Data Mining di Bidang Pendidikan untuk Analisa Prediksi Kinerja Mahasiswa dengan Komparasi 2 Model Klasifikasi pada STMIK Jabar

Galih

Sistem Informasi, STMIK JABAR, *Jl. Soekarno Hatta no.775-777, Kota Bandung, 40292*
E-mail: galih@stmikjabar.ac.id

Abstract

Until now many universities that helped the government in accelerating improving the quality of education so the creation of a competitive environment.. The abundance of data contained in the College can be put to good use in accordance with the needs and processed into useful information so as to find out the relationship between the attributes of the data in it can be on the analysis and expected output in the form of student performance has related to the period of study i.e. can be categorized into appropriate or too late in the anticipated period of study. Data mining can be used for educational institutions or institutions and often called as Educational Data Mining (EDM). In the study carried out using two models of Naive Bayes Classifier i.e. algorithms and C 4.5. as for the value of the best accuracy in the Naive Bayes Classifier algorithm model (NBC) was 86.83% with ratio 80% training data, whereas in model algorithm C 4.5 was 88.10% with 90% training data ratio. Application of EDM and expected to be maximized and developed so that it can contribute to and progress in the world of education especially in data mining.

Keywords : Educational Data Mining (EDM), Naive Bayes Classifier (NBC), Decision Tree C4.5, Classification

1. Pendahuluan

Pendidikan sangat berpengaruh terhadap kemajuan sebuah negara dengan menghasilkan sumber daya manusia yang dapat mengembangkan peradaban menjadi lebih baik. Adapun peran perguruan tinggi salah satunya adalah berperan dalam menyumbang sumber daya manusia yang handal agar dapat mendorong pertumbuhan dan menciptakan kehidupan manusia yang lebih baik dari sebelumnya.

Sampai sekarang banyak sekali perguruan tinggi yang ikut membantu pemerintah dalam percepatan peningkatan mutu pendidikan sehingga terciptanya sebuah lingkungan yang kompetitif. Adapun salah satu cara didalam perguruan tinggi yaitu dengan menggunakan data pendidikan dan menggali pengetahuan didalamnya sehingga dapat ditemukan mengenai atribut utama yang dapat berpengaruh terhadap kualitas kinerja mahasiswa (Tair & El-halees, 2012) salah satu atribut yang dapat digunakan sebagai parameter keberhasilan sebuah perguruan tinggi yaitu mahasiswa dapat melaksanakan masa studi secara tepat pada waktunya sesuai jenjang yang diambil. Berlimpahnya data pada perguruan tinggi dapat digunakan secara maksimal sesuai dengan kebutuhan dan mampu diolah menjadi informasi yang bermanfaat sehingga dapat mengetahui hubungan antara atribut data yang di dalamnya dapat dianalisis dan diharapkan memiliki keluaran berupa kinerja mahasiswa yang berhubungan dengan masa studi yaitu dapat dikategorikan menjadi tepat atau terlambat dalam menempuh masa studi. Untuk mengolah data dapat dilakukan dengan memanfaatkan data mining dalam menyelesaikan masalah terkait pendidikan. Data mining adalah sebuah metode dalam menemukan informasi berharga dari sejumlah data yang dilakukan dengan memanfaatkan ilmu lain seperti statistik, matematika, pengenalan pola (Larose & Larose, 2014).

Data Mining sangat bermanfaat dan banyak diterapkan pada berbagai bidang, misalnya bidang analisis pemasaran, kedokteran, rekayasa manufaktur, pendidikan dan lain lain. Data mining dapat diaplikasikan pada bidang lembaga atau institusi Pendidikan dan sering disebut juga dengan *Educational Data Mining (EDM)* yaitu sebuah pengembangan metode dalam mengeksplorasi jenis tipe data pendidikan yang bersifat unik yang bertujuan untuk mempelajari dalam memahami kinerja siswa dan pengaturan lingkungan di tempat siswa belajar (Romero & Ventura, 2010). *Educational Data Mining (EDM)* dapat diterapkan dengan banyak

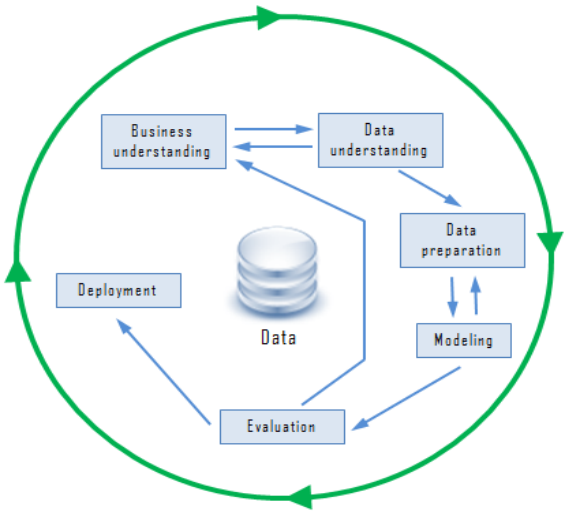
metode seperti menggunakan teknik Pohon Keputusan (Decision Tree), Jaringan Syaraf Tiruan, Naive Bayes Classifier, dan lain-lain.

Pada penelitian bidang pendidikan telah dilakukan oleh Badr, Din dan Elaraby (Badr, Din, & Elaraby, 2014) dalam menganalisis prediksi kualitas prestasi siswa dengan menggunakan algoritma Pohon Keputusan ID3, adapun output nya adalah menghasilkan model yang dapat digunakan untuk memprediksi nilai akhir dari siswa. Devasia Tismy , T P Vinushree dan Hegde Vinayak (Devasia, Vinushree, & Hegde, 2016) dalam penelitiannya melakukan prediksi kinerja siswa dengan melakukan komparasi beberapa algoritma yaitu Naive Bayes Classifier (NBC), Regression, Pohon Keputusan dan Jaringan Syaraf Tiruan dengan percobaan dilakukan terhadap 700 siswa dengan memiliki 19 atribut di dalamnya dan menghasilkan bahwa Naive Bayes memberikan hasil akurasi lebih baik dibandingkan algoritma lainnya. Penelitian dilakukan oleh Shakeel dan Anwer Butt (Shakeel & Anwer Butt, 2015) mengenai klasifikasi mahasiswa putus sekolah dengan *Educational Data Mining (EDM)* menggunakan empat model klasifikasi yaitu Naive Bayes Classifier, Random Forest, J48 graft dan Bayessian Logistic yang bertujuan untuk membantu para mahasiswa dan administrasi akademik dalam melakukan identifikasi dan perbaikan agar mahasiswa dapat memperbaiki kinerjanya dan membantu mengurangi mahasiswa putus sekolah *(Drop Out).* Dalam penelitiannya menghasilkan algoritma Naive Bayes Classifier (NBC) dengan tingkat akurasi tertinggi dibandingkan dengan tiga algoritma lainnya yaitu sebesar 91,9355%. Penelitian yang dilakukan oleh Makhtar, Nawang dan Nor (Makhtar, Nawang, & Nor, 2017) melakukan analisa kinerja siswa dengan data total sebanyak 488 siswa dengan memiliki atribut berupa 7 nilai mata pelajaran serta melakukan komparasi algoritma pengelompokan yaitu Naive Bayes Classifier, Random Tree, Nearest Neighbourhood (IB1), Multi Class Classifier, Conjunctive Rule yang menghasilkan keluaran berupa pengaruh nilai mata kuliah terhadap kinerja siswa dan menghasilkan Naive Bayes Classifier sebagai algoritma dengan nilai akurasi terbaik dibandingakan empat algoritma lainnya. Ihsan A Abu Amra dan Ashraf Y.A. Maghari (Amra & Maghari, 2017) melakukan penelitian dalam memprediksi performa siswa pada sekolah menengah, data dikumpulkan dari kementrian pendidikan di jalur Gaza pada tahun 2015 dan menghasilkan dataset sebanyak 500 *records* setelah di seleksi dari total data siswa sebanyak 33294. Adapun dalam eksperimen penelitian menghasilkan Naïve Bayes Classifier lebih unggul dari KNN dengan menghasilkan nilai dengan akurasi tertinggi yaitu 93,6%.

Berdasarkan pada penelitian sebelumnya, pemanfaatan EDM banyak dilakukan dengan menggunakan algoritma di antaranya Decision Tree, Naive Bayes Classifier (NBC) Neural Network dan lain-lain. Menghasilkan NBC dan Decision Tree sebagai algoritma dengan nilai akurasi yang paling baik. Sehingga dapat disimpulkan untuk melakukan penelitian dengan menggunakan kedua algoritma tersebut yaitu Naïve Bayes Classifier dan Decision Tree untuk dilakukan perbandingan algoritma mana yang lebih baik sehingga menjadi landasan dan masukan untuk LABKOM STMIK Jabar agar diterapkan pada pengembangan sistem informasi akademik cerdas kedepannya yang dapat membantu lebih cepat dalam membuat keputusan bagi pihak kampus terhadap penanganan mahasiswa yang terindikasi terlambat dalam masa studi nya.

2. Metode Penelitian

Dalam penelitian ini akan mengimplementasikan model standar dalam proses penggalian data yaitu CRISP-DM dalam memecahkan masalah yang terjadi contohnya pada bidang bisnis ataupun dalam bidang penelitian, adapun proses yang terjadi dalam penggalian data memiliki siklus hidup *(life cycle)* yang dibagi menjadi enam tingkatan masa atau tahapan proses (Larose, 2006)



Gambar 1 Tahapan CRISP-DM

2.1. Tahapan Pemahaman Bisnis

Institusi pendidikan pada jenjang perguruan tinggi sangat diharapkan agar selalu memberikan peningkatan kualitas baik dari sisi institusi dan juga lulusannya. Adapun yang dapat menjadi keberhasilan sebuah institusi perguruan tinggi dapat terlihat dari peran mahasiswa yaitu dalam ketepatan waktu menyelesaikan masa studi yang menjadi salah satu elemen penting didalam borang dan evaluasi untuk akreditasi perguruan tinggi (BAN-PT, 2011) sehingga berdampak juga kepada kualitas sebuah perguruan tinggi.

2.2. Tahapan Pemahaman Data

Pada penelitian ini dataset yang akan dimanfaatkan merupakan database yang di dapatkan dari bagian akademik pada STMIK JABAR untuk jurusan Sistem Informasi dan Teknik Informatika jenjang S-1 angkatan 2010, 2011, 2012, 2013, 2014, 2015, 2016 dengan jumlah *records* 946. Data yang digunakan merupakan profil mahasiswa dan akademik mahasiswa berupa nilai IPK dan keterangan Masa Studi serta penambahan atribut jarak yang dibagi menjadi 2 kategori yaitu : <10 km dan >10 km yang diharapkan dapat berpengaruh terhadap hasil analisa.

2.3. Tahapan Pengolahan Data

Dataset pada penelitian harus dilakukan penyederhanaan terlebih dahulu agar dapat dimanfaatkan sesuai dengan metode yang diusulkan, adapun dalam pengolahan awal data adalah sebagai berikut:

Tabel 1. Atribut Data Penelitian.

Atribut	Deskripsi	Nilai
NIM	Nomor Induk Mahasiswa	ID
Jenis Kelamin	Jenis kelamin mahasiswa : Pria atau Wanita	Pria atau Wanita
Kategori Sekolah Asal	Kategori sekolah asal mahasiswa sebelum melanjutkan ke Perguruan Tinggi	Sekolah Menegah Atas, Sekolah Menegah Kejuruan, Madrasyah Aliyah
Asal Kota	Menjelaskan asal kota lahir mahasiswa bersangkutan	Bandung, Bekasi, Ciamis, Cianjur, Dili, Garut, Jakarta, Palembang, Pangalengan, Semarang, Sukabumi, Tasikmalaya
Jarak Tempat Tinggal ke Kampus	Jarak tempat tinggal mahasiswa ke kampus yaitu : <10km atau >10km	< 10km atau > 10km
Pekerjaan Orangtua	Status pekerjaan orangtua atau wali mahasiswa	Petani, PNS, POLRI, Swasta, TNI, Wiraswasta
IPK	Indeks Prestasi Kumulatif	0,00 sampai 4,00
Keterangan Lulus	Variabel yang dijadikan sebagai label. Jika mahasiswa menempuh kuliah S1 dalam waktu kurang dari sama dengan empat tahun maka memiliki status Tepat , jika lebih maka status Terlambat	Tepat atau Terlambat

Pada Tabel 1 merupakan database mahasiswa di STMIK JABAR yang akan dilakukan pemilihan dan pengurangan data mahasiswa yang didalamnya memiliki atribut dan jumlah *record* yang diperlukan sesuai dengan kebutuhan. Untuk tahapan eliminasi beberapa atribut dilakukan terhadap nilai yang diduga tidak terlalu berpengaruh terhadap proses pengklasifikasian yaitu pada kolom nama, agama, dan nomor ijazah. sedangkan nilai atribut yang sering kosong pada data mahasiswa yaitu terjadi pada kolom nilai gaji orangtua sehingga pada atribut ini dilakukan proses eliminasi. Setelah dilakukan pengolahan data awal *(preprocessing)* maka dihasilkan data valid dengan jumlah 836 *records*.

2.4. Tahapan Pemodelan

Model yang diusulkan akan menggunakan 2 teknik klasifikasi dan melakukan komparasi terhadap keduanya yaitu NBC dan Decision Tree

untuk menghasilkan nilai akurasi tertinggi dengan menggunakan *tools* data mining yaitu RapidMiner Studio 9.2.000

2.5. Tahapan Evaluasi

Setelah dilakukan tahapan pengujian model selanjutnya akan diberlakukan pengukuran nilai untuk menguji tingkat akurasi dengan memanfaatkan Confusion Matrix dan Split Validation sebagai proses validasinya.

2.6. Tahapan Penyebaran

Diharapkan pada penelitian ini dapat menghasilkan sebuah analisa dengan akurasi yang baik, sehingga dapat membantu institusi perguruan tinggi dalam menghadapi masalah keterlambatan masa studi mahasiswa yang berdampak pada kualitas pendidikan dan penilaian akreditasi. Hasil penelitian dapat dijadikan sebagai bahan penelitian lanjutan khususnya pada bidang sistem penunjang keputusan yang dapat membantu dalam pengidentifikasian potensi mahasiswa yang terlambat dalam menyelesaikan masa studi nya.

Berdasarkan pada penelitian sebelumnya, pemanfaatan EDM banyak dilakukan dengan menggunakan algoritma di antaranya Decision Tree, Naive Bayes Classifier (NBC) Neural Network dan lain lain. Menghasilkan NBC dan Decision Tree sebagai algoritma dengan nilai akurasi yang paling baik. Sehingga diputuskan untuk melakukan perbandingan dengan kedua algoritma tersebut dan memilih hasil yang paling baik.

2.7. Naive Bayes Classifier (NBC)

NBC adalah metode pada *probabilistic reasoning* dengan cara menghitung beberapa kumpulan probabilitas dan melakukan penjumlahan frekuensi dan kombinasi nilai dari sebuah dataset (Tempola, Muhammad, & Khairan, 2018)

Metode ini mempunyai fungsi dalam mendapatkan nilai probabilitas untuk menentukan keputusan karena terdapat penghitungan yang dapat menentukan resiko dari setiap kasus. Bentuk umum persamaannya seperti berikut (Bustami, 2014):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Dimana :

X = Data yang class nya belum diketahui

H = Hipotesis data merupakan suatu class spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X *(Posteriori Probability)*

P(H) = Probabilitas hipotesis H *(Prior Probability)*

P(X|H) = Probabilitas hipotesis X berdasarkan kondisi H

P(X) = Probabilitas hipotesis X

2.8. Decision Tree C4.5

Decision tree memodelkan klasifikasi sampel secara *top-down*, yaitu dimulai dari simpul akar *(root)* dengan menjaga jarak dengan hasil dari tes *node* internal, sampai simpul daun yang dicapai oleh kelas label yang ditugaskan. Adapun keuntungan yang paling penting dari Decision Tree atau Pohon Keputusan adalah bahwa pengetahuan dapat dipisahkan atau dilakukan proses ekstraksi dan dapat diwakili dalam sebuah bentuk aturan klasifikasi yaitu percabangan *if-then* oleh Yu, Huang Hu, dan Cai (Yu, Huang, Hu, & Cai, 2010) Algoritma C4.5 berbentuk struktur pohon di dalamnya terdapat simpul atau biasa disebut juga dengan *node* yang menggambarkan setiap atribut, kemudian cabang yaitu merupakan hasil dari atribut yang telah diuji dan ada juga yang disebut dengan daun yang merupakan representasi dari kelas. Algoritma C4.5 memiliki proses kerja yaitu memasuki setiap simpul keputusan dan melakukan proses pembagian yang optimal sampai tidak bisa dibagi lagi dengan menggunakan konsep information gain atau entropy reduction. Adapun tahapan yang harus dilakukan untuk membentuk pohon keputusan dengan algoritma C4.5 adalah sebagai berikut :

2.8.1. Menyiapkan data latih

Data latih merupakan data histori yang sudah pernah diolah sebelumnya dan sudah dibagi menjadi beberapa kelas yang telah ditentukan.

2.8.2. Menentukan akar dari pohon

Dalam menentukan akar diperoleh dari atribut yang telah terpilih, yang dilakukan dengan proses menghitung nilai gain yang ada pada tiap-tiap atribut dan jika nilai gain paling tinggi yang dihasilkan maka akan menjadi akar pertama. Sebelum menentukan nilai gain, tentukan terlebih dahulu nilai entropy yaitu merupakan distribusi *probabilitas* pada teori informasi yang digunakan oleh Pohon Keputusan (C4.5) dalam menentukan tingkat kesamaan *(homogenitas)* distribusi kelas

dari sebuah dataset. Jika pada sebuah dataset memiliki nilai yang tinggi tingkat entropy nya maka semakin homogen distribusi kelas tersebut. Untuk menghitung nilai entropy digunakan rumus:

$$Entropy(\mathrm{S}) = \sum_{i=1}^{n} -pi * log2pi$$

Keterangan :
S = himpunan (dataset) kasus
n = banyaknya partisi S
pi = proporsi Si terhadap S

2.8.3. Menghitung nilai *Gain* dengan persamaan

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

Keterangan :
S = himpunan (dataset) kasus
A = fitur
n = jumlah partisi atribut A
|Si| = proporsi Si terhadap S
|S| = jumlah kasus dalam S

2.8.4. Eksperimen Dan Pengujian Model

Pada eksperimen dan pengujian model penelitian tahapannya adalah sebagai berikut:

1. Menyiapkan data untuk dijadikan sebagai bahan eksperimen menggunakan software microsoft excel 2010 dengan format file .xls
2. Kemudian masuk ke pengolahan awal data yang dilakukan dengan menghapus data-data yang kosong dan diduga tidak berpengaruh terhadap hasil klasifikasi.
3. Implementasi pengujian model pada penelitian ini dilakukan dengan menggunakan software Rapidminer versi 8.2.
4. Pada pengujian ini dilakukan menggunakan dua model algoritma yaitu Naive Bayes Classifier dan C4.5 untuk mendapatkan hasil nilai tingkat akurasi, AUC dan waktu eksekusi pada masing-masing algoritma sehingga dapat dibandingkan nilai akurasi terbaik pada kedua model algoritma tersebut.

2.9. Confusion Matrix

Pengujian pada sistem pengklasifikasi harus berjalan sesuai dengan yang diharapkan, sehingga memberikan hasil nilai dengan tingkat akurasi tertinggi dan menghasilkan nilai error serendah mungkin. Maka pengujian itu dapat dilakukan dengan menggunakan matriks konfusi *(confusion matrix)* (Larose & Larose, 2014).

Tabel 2. Confusion Matrix 2 kelas

Classification	*Predicted Class*	
	Prediction = Yes	Prediction = No
Actual = Yes	a (*true positive* - TP)	b (*false negative* - FN)
Actual = No	c (*false positive* - FP)	d (*true negative* - TN)

Rumus yang digunakan dalam menghitung persamaan model confusion matrix untuk menghasilkan nilai akurasi, recall dan precision:

$$Accuracy = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}}$$

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

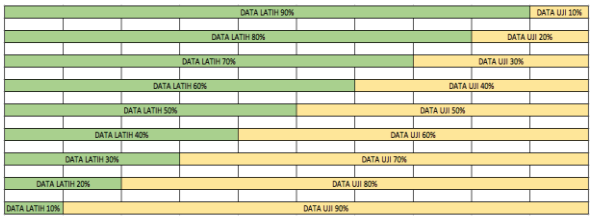
$$Precision\ (positive) = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

$$Recall\ (negative) = \frac{TN}{FN+TN}$$

$$Precision\ (negative) = \frac{\mathrm{TN}}{\mathrm{FN+TP}}$$

2.10. Split Validation

Split validation adalah salah satu operator di dalam Rapidminer yang memiliki fungsi dalam melakukan validasi sederhana secara acak kemudian membagi dataset menjadi 2 bagian yaitu data latih dan data uji dalam mengevaluasi model. Hal ini digunakan terutama dalam memperkirakan seberapa akurat model yang akan diujikan (Rapidminer Inc, 2019) seperti pada ilustrasi gambar 2 berikut:



Gambar 2 Ilustrasi Split Validation

2.11. Kurva *Receiver Operating Characteristic (ROC)*

Dapat dimanfaatkan untuk melihat informasi nilai akurasi dan dapat membantu

dalam membandingkan klasifikasi secara visual sehingga memberikan kemudahan dalam pembacaan. ROC mengekspresikan confusion matrix. Adapun kurva ROC dapat digunakan dalam menilai nilai AUC (Area Under Curve) yaitu nilai untuk menentukan nilai akurasi klasifikasi pengujian diagnostik, sebagai berikut (Gorunescu, 2011)

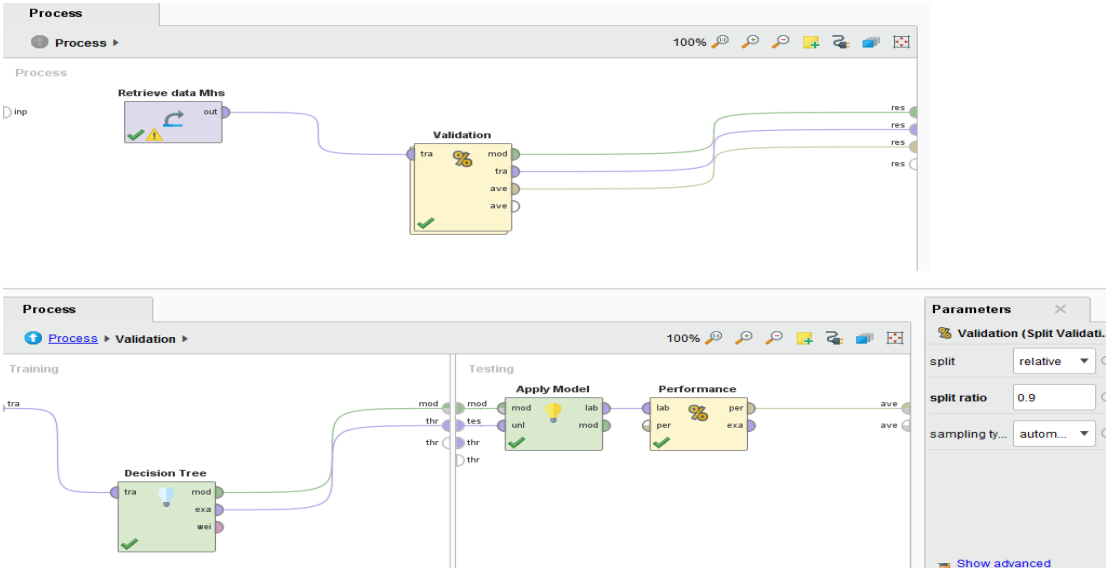
Tabel 3. Nilai AUC (Area Under Curve)

Nilai Akurasi	**Hasil Klasifikasi**
Batas Nilai 0.90 - 1.00	Sangat Baik
Batas Nilai 0.80 - 0.90	Baik
Batas Nilai 0.70 – 0.80	Sedang
Batas Nilai 0.60 – 0.70	Lemah
Batas Nilai 0.50 – 0.60	Sangat Lemah

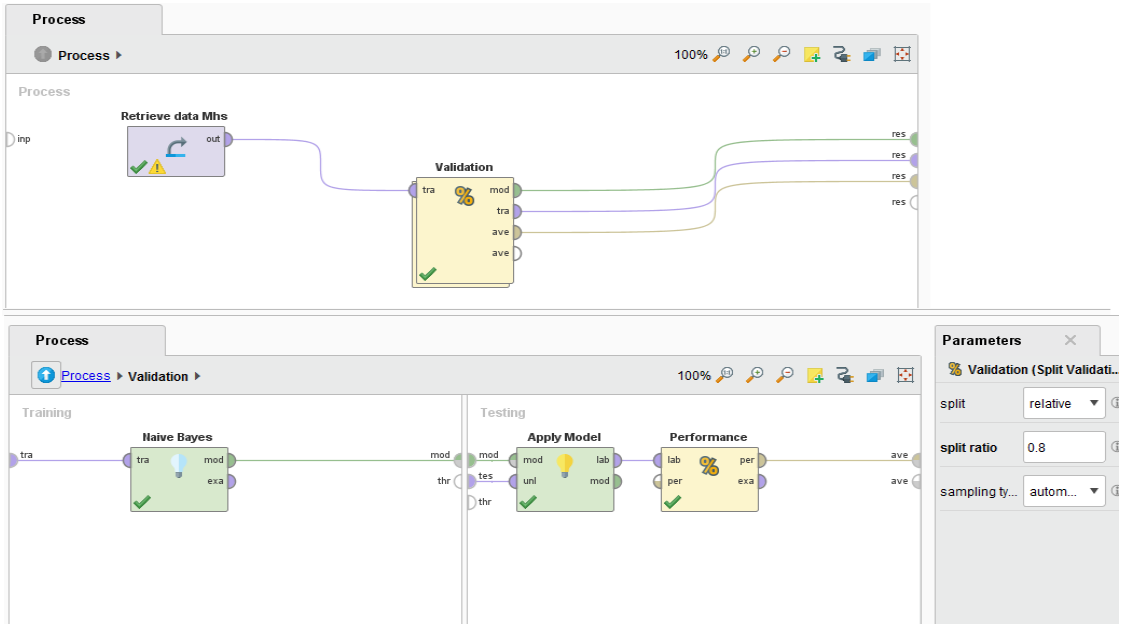
3. Pembahasan dan Hasil Penelitian

Pada proses pengujian validasi terhadap dataset dilakukan menggunakan Rapid Miner Studio seperti pada Gambar 3 dan 4.

3.1. Validasi dan Evaluasi



Gambar 3 Desain Model C4.5



Gambar 4 Desain Model Naïve Bayes Classification (NBC)

Retrieve Data : Operator ini berfungi untuk mengimport data yang akan diujikan terhadap model klasifikasi dengan format file excel
Validation : Penelitian dilakukan validasi dengan menggunakan Split Validation
Navie Bayes : Model klasifikasi yang digunakan
Decision Tree : Model klasifikasi yang digunakan
Apply Model : Operator yang berfungsi dalam menjalankan model klasifikasi
Performance : Jenis operator untuk mendapatkan nilai akurasi dan performa

3.2. Hasil Penelitian

Hasil dari implementasi data mining dengan menggunakan software Rapidminer, dilakukan terhadap dua model algoritma klasifikasi yaitu C4.5 dan NBC kemudian memasukan dataset sebagai bahan uji untuk kedua model tersebut yang di dalamnya terdapat data latih dan data uji. Adapun fungsi dari data latih adalah sebagai pembentuk model klasifikasi, sedangkan data uji memiliki fungsi sebagai pengujian dalam mengevaluasi kualitas hasil klasifikasi dari algoritma yang digunakan.

Tabel 1. Hasil Akurasi Komparasi Dua Model Algoritma

Naive Bayes Classifier (NBC)	**Ratio Data Training (%)**								
	10	20	30	40	50	60	70	80	90
Akurasi (%)	77,66	81,76	84,44	84,26	84,21	84,73	84,46	**86,23**	82,14
AUC	0,874	0,907	0,925	0,917	0,906	0,928	0,923	**0,940**	0,938
C4.5	**Ratio Data Training (%)**								
	10	20	30	40	50	60	70	80	90
Akurasi (%)	78,59	83,71	83,42	85,66	80,62	86,23	84,86	88,02	**88,10**
AUC	0,855	0,851	0,846	0,856	0,824	0,861	0,845	0,882	**0,886**

4. Kesimpulan

Setelah dilakukan penelitian maka dihasilkan beberapa kesimpulan yaitu sebagai berikut:

a. Berdasarkan pengujian pada kedua model algoritma tersebut dengan menggunakan *ratio data training* dapat mempengaruhi hasil dari nilai akurasi masing-masing, adapun nilai akurasi terbaik pada model algoritma Naive Bayes Classifier (NBC) adalah 86,83% dengan ratio data training 80%, sedangkan pada model algoritma C4.5 adalah 88,10 % dengan ratio data training 90%.
b. Dapat disimpulkan nilai akurasi terbaik dari hasil komparasi kedua model algoritma diperoleh oleh model algoritma C4.5 dengan nilai akurasi 88,10%. Dan penerapan EDM diharapkan dapat dimaksimalkan dan dikembangkan sehingga bisa memberikan kontribusi dan kemajuan di dunia pendidikan khususnya pada bidang data mining.
c. Untuk pengembangan Sistem Informasi Akademik Cerdas dalam memprediksi kelulusan masa studi mahasiswa akan menggunakan model algoritma C4.5

Daftar Pustaka

Amra, I. A. A., & Maghari, A. Y. A. (2017). Students performance prediction using KNN and Naïve Bayesian. *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, 909–913. https://doi.org/10.1109/ICITECH.2017.8079967

Badr, A., Din, E., & Elaraby, I. S. (2014). *Data Mining : A prediction for Student ' s Performance Using Classification Method*. *2*(2), 43–47. https://doi.org/10.13189/wjcat.2014.020203

BAN-PT. (2011). *6 Buku 6 Matriks Penilaian Borang Dan Evaluasi Diri Aipt 2011* (pp. 21–23). pp. 21–23. Retrieved from https://banpt.or.id/instrumen/APT.rar

Bustami. (2014). Penerapan Algoritma Naive Bayes. *Jurnal Informatika*, *8*(1), 884–898.

Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, *1*(3), 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

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

Larose, D. T. (2006). DATA MINING METHODS AND MODELS. In *Contemporary Psychology: A Journal of Reviews* (Vol. 21). https://doi.org/10.1037/014836

Larose, D. T., & Larose, C. D. (2014). Discovering Knowledge in Data. In *Discovering Knowledge in Data*. https://doi.org/10.1002/9781118874059

Makhtar, M., Nawang, H., & Nor, S. W. S. (2017). ANALYSIS ON STUDENTS PERFORMANCE USING NAÏVE BAYES CLASSIFIER. *Journal of Theoretical and Applied Information Technology*, *95*(16), 1–8. Retrieved from www.jatit.org

Rapidminer Inc. (2019). Split Validation (RapidMiner Studio Core). Retrieved May 8, 2019, from https://docs.rapidminer.com/latest/studio/operators/validation/split_validation.html

Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Shakeel, K., & Anwer Butt, N. (2015). Educational Data Mining to Reduce Student Dropout Rate by Using Classification. *253rd OMICS International Conference on Big Data Analysis & Data Mining*, (May). Retrieved from https://www.researchgate.net/publication/281149091_Educational_Data_Mining_to_Reduce_Student_Dropout_Rate_by_Using_Classification?enrichId=rgreq-95ac86e292913b9c0a37d546c75096e5-XXX&enrichSource=Y292ZXJQYWdlOzI4MTE0OTA5MTtBUzoyNjUyNzk2MzI1NzI0MTZAMTQ0MDI1

Tair, M. M. A., & El-halees, A. M. (2012). Mining Educational Data to Improve Students 'Performance : A Case Study. *Journal of Theoretical and Applied Information Technology*, *2*(2), 140–146.

Tempola, F., Muhammad, M., & Khairan, A. (2018). *Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan K-Fold Cross Validation Comparison of Classification Between Knn and Naive Bayes At the Determination of the Volcanic Status With K-Fold Cross*. *5*(5), 577–584. https://doi.org/10.25126/jtiik20185983

Yu, H., Huang, X., Hu, X., & Cai, H. (2010). *A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation*. https://doi.org/10.1109/ICMeCG.2010.16