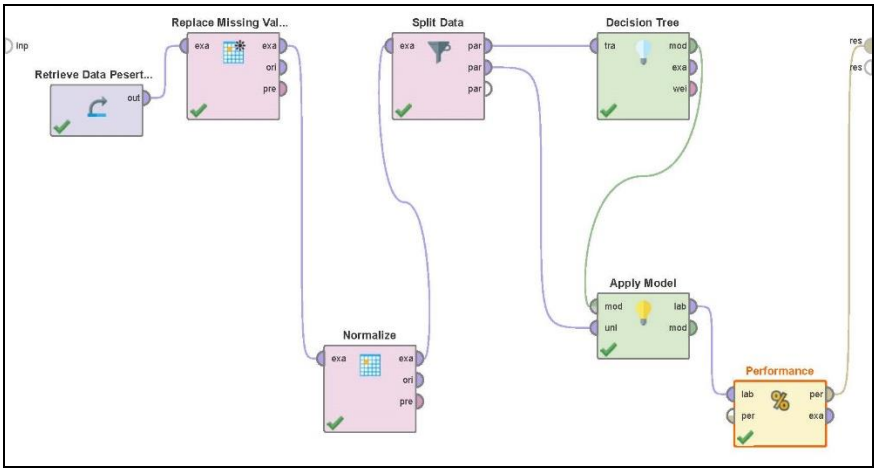


Gambar 10. Grafik *Decision Tree* pada RapidMiner

Dari grafik *decision tree* di atas, nampak bahwa atribut Angket_Ortu mempunyai pengaruh yang besar terhadap menentukan minat dari peserta didik dan dari atribut Angket_Ortu dapat dibagi lagi menjadi cabang lagi yaitu Raport_IPA dan Raport_IPS yang dapat mempengaruhi penentuan minat peserta didik. Sehingga dari grafik *decision tree* tersebut dapat diketahui atribut-atribut yang berpengaruh terhadap penentuan peminatan peserta didik.

4) Evaluasi dan Validasi

Untuk menguji atau mengevaluasi keakuratan dan performansi dari metode *data mining classification decision tree* dengan algoritma C4.5, digunakan *Confusion Matrix* dan ROC *Curve* (AUC). Namun sebelum itu, *dataset* yang telah dinormalisasi dibagi menjadi *data training* (data latih) dan *data testing* (data uji) terlebih dahulu dengan menggunakan operator *split data*. Adapun prosesnya ditunjukkan pada gambar berikut ini:



Gambar 11. Proses Evaluasi dan Validasi pada RapidMiner

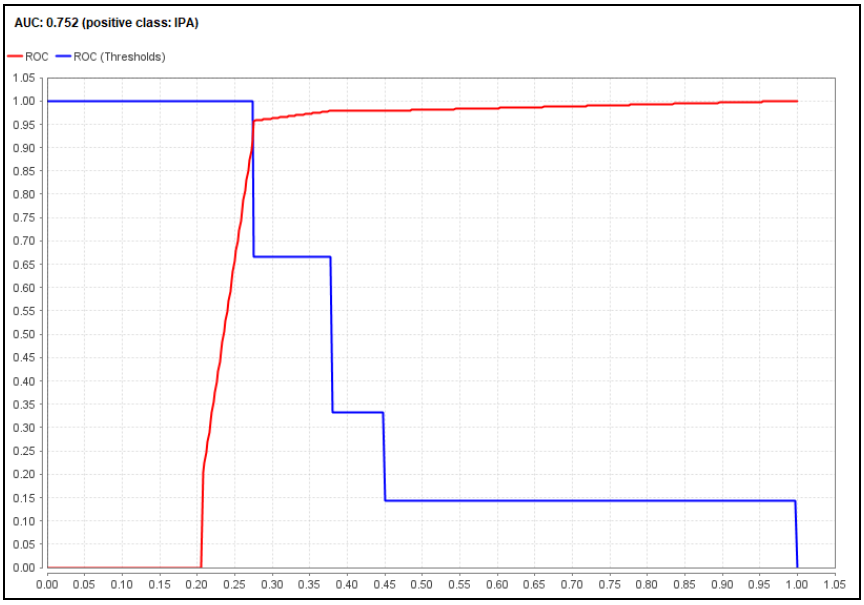
Berikut ini tingkat akurasi yang diperoleh dari proses *data mining* metode classification *decision tree* dengan algoritma C4.5.

Tabel 2. Akurasi *Confusion Matrix* pada RapidMiner

Accuracy : 86.84%			
	True IPS	True IPA	Class Precision
Pred. IPA	21	2	91.30%
Pred. IPS	8	45	84.91%
Class Recall	72.41%	95.74%	

Dari tabel *confusion matrix* di atas diketahui bahwa tingkat akurasi yang dihasilkan dari metode atau teknik *data mining* yang diterapkan untuk menentukan minat dari peserta didik sebesar 86,84%. Dari data *testing* yang berjumlah 76, prediksi yang masuk IPA dengan realita masuk IPA sebanyak 45 dan jumlah prediksi

yang masuk IPS dengan realita masuk IPS sebanyak 21. Sedangkan sisanya, hasil prediksi dengan realita tidak sesuai.



Gambar 12. ROC *Curve* (AUC)

Pada gambar ROC (*Receiver Operting Characteristic*) *Curve* di atas, nilai AUC sebesar 0,752 mendekati 1 yang berarti model prediksi yang dihasilkan semakin baik.

Simpulan

Berdasarkan pembahasan dan hasil dari penelitian yang telah selesai dilakukan, dapat diambil kesimpulan bahwa tingkat akurasi dari model prediksi yang dihasilkan sebesar 86,84% dan nilai AUC sebesar 0,752 termasuk dalam kategori *fair classification* (Gorunescu, 2011). Dan penentuan minat dari peserta didik pada penelitian sebelumnya oleh (Setiawati, 2016) yang menggunakan *dataset* yang sama dengan model *hybrid* metode SAW dan TOPSIS menghasilkan nilai akurasi sebesar 80,3%. Dengan demikian, model prediksi menggunakan *data mining metode classification decision tree* dengan algoritma C4.5 menghasilkan tingkat atau nilai akurasi yang lebih tinggi atau lebih baik, sehingga dapat dijadikan sebagai acuan dalam prediksi penentuan peminatan peserta didik.

Daftar Pustaka

Aprilla, C. D., Baskoro, D. A., & dkk. (2013). *Belajar Data Mining dengan Rapid Miner.* Jakarta: Open Content Model.

Arcinas, M. M., Sajja, G. S., Asif, S., & dkk. (2021). Role of Data Mining in Education for Improving Students Performance for Social Change. 6519–6526.

Głębockaa, A. D., & Zdrodowskab, M. (2021). Analysis Children with Disabilities Self Care Problems based on Selected Data Mining Techniques. 2854–2862.

Gorunescu, F. (2011). *Data Mining Concepts, Models and Techniques.* Verlag Berlin Heidelberg: Springer.

Hoffler, T. N., Kohler, C., & Parchmann, I. (2019). Scientist of the Future: An Analysis of Talented Students Interests. 1–8.

Irawan, Y. (2019). Implementation of Data Mining for Determining Majors using k-Means Algorithm in Students of SMA Negeri 1 Pangkalan Kerinci. *J. Appl. Eng. Technol. Sci*, 17–29.

Ito, T. A., & McPherson, E. (2018). Factors Influencing High School Students Interest in pSTEM. 1–13.

J. Han, M. K., & Pei, J. (2012). *Data Mining, Concepts and Techniques.* Massachusetts: Morgan Kaufmann Publishers.

Kamath, P., P. Patil, S. S., & S, S. S. (2021). Crop Yield Forecasting using Data Mining. 1–7.

Larose, D. T. (2006). *Data Mining Methods and Models.* New Jersey: ohn Wiley & Sons, Inc.

Martiana, E. (2003). Data Preprocessing. 1–13.

Matzavela, V., & Alepis, E. (2021). Decision Tree Learning Through a Predictive Model for Student Academic Performance in Intelligent M-Learning Environments. 1–12.

Pemerintah, P. (2003). Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional. 1–33.

Pendidikan, B. P. (2013). *Pedoman Peminatan Peserta Didik.* Jakarta: Kementerian Pendidikan dan Kebudayaan.

Saranya, C., & Manikandan, G. (2013). A Study on Normalization Techniques for Privacy Preserving Data Mining. 2701–2704.

Setiawati, I. (2016). *Model Hybrid Metode SAW dan TOPSIS untuk Menentukan Peminatan Peserta Didik SM.* Universitas Dian Nuswantoro Semarang.

Viloria, A. (2019). Data Mining Techniques and Multivariate Analysis to Discover Patterns in University Final Researches. 581–586.

Viloria, A. (2019). Determinating Student Interactions in a Virtual Learning Environment using Data Mining. 587–592.

Wood, D. A. (2021). Prediction and Data Mining of Burned Areas of Forest Fires: Optimized Data Matching and Mining Algorithm Provides Valuable Insight. 24–42.