IMPLEMENTASI *MULTILAYER PERCEPTRON* UNTUK PREDIKSI KEGAGALAN STUDI MAHASISWA

Cosmas Haryawan[1)] dan Maria Mediatrix Sebatubun[2)]
[1)] Sistem Informasi, STMIK AKAKOM Yogyakarta
[2)] Informatika, STMIK AKAKOM Yogyakarta
Jalan Raya Janti 143, Karangjambe, Banguntapan, Yogyakarta
e-mail: cosmas@akakom.ac.id[1)], memey@akakom.ac.id[2)]

ABSTRAK

Perguruan Tinggi (PT) merupakan salah satu lembaga yang bergerak di bidang pendidikan dan dapat didirikan oleh Pemerintah maupun swasta. Pada saat sekarang ini, Indonesia telah memiliki ratusan Perguruan Tinggi yang tersebar di seluruh wilayah. Sebagai lembaga pendidikan, tentu saja sebuah Perguruan Tinggi harus mampu mendidik mahasiswanya dan mengeluarkan lulusan-lulusan yang berkualitas secara akademik maupun non akademik. Dalam pelaksanaannya, tentu saja terdapat banyak masalah yang harus diselesaikan dengan sebaik-baiknya, salah satunya adalah ketika terdapat mahasiswa yang dengan sengaja berhenti atau menghilang sebelum menyelesaikan pendidikannya atau bahkan tidak sanggup lagi menyelesaikan pendidikan dan dikeluarkan oleh PT (dropout).

Berdasarkan masalah tersebut, maka penelitian ini membuat suatu model untuk melakukan prediksi terhadap mahasiswa yang berpotensi gagal atau dropout selama masa studinya menggunakan salah satu metode data mining yaitu MultiLayer Perceptron (MLP) dengan mengacu pada data pribadi dan data akademik mahasiswa tersebut. Hasil yang diperoleh dari penelitian ini yaitu tingkat akurasi sebesar 86,9%, sensitivitas 54,7% dan spesifisitas 95,4%. Penelitian ini diharapkan dapat digunakan untuk menentukan strategi yang perlu dilakukan untuk meminimalisir jumlah mahasiswa yang berhenti atau dropout.

***Kata kunci**: Data mining, klasifikasi, multilayer perceptron.*

IMPLEMENTATION OF MULTILAYER PERCEPTRON FOR STUDENT FAILURE PREDICTION

Cosmas Haryawan[1)] and Maria Mediatrix Sebatubun[2)]
[1)] Information System, STMIK AKAKOM Yogyakarta
[2)] Informatics, STMIK AKAKOM Yogyakarta
Jalan Raya Janti 143, Karangjambe, Banguntapan, Yogyakarta
e-mail: cosmas@akakom.ac.id[1)], memey@akakom.ac.id[2)]

ABSTRACT

University is one of the educational institutions and can be established by the government or the individual. At this time, Indonesia has hundreds of universities spread throughout the region. As an educational institution, university of course must be able to educate its students and issue quality graduates with the academically and non-academically qualified. In its implementation, there are many problems that should be resolved as well as possible, such as when there are students who intentionally stop or disappear before completing their education or are even unable to complete their education and issued by institution (dropout).

Based on these problems, this research makes a model for predicting students who have the potential to fail or dropout during their studies using one of the data mining methods namely Multilayer Perceptron by referring to personal and academic data. The results obtained from this research are 86.9% an accuracy rate with the 54.7% sensitivity, and 95.4% specificity. This research is expected to be used to determine the need strategies to minimize the number of students who stop or dropout.

***Keywords**: Classification, data mining, multilayer perceptron.*

I. PENDAHULUAN

PERGURUAN Tinggi (PT) merupakan salah satu lembaga yang bergerak di bidang pendidikan dan dapat didirikan oleh Pemerintah maupun swasta. Pada saat sekarang ini, Indonesia telah memiliki ratusan Perguruan Tinggi yang tersebar di seluruh wilayah. Sebagai lembaga pendidikan, tentu saja sebuah Perguruan Tinggi harus mampu mendidik mahasiswanya dan mengeluarkan lulusan-lulusan yang berkualitas secara akademik maupun non akademik. Untuk itu, berbagai upaya dilakukan oleh pihak Perguruan Tinggi agar dapat menghasilkan generasi-generasi yang berguna bagi Negara. Meskipun telah berupaya semampunya, sebuah

DOI: http://dx.doi.org/10.12962/j24068535.v18i2.a990

Perguruan Tinggi tentu saja memiliki masalah terkait dengan pengelolaan mahasiswanya. Salah satu masalah yang paling sering terjadi adalah ketika terdapat mahasiswa yang berhenti atau menghilang sebelum menyelesaikan pendidikannya atau bahkan tidak sanggup lagi menyelesaikan pendidikan dan dikeluarkan oleh Perguruan Tinggi. Hal ini akan berdampak bagi mahasiswa itu sendiri maupun bagi PT karena dapat dianggap tidak bisa mendidik mahasiswa, atau bahkan mempersulit mereka selama masa pendidikan. Sementara terdapat ratusan mahasiswa yang berhasil menyelesaikan pendidikan mereka tepat waktu bahkan lebih cepat dari waktu normalnya. Meskipun demikian, kasus seperti diatas akan berpengaruh terhadap reputasi dan akreditasi PT tersebut, padahal PT telah mengupayakan berbagai cara untuk dapat mendidik dan mengayomi mahasiswanya dengan baik.

Oleh karena itu, PT juga perlu memikirkan cara dan strategi untuk menyelesaikan permasalahan tersebut. Hasil akhir dari penelitian ini yaitu dapat memberikan kontribusi bagi Perguruan Tinggi untuk memprediksi mahasiswa/i yang berpotensi *dropout* dan menerapkan strategi yang diperlukan sehingga semua mahasiswa/i dapat menyelesaikan pendidikan mereka. Salah satu langkah yang dapat dilakukan adalah dengan menganalisis faktor-faktor penyebab mahasiswa yang masuk ke PT tersebut berhenti sebelum lulus atau bahkan di-*dropout* oleh PT. Hal ini dapat dilakukan secara manual oleh pihak PT, tetapi juga memiliki kendala dikarenakan suatu PT dimungkinkan memiliki ratusan bahkan ribuan mahasiswa yang tentu saja memiliki datanya masing-masing. Untuk melakukan analisis secara manual akan membutuhkan waktu yang cukup lama dan kemungkinan juga tidak dapat memberikan hasil yang maksimal. Berdasarkan permasalahan dan upaya untuk meminimalisir permasalahan tersebut, maka proses analisis ratusan bahkan ribuan data yang membutuhkan waktu lama dapat dilakukan oleh komputer melalui sebuah sistem. Agar sistem dapat melakukan analisis seperti manusia, maka perlu diberikan pengetahuan seperti yang dimiliki oleh manusia sehingga hasil analisisnya akurat dan dapat digunakan sebagai bahan pertimbangan dalam membuat strategi untuk mengatasi mahasiswa yang tidak menyelesaikan studinya.

Salah satu cara yang sedang popular saat ini adalah dengan menggunakan teknik *data mining*. *Data mining* merupakan proses untuk mendapatkan informasi atau biasanya dikenal dengan *knowledge discovering* yang berguna dari sebuah basis data yang besar [1]. Pengetahuan yang diperoleh tersebut nantinya dapat digunakan untuk mengambil keputusan di masa mendatang. Penelitian dalam bidang *data mining* telah banyak dilakukan untuk mengatasi berbagai masalah yang berkaitan dengan data dalam jumlah besar, untuk membangun sebuah model kemudian menggunakan model tersebut untuk mengenali pola data yang lain yang tidak berada dalam basis data. Teknik ini dapat dilakukan dengan menggunakan berbagai metode seperti *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM), *Naïve Bayes*, *Decision Tree* dan lain sebagainya.

Berdasarkan permasalahan tersebut, maka penelitian ini akan melakukan prediksi mahasiswa yang gagal atau *dropout* selama masa studinya menggunakan metode *data mining* yang belum pernah digunakan untuk menyelesaikan permasalahan diatas yaitu *MultiLayer Perceptron* (MLP). Karena metode ini telah banyak digunakan untuk menyelesaikan banyak permasalahan baik yang datanya bersifat linear maupun data nonlinear. Penggunaan metode MLP ini dimaksudkan dengan tujuan agar model yang dihasilkan dapat digunakan untuk melakukan prediksi dengan lebih akurat.

II. TINJAUAN PUSTAKA

Terdapat beberapa penelitian yang telah dilakukan terkait dengan penelitian ini yaitu dengan menggunakan teknik data mining. Salah satu penelitian [2], yang mengusulkan model SVM untuk memprediksi mahasiswa yang berhenti dalam masa studinya di salah satu institusi pendidikan tinggi di Malaysia. Penelitian tersebut menggunakan beberapa faktor yang secara signifikan mempengaruhi permasalahan tersebut. Penelitian tersebut menggunakan beberapa atribut yaitu: jenis kelamin, usia, daerah asal, pekerjaan orang tua, pendapatan orang tua, kesehatan, keterlibatan sosial media, pilihan jurusan, status studi, dan pengetahuan dasar. Tahap awal yang dilakukan adalah praproses dan transformasi data, karena SVM hanya mendukung data dalam bentuk numerik dan biner sementara data yang digunakan adalah data teks. Setelah data ditransformasi, model SVM dapat digunakan untuk klasifikasi. Sementara di tahun 2015, terdapat penelitian lain [3] yang mengusulkan penggunaan Teknik data mining untuk prediksi kegagalan mahasiswa. Penelitian tersebut membandingkan algoritma Decision Tree dan Naïve Bayes dengan menggunakan 11 atribut yang bervariasi yaitu data pribadi, data terkait Pendidikan sebelumnya, dan data akademik mencakup nilai-nilai yang diperoleh selama kuliah. Hasil yang diperoleh menunjukkan bahwa Naïve Bayes memberikan akurasi yang lebih tinggi dalam melakukan prediksi dibandingkan Decision Tree.

Penelitian lain [4] juga menggunakan teknik *data mining* untuk memprediksi mahasiswa yang gagal atau *dropout* selama masa studi. Penelitian ini menggunakan dua dataset yaitu: dataset yang pertama dengan fitur demografis, fitur kognitif dan fitur non-kognitif. Fitur demografis seperti jenis kelamin, status pekerjaan, pendidikan ibu, sementara fitur kognitif seperti nilai-nilai yang diperoleh dari kuis, tugas, proyek, dll. Fitur non-kognitif seperti manajemen waktu, konsep diri, penilaian diri, kepemimpinan, nilai ujian, dukungan masyarakat. Dataset

yang kedua juga menggunakan tiga fitur, fitur demografis seperti jenis kelamin, status pernikahan orang tua dan wali. Fitur kognitif seperti ketidakhadiran, nilai ujian pertama dan kedua. Fitur non-kognitif seperti preferensi studi, waktu belajar, waktu luang, kemandirian, kedekatan dengan PT, kesehatan, waktu keluar, dukungan universitas, perencanaan studi di waktu yang akan datang. Tahap awal yang dilakukan adalah praproses dan transformasi data. Setelah itu, proses klasifikasi dilakukan dengan beberapa metode yaitu *Decision Tree* dengan algoritma J48, *Naïve Bayes*, *Logistic Regression* dan *Neural Network*. Hasil yang diperoleh, dataset pertama yang terdiri dari 113 data, *Decision Tree* menghasilkan akurasi sebesar 65% (menggunakan tiga fitur), *Logistic Regression* menghasilkan akurasi sebesar 54% (tanpa fitur non-kognitif), *Naïve Bayes* menghasilkan akurasi sebesar 61% (menggunakan tiga fitur), dan *Neural Network* menghasilkan 54% (tanpa fitur non-kognitif). Untuk dataset kedua, *Decision Tree* menghasilkan akurasi sebesar 84% (menggunakan tiga fitur), *Logistic Regression* menghasilkan akurasi sebesar 84% (tanpa fitur non-kognitif), *Naïve Bayes* menghasilkan akurasi sebesar 84% (menggunakan tiga fitur), dan *Neural Network* menghasilkan 82% (tanpa fitur non-kognitif).

Di tahun yang sama terdapat penelitian [5], yang juga mengembangkan sebuah model untuk melakukan prediksi terhadap mahasiswa yang gagal melewati tahun pertama pada jurusan kedokteran. Penelitian tersebut menggunakan 1819 data mahasiswa dan dengan analisis regresi logistik, prediksi kegagalan dalam kurikulum tahun pertama dibuat pada bulan ke 0, 4, 6, 8, 10, dan 12. Variabel yang digunakan adalah variabel sebelum diterima di universitas tersebut seperti variabel usia, jenis kelamin, indeks prestasi di bangku SMA/sederajat, cara seleksi dan variabel setelah diterima seperti jumlah SKS yang diperoleh, tingkat partisipasi dalam ujian, dan tingkat keberhasilan ujian. Mahasiswa yang lulus semua ujian pada bulan ke-4, ke-6 atau ke-8 memiliki peluang 99% lulus dari kurikulum pertama (dapat dikatakan optimal). Dalam kelompok non-optimal, pada bulan ke-6, kegagalan untuk lulus kurikulum tahun pertama dapat diprediksi dengan spesifisitas 66,7% dan sensitivitas 84,5% dengan menggunakan variabel "lulus 0 ujian pada bulan ke-4 dan ke-6".

Di tahun 2018, terdapat penelitian [6] yang juga melakukan prediksi *student failure* dengan memanfaatkan dataset dari Brunel University, London. Atribut yang diambil terkait informasi penerimaan mahasiswa, nilai modul tertentu yang dipilih, dan nilai akhir tahun pertama. Penelitian tersebut juga mengusulkan tiga metode yaitu J48, Decision Tree dan Naïve Bayes. Berdasarkan hasil penelitian, dapat disimpulkan bahwa Naïve Bayes memberikan hasil yang lebih baik daripada algoritma Decision Tree dalam memprediksi mahasiswa yang memiliki beresiko gagal berdasarkan modul. Akurasi untuk Naïve Bayes sebesar 88,48% sedangkan Decision Tree sebesar 84,29%. Kemudian di tahun 2019 terdapat penelitian [7] yang mengusulkan tiga metode *feature selection* yaitu *correlation Based*, *Information Gain Based*, dan *Learner Based/Wrapper* untuk mencari fitur yang paling berpengaruh untuk memprediksi lama studi mahasiswa. Dataset yang digunakan adalah data nilai mahasiswa STIKOM Bali program studi Sistem Informasi. Data yang diambil yaitu nilai matakuliah dari semester I sampai dengan semester VII yang terdiri dari 40 atribut dan 1246 baris. Penelitian tersebut menggunakan WEKA dan pengujian metode *feature selection* dilakukan dengan menggunakan metode Naïve Bayes. Teknik validasi yang digunakan adalah *10 folds cross-validation*. Hasil yang diperoleh menunjukkan bahwa Teknik *Wrapper* memberikan akurasi tertinggi yaitu 77,83% dan akurasi terendah diperoleh dengan Teknik *Information Gain* sebesar 73,65%. Tingkat kesalahan pada tiap-tiap metode dalam eksperimen cukup rendah dilihat dari angka *Mean Absolute Error* yang berkisar antara 0,1-0,2.

Berdasarkan tinjauan Pustaka yang dilakukan, diperoleh bahwa penelitian-penelitian sebelumnya cenderung menggunakan metode Naïve Bayes meskipun dengan atribut yang berbeda-beda. Kemudian, mengacu pada penelitian terbaru [7] yang menyebutkan bahwa Teknik *Wrapper* dapat memberikan hasil seleksi fitur yang cukup bagus sehingga akurasi juga lebih baik. Penelitian ini juga mengusulkan metode seleksi fitur yang berbeda yang diharapkan mampu memberikan akurasi yang lebih tinggi. Untuk klasifikasi, metode yang akan digunakan adalah *MultiLayer Perceptron* (MLP) yang merupakan turunan dari *Artificial Neural Network* (ANN). Metode ini telah digunakan untuk menyelesaikan banyak masalah terkait *Data Mining* dan terbukti mampu memberikan akurasi yang tinggi, tetapi belum diterapkan pada data mahasiswa untuk prediksi mahasiswa/i yang berpotensi *dropout*. Selain itu, pada penelitian-penelitian sebelumnya, akurasi yang diperoleh cukup tinggi tetapi atribut yang digunakan cukup banyak. Hal ini akan berpengaruh pada kecepatan proses *training*. Oleh karena itu, penelitian ini mengusulkan metode MLP dengan metode seleksi fitur yang dapat mencari atribut yang paling berpengaruh dalam proses klasifikasi. Tujuannya adalah agar memperoleh akurasi yang tinggi dan proses *training* dapat berlangsung dengan cepat.

III. LANDASAN TEORI

A. Data mining dan Knowledge Discovery

Data adalah kumpulan sesuatu dan *knowledge* (pengetahuan) adalah sesuatu yang dapat membantu manusia untuk membuat keputusan yang tepat. Ekstraksi pengetahuan dari sebuah data disebut sebagai *data mining*. *Data mining* juga dapat didefinisikan sebagai proses mengeksplorasi dan analisis sejumlah data yang besar dengan

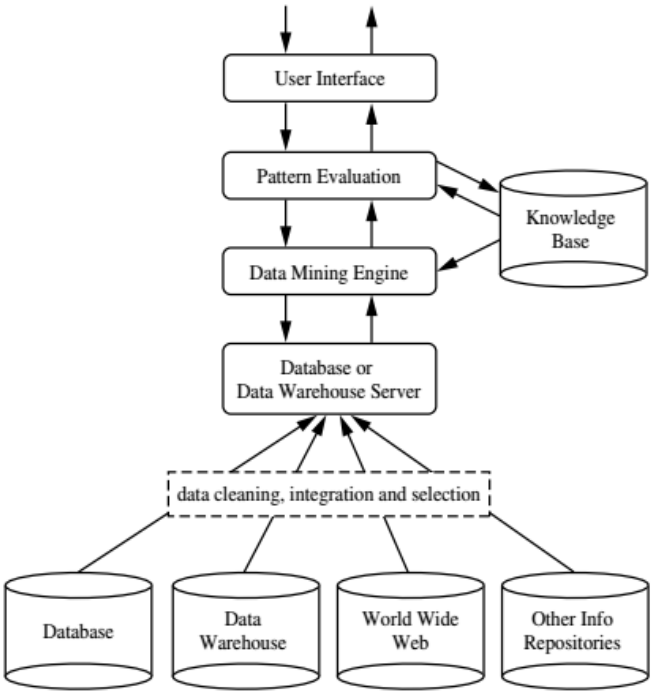
tujuan untuk menemukan pola dan aturan [8]. Secara sederhana, tujuan utama *data mining* adalah menemukan pengetahuan dan merupakan salah satu tahap dalam proses tersebut. Berdasarkan [9], *knowledge discovery* terdiri dari 7 tahap yaitu :

1. *Data cleaning*, proses untuk menghilangkan *noise* (derau) dan data yang tidak konsisten.
2. *Data integration*, proses penggabungan data, jika terdapat lebih dari satu sumber data.
3. *Data selection*, proses analisis data yang relevan yang diambil dari *database*.
4. *Data transformation*, proses transformasi atau penggabungan data ke bentuk yang sesuai untuk proses *mining*, misalnya dengan melakukan agregasi.
5. *Data mining*, proses utama dimana *intelligent methods* diterapkan untuk mengekstrak pola data.
6. *Pattern evaluation*, proses mengidentifikasi pola yang benar-benar dapat menggambarkan pengetahuan berdasarkan beberapa pengukuran.
7. *Knowledge presentation* teknik visualisasi dan representasi pengetahuan yang dilakukan menyajikan hasil pengetahuan yang diperoleh kepada pengguna.

Tahap 1 sampai 4 adalah merupakan tahap praproses data, dimana data dipersiapkan sebelum proses *mining*. Tahap data mining dapat berinteraksi dengan pengguna ataupun basis pengetahuan.

Menurut [9], *data mining* adalah proses menemukan pengetahuan yang penting dari sejumlah data yang besar yang disimpan dalam *database*, *data warehouse* (gudang data) atau repositori informasi lainnya *Database, Data Warehouse, World Wide Web*, atau repository lainnya, merupakan sumber data dimana proses data cleaning, integration dan selection dapat dilakukan. Gambaran arsitektur sistem data mining dapat dilihat pada Gambar 1. Berdasarkan pendapat tersebut maka arsitektur sistem data mining dapat memiliki komponen-komponen utama yaitu :

- *Database* atau *Data Warehouse server* bertanggung jawab untuk mengambil data yang relevan, berdasarkan permintaan pengguna
- *Knowledge Base*, domain pengetahuan yang digunakan untuk pencarian evaluasi ketertarikan pola yang dihasilkan.
- *Data Mining engine*, bagian penting dalam sistem *data mining* dan idealnya terdiri dari satu set modul fungsional seperti karakterisasi, asosiasi dan analisis korelasi, klasifikasi, prediksi, *cluster analysis*, analisis outlier, dan evolusi.
- *Pattern Evaluation Module*, biasanya digunakan untuk melakukan pengukuran dan berinterasksi dengan modul *data mining*.
- *User Interface,* modul ini berkomunikasi dengan pengguna dan system data mining, memungkinkan pengguna untuk berinteraksi dengan system dengan menentukan *query* terhadap *data mining*, memberikan informasi untuk membantu fokus pencarian, dan melakukan eksplorasi *data mining*.



Gambar 1. Arsitektur sistem *data mining*.

B. Multilayer Perceptron

Multilayer Perceptron adalah contoh dari ANN yang biasanya digunakan untuk memberikan solusi untuk masalah yang berbeda, misalnya untuk pengenalan pola dan interpolasi [10]. Diperlukan beberapa langkah untuk menjalankan klasifikasi ANN menggunakan arsitektur MLP, yaitu dimulai dengan pengumpulan data, kemudian membuat dan mengkonfigurasi jaringan. Selanjutnya menginisialisasi bobot dan bias. Setelah jaringan dapat melakukan pelatihan, validasi data dan digunakan selama klasifikasi. Kelemahan *perceptron* adalah tempat *perceptron* hanya dapat memecahkan masalah yang dapat dipisahkan secara linear [11].

Secara umum ANN adalah sebuah unit proses yang memiliki input dan mengeluarkan output, dengan neuron-neuron diorganisasikan sebagai layer. Output dari ANN dapat dikomputasi dengan rumus sebagai berikut

$$O = f(IW_{io}), \tag{1}$$

di mana W_{io} merupakan matriks beban (*weight matrix*) dengan ukuran *i* x *o*, dengan *i* jumlah input node*, o* jumlah *output node*, *input vector I,* dan *output vector O.*

Secara umum data dipresentasikan dalam input layer, kemudian *network* akan melakukan proses *input* dengan melakukan mengalikan *input* dengan layer *weight* (beban). Adapun untuk mempermudah dalam memahami cara kerja MLP, dapat digunakan algoritma yang telah disampaikan pada [10], sebagai berikut.

1. Melakukan inisialisasi network, dengan semua beban diset secara random antara angka -1 sampai dengan +1.
2. Mempresentasikan pola pelatihan pertama pada network yang ada, dan menyimpan hasil output.
3. Membandingkan output network tersebut dengan output target yang ada.
4. Memperbaiki eror secara *backward*.
   a. Perbaikan pada layer beban dari output:

   $$\omega_{h0} = \omega_{h0} + (\eta \delta_0 o_h), \tag{2}$$

   di mana ω_{ho} merupakan nilai beban dari unit *h* yang tersembunyi dengan output unit *o*, η merupakan rasio pelatihan, dan o_h merupakan keluaran dari unit *h* yang tersembunyi. Pada persamaan tersebut,

   $$\delta_0 = o_0(1-o_0)(t_0 - o_0). \tag{3}$$

   di mana O_0 merupakan node *o* dari output layer dan t_0 merupakan target output untuk node tersebut.
   b. Perbaikan pada beban input:

   $$\omega_{ih} = \omega_{ih} + (\eta \delta_h o_i), \tag{4}$$

   di mana ω_{ih} merupakan nilai beban dari unit *h* yang tersembunyi dengan input unit *i*, η merupakan rasio pelatihan, dan o_i merupakan *input dari node i*, dengan

   $$\delta_h = o_h(1-o_h)\sum_o(\delta_0 \omega_{h0}). \tag{5}$$

5. Melakukan perhitungan eror, dengan menghitung rata-rata dari nilai target dan output vector. Fungsi berikut dapat digunakan untuk menghitung eror tersebut.

   $$E = \frac{\sqrt{\sum_{n=1}^{p}(t_0 - o_0)^2}}{p}, \tag{6}$$

   di mana *P* merupakan jumlah unit pada *output layer.*
6. Mengulangi langkah 2 untuk setiap pola pada dataset pelatihan untuk melengkapi satu *epoch*.
7. Melakukan pertukaran dataset pelatihan secara random. Hal ini untuk mengurangi kemungkinan *network* dipengaruhi oleh urutan pada data.
8. Mengulangi langkah 2 untuk sejumlah *epochs* atau hingga *error* mulai berubah.

C. GainRatio Attribute Evaluation

Metode yang digunakan untuk reduksi data dibagi menjadi dua yaitu : Wrapper dan Filter. Model pendekatan Wrapper menggunakan metode klasifikasi itu sendiri untuk mengukur tingkat kepentingan dari sekumpulan fitur, selanjutnya fitur dipilih tergantung pada model pengklasifikasi yang digunakan. Pendekatan filter mendahului proses klasifikasi yang sebenarnya. Pendekatan ini bersifat independen dari algoritme pembelajaran, komputasi yang sederhana, cepat dan dapat terukur. Dengan menggunakan metode filter, proses seleksi fitur hanya dilakukan sekali dan kemudian dapat digunakan sebagai input untuk pengklasifikasi yang berbeda. Berbagai feature ranking dan seleksi fitur telah digunakan seperti *Correlation-based Feature Selection* (CFS), *Principal Component Analysis* (PCA), *Gain Ratio* (GR) *attribute evaluation*, *Chi-square feature evaluation*, *Fast Correlation-based Feature Selection* (FCBF), *Information gain*, *Euclidean distance*, *i-test*, *Markov blanket filter*. Beberapa dari metode filter ini tidak melakukan seleksi fitur tetapi hanya *feature ranking* dan oleh karena itu, metode tersebut dikombinasikan dengan metode pencarian untuk mengetahui jumlah atribut [12].

$$I(S) = -\sum_{i=1}^{m} p_i \, log_2(p_i), \quad (7)$$

di mana p_i adalah probabilitas sampel acak merupakan milik kelas C_i dan diperkirakan dengan s_i/s.

Atribut A memiliki nilai v yang berbeda dan s_{ij} merupakan jumlah sampel kelas C_i dalam S_j. S_j berisi *sample* S yang memiliki nilai a_j dari A. Entropi atau informasi yang diharapkan berdasarkan partisi menjadi himpunan bagian A, yang dirumuskan sebagai berikut

$$E(A) = -\sum_{i=1}^{m} I(S) \frac{s_{1i} + s_{2i} + \cdots s_{mi}}{s}. \quad (8)$$

Pengkodean informasi diperoleh dengan

$$Gain(A) = I(S) - E(A) \quad (9)$$

Gain ratio digunakan untuk *normalisasi information gain* menggunakan nilai yang dirumuskan sebagai berikut

$$SplitInfo_A(S) = -\sum_{i=1}^{v} (|S_i|I|S|) log_2(|S_i|I|S|) \quad (10)$$

Nilai pada persamaan (10) menunjukkan informasi dihasilkan dengan memisahkan data uji S menjadi bagian v sesuai dengan hasil v dari uji atribut A. Gain ratio dapat ditentukan dengan

$$Gain\ Ratio(A) = Gain(A)/SplitInfo_A(S) \quad (11)$$

Atribut yang memiliki *gain ratio* tertinggi dipilih sebagai *splitting attributes*.

D. Indeks Pengukuran

Pengukuran yang dilakukan dalam penelitian ini adalah untuk mengetahui tingkat kesuksesan terhadap proses-proses yang telah dilakukan. Pengukuran yang akan dilakukan yaitu pengukuran terhadap kinerja dari metode ekstraksi fitur maupun metode klasifikasi. Pengukuran dari proses klasifikasi ditentukan dengan nilai-nilai berikut:

1. Akurasi

Nilai akurasi dari hasil klasifikasi dapat diperoleh dengan menghitung jumlah klasifikasi yang benar dan sesuai target dibagi dengan jumlah klasifikasi yang berbeda dengan target dari semua kelas. Akurasi dirumuskan dengan

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}, \quad (12)$$

dengan TP (*True Positive*) adalah jumlah data benar pada target yang terklasifikasi benar pada sistem, TN (*True Negative*) adalah jumlah data salah pada target yang terklasifikasi salah pada sistem, FP (*False Positive*) merupakan representasi jumlah data salah pada target yang terklasifikasi benar pada sistem dan FN (*False Negative*) merupakan representasi jumlah data benar pada target yang terklasifikasi salah pada sistem. Nilai-nilai tersebut akan tampil dalam bentuk *confusion matrix*.

2. Sensitivitas

Sensitivitas merupakan ukuran kemampuan sistem untuk melakukan prediksi terhadap data yang dianggap benar sesuai dengan TPR (*True Positive Rate*). Sensitivitas dapat dirumuskan sebagai berikut [11].

$$sensitivity = \frac{TP}{TP + FN}. \quad (13)$$

3. Spesifisitas

Spesifisitas berkebalikan dengan sensitivitas yaitu kemampuan sistem untuk melakukan prediksi terhadap data yang dianggap salah sesuai dengan TNR (*True Negative Rate*). Spesifisitas dapat dirumuskan dengan

$$specificity = \frac{TN}{TN + FP}. \quad (14)$$

IV. METODE PENELITIAN

A. Teknik Pengumpulan dan Analisis Data

Penelitian ini memperoleh data dengan cara pengambilan langsung dari bagian sistem informasi STMIK AKAKOM yang berbentuk file excel. Selain itu, pengumpulan literatur juga dilakukan seperti studi pustaka yang mencakup buku-buku teks, makalah, artikel baik nasional maupun internasional.

Data yang digunakan adalah data mahasiswa angkatan 2008 – 2011 dari lima jurusan yaitu Teknik Informatika, Sistem Informasi, Manajemen Informatika, Komputerisasi Akuntansi, dan Teknik Komputer. Data mahasiswa yang diambil terdiri dari nim, nama, daerah asal, nilai SMA, dan Indeks Prestasi Semester (IPS). Untuk tahap awal, jumlah data yang digunakan sebanyak 1.265 data mahasiswa yang terdiri dari 265 data mahasiswa dengan

status K (Keluar) dan 1.000 data mahasiswa dengan status L (Lulus).

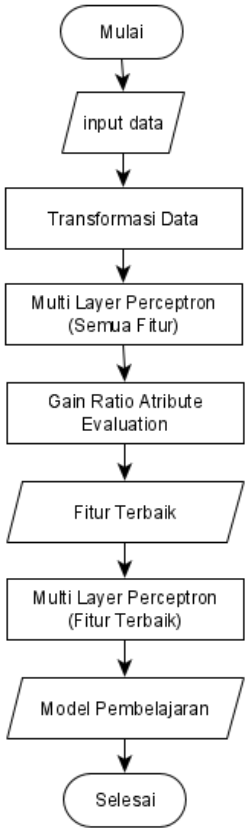
B. Rancangan Penelitian

Data-data yang diperoleh tersebut terdiri dari berbagai tipe data, baik yang berupa angka maupun berupa kalimat. Dalam data mining, data yang bisa diolah adalah data angka sehingga perlu adanya proses transformasi data untuk mengubah data kalimat menjadi angka agar dapat diproses di tahap selanjutnya. Oleh karena itu, tahap awal yang dilakukan adalah melakukan transformasi data non-angka menjadi data angka. Tahap selanjutnya adalah proses klasifikasi dengan menggunakan atribut data mahasiswa yaitu asal daerah, nilai SMA, IP Semester 1, Semester 2, Semester 3 dan Semester 4, untuk menguji kemampuan MLP dalam mengenali data yang statusnya K atau L. Berdasarkan hasil klasifikasi yang diperoleh, tingkat akurasi masih cukup rendah sehingga dilakukan seleksi fitur yang bertujuan untuk mencari fitur-fitur atau atribut yang memiliki korelasi paling tinggi dan mempengaruhi hasil klasifikasi. Setelah proses seleksi fitur, maka diperoleh atribut yang berpengaruh secara signifikan terhadap hasil klasifikasi. Atribut tersebut kemudian digunakan sebagai masukkan untuk proses klasifikasi ulang. Alur penelitian dapat dilihat pada Gambar 2.

Proses klasifikasi dilakukan menggunakan WEKA 3.6 dengan test mode *10-fold cross-validation* yang berarti data untuk *training* dan *testing* tidak akan dibedakan (digabungkan). Data akan dibagi menjadi 10 bagian secara acak, kemudian dilakukan 10 kali eksperimen dimana masing-masing eksperimen menggunakan 10 data sebagai data uji dan sisanya sebagai data latih.

V. HASIL PENELITIAN

Berdasarkan data yang diperolah, terdapat enam fitur/ciri yang diambil untuk tahap awal yaitu nilai ip semester 1-4, nilai SMA, dan asal daerah. Keenam fitur tersebut akan diklasifikasi menjadi dua kelas yaitu Lulus (L) dan Keluar (K). Tahap klasifikasi dilakukan dengan menggunakan *10-fold-cross validation*. Selanjutnya akan dilakukan evaluasi terhadap hasil klasifikasi. Tabel I menampilkan *confusion matrix* dari proses klasifikasi menggunakan Weka. TABEL I merupakan *confusion matrix* untuk klasifikasi dan diperoleh nilai *True Positive* (TP) = 140, *True Negative* (TN) = 952, *False Negative* (FN) = 125, dan *False Positive* (FP) = 48. Artinya dari 265 data mahasiswa yang Keluar, MLP mampu mengenali sebagai Keluar sebanyak 140 data sedangkan 125 data dikenali sebagai kelompok mahasiswa yang Lulus. Selanjutnya dari 1.000 data mahasiswa yang berhasil Lulus, MLP mampu mengenali sebagai kelompok Lulus sebanyak 952 data sedangkan 48 data dikenali sebagai kelompok yang Keluar. Berdasarkan *confusion matrix* maka dapat dihitung tingkat akurasi, sensitivitas dan spesifisitas. Untuk tahap klasifikasi pertama, diperoleh akurasi 86,3% dengan sensitivitas 52,8% dan spesifisitas 95,2%.



Gambar 2. *Flowchart* penelitian.

Jika dilihat, nilai sensitivitas yang diperoleh cukup rendah. Untuk meningkatkan tingkat sensitivitas, proses selanjutnya yang dilakukan adalah seleksi fitur untuk mencari fitur yang berpengaruh secara signifikan terhadap hasil klasifikasi. Seleksi fitur menggunakan metode *Gain Ratio Attribute Evaluation* dan diperoleh hasil seperti pada TABEL II.

Berdasarkan hasil seleksi fitur, diperoleh hasil yang menunjukkan bahwa atribut IP Semester 4 yang memiliki dampak yang paling signifikan dalam menentukan tingkat akurasi. Setelah selesai seleksi fitur, maka proses klasifikasi dilakukan ulang dengan menggunakan atribut IP Semester 4 dan diperoleh *confusion matrix* seperti pada TABEL III yang merupakan *confusion matrix* untuk klasifikasi dan diperoleh nilai *True Positive* (TP)=145, *True Negative* (TN)=954, *False Negative* (FN)=120 dan *False Positive* (FP)=46. Artinya dari 265 data mahasiswa yang Keluar, MLP mampu mengenali sebagai Keluar sebanyak 145 data sedangkan 120 data dikenali sebagai kelompok mahasiswa yang Lulus. Selanjutnya dari 1.000 data mahasiswa yang berhasil Lulus, MLP mampu mengenali sebagai kelompok Lulus sebanyak 954 data sedangkan 46 data dikenali sebagai kelompok yang Keluar. Berdasarkan *confusion matrix* pada TABEL III tersebut maka dapat dihitung tingkat akurasi, sensitivitas dan spesifisitas. Untuk tahap klasifikasi pertama, diperoleh akurasi 86,9% dengan sensitivitas 54,7% dan spesifisitas 95,4%. Jika dibandingkan, untuk akurasi hanya meningkat 0,6% sedangkan sensitivitas meningkat 1,9% dan spesifisitas 0,2%. Hal ini mungkin disebabkan karena perbandingan jumlah data mahasiswa yang Keluar jauh lebih sedikit dibanding data mahasiswa yang Lulus.

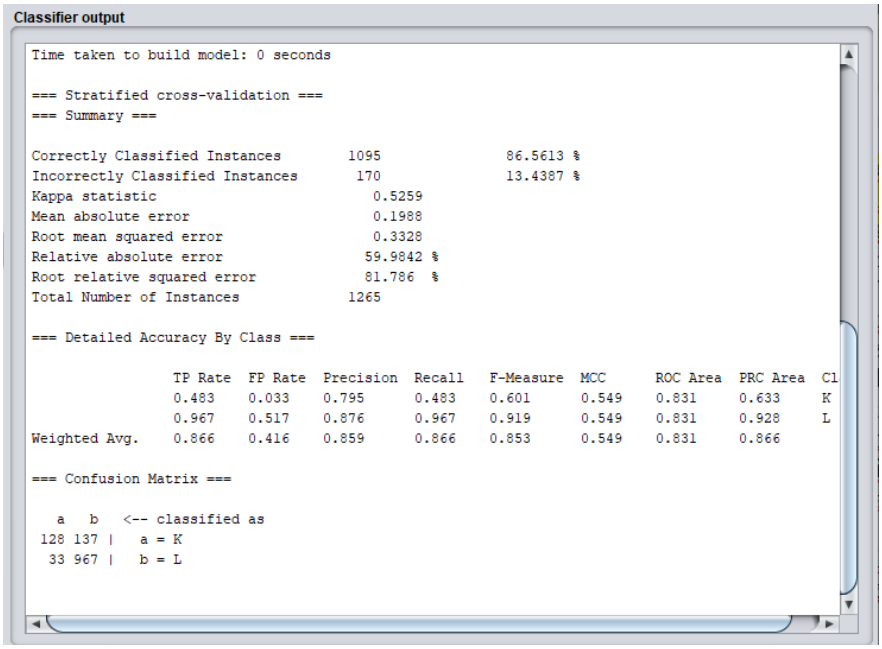
Atribut yang digunakan sebagai masukkan kemudian digunakan untuk klasifikasi menggunakan metode lain yaitu *Naïve Bayes*. Hasil klasifikasi menggunakan *Naïve Bayes* pada Gambar 3 menunjukkan akurasi sebesar 86,56%. Atribut yang sama juga digunakan sebagai masukkan dalam *Decision Tree* dengan algoritma J48 yang ditunjukkan pada Gambar 4. Dapat dilihat bahwa tidak banyak perbedaan akurasi dengan metode MLP.

TABEL I
CONFUSION MATRIX.

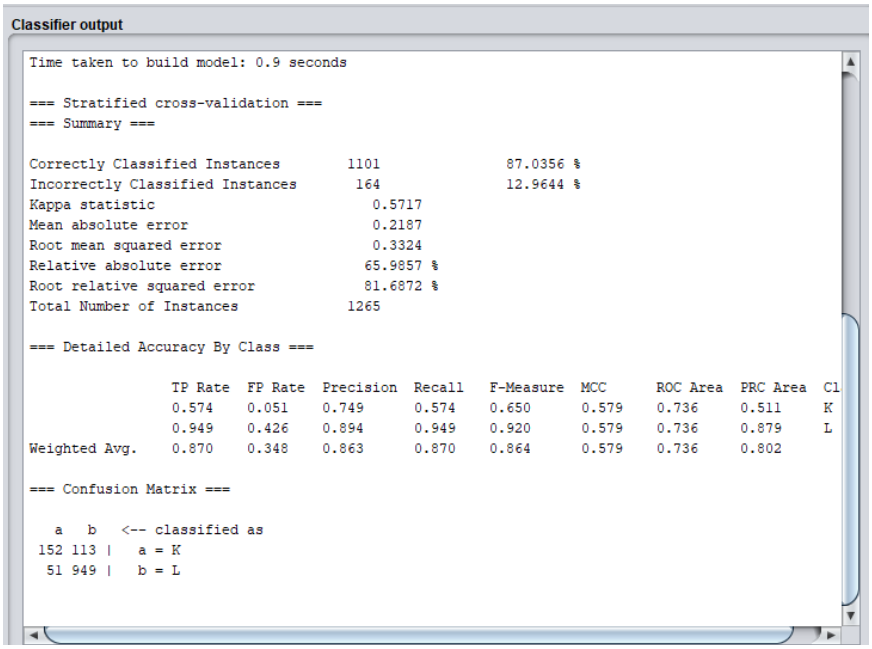
		Target	
		Keluar	**Lulus**
Prediksi	**Keluar**	140	125
	Lulus	48	952

TABEL II
HASIL SELEKSI FITUR.

Rank	Attributes
0.153	Sem4
0.1227	Sem3
0.1225	Sem1
0.1032	Sem2
0.0152	NILAI SMA
0	PROP



Gambar 3. Hasil klasifikasi menggunakan *Naïve Bayes*.



Gambar 4. Hasil klasifikasi menggunakan J48.

TABEL III
CONFUSION MATRIX.

		Target	
		Keluar	**Lulus**
Prediksi	**Keluar**	145	120
	Lulus	46	954

VI. KESIMPULAN

Penelitian ini menghasilkan sebuah *rule model* yang dapat digunakan untuk melakukan prediksi kegagalan studi mahasiswa. Berdasarkan tahap-tahap yang telah dilakukan maka dapat disimpulkan bahwa metode MultiLayer Perceptron juga dapat digunakan untuk memprediksi kegagalan studi mahasiswa. Dengan adanya prediksi terhadap kegagalan studi mahasiswa, diharapkan penelitian ini dapat membantu Perguruan Tinggi untuk mengetahui mahasiswa/i yang berpotensi akan di*dropout*. Dengan demikian Perguruan Tinggi dapat menentukan strategi dan langkah yang harus dilakukan agar dapat meminimalisir jumlah mahasiswa yang *dropout*.

Tahap awal yang dilakukan yaitu pengumpulan data, transformasi data, dan klasifikasi, kemudian seleksi fitur untuk mendapat fitur yang paling berpengaruh dalam proses klasifikasi. Berdasarkan seleksi fitur ini, dapat dilihat bahwa IP Semester 4 yang paling berpengaruh. Oleh karena itu proses klasifikasi dilakukan ulang dan akurasi yang dicapai cukup baik yaitu 86,9% dengan spesifisitas yang cukup tinggi yaitu 95,4%. Meskipun demikian, nilai sensitivitas masih rendah. Hal ini mungkin disebabkan karena jumlah data mahasiswa yang Keluar memang lebih sedikit dibanding mahasiswa yang berhasil lulus. Oleh karena itu, untuk penelitian selanjutnya perlu menambahkan lebih banyak data dan mencari atribut yang lebih signifikan dalam mempengaruhi hasil klasifikasi sehingga diharapkan dapat menghasilkan akurasi, sensitivitas dan spesifisitas yang lebih tinggi.

Atribut dari hasil seleksi fitur yaitu IP Semester 4 kemudian digunakan sebagai masukkan dalam metode *Naïve Bayes* dan J48. Hasil yang diperoleh menunjukkan selisih tingkat akurasi yang sangat sedikit bahkan algoritma J48 memberikan akurasi yang lebih tinggi dibandingkan dengan MLP. Hal ini menunjukkan bahwa metode *Gain Ratio Attribute Evaluation* mampu mencari atribut yang paling berpengaruh dalam proses klasifikasi dan atribut tersebut juga dapat memberikan tingkat akurasi yang cukup stabil ketika diuji menggunakan metode yang berbeda.

UCAPAN TERIMA KASIH

Ucapan terima kasih diberikan kepada Kementerian Riset dan Teknologi – Badan Riset dan Inovasi Nasional yang telah menyediakan dana untuk kelancaran penelitian ini. Terima kasih juga diberikan kepada STMIK AKAKOM Yogyakarta yang telah bersedia memberikan data untuk penelitian ini.

DAFTAR PUSTAKA

[1] J. P. Jiawei Han dan Micheline Kamber, *Data Mining – Concepts & Techniques*. 2011.
[2] A. Sangodiah dan B. Balakrishnan, "Holistic Prediction of Student Attrition in Higher Learning Institutions in Malaysia Using Support Vector Machine Model," *Int. J. Res. Stud. Comput. Sci. Eng.*, vol. 1, no. 1, hal. 29–35, 2014.
[3] L. P. Khobragade dan P. P. Mahadik, "Students ' Academic Failure Prediction Using Data Mining," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 11, hal. 290–298, 2015.
[4] S. Sultana, S. Khan, dan M. A. Abbas, "Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts," *Int. J. Electr. Eng. Educ.*, vol. 54, no. 2, hal. 105–118, 2017.
[5] G. J. A. Baars, T. Stijnen, dan T. A. W. Splinter, "A Model to Predict Student Failure in the First Year of the Undergraduate Medical Curriculum," *Heal. Prof. Educ.*, vol. 3, no. 1, hal. 5–14, 2017.
[6] M. Al, A. Tucker, dan L. Yousefi, "The Prediction of Student Failure using Classification Methods : A Case Study," dalam *Proc. Int. Conf. Image Process. Pattern Recognit.*, hal. 79–90, 2018.
[7] I. made B. Adnyana, "Penerapan Feature Selection untuk Prediksi Lama Studi Mahasiswa," *J. Sist. Dan Inform.*, vol. 13, hal. 72–76, 2019.
[8] M. Negnevitsky, *Artificial intelligence*, vol. VII. 2001.
[9] J. Han dan M. Kamber, *Data Mining: Concepts and Techniques*, vol. 12. 2011.
[10] L. Noriega, "Multilayer Perceptron Tutorial." 2005.
[11] M. M. Sebatubun dan M. A. Nugroho, "Ekstraksi Fitur Circularity untuk Pengenalan Varietas Kopi Arabika," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 4, no. 4, hal. 283–289, 2017.
[12] A. G. Karegowda, A. S. Manjunath, dan M. A. Jayaram, "Comparative Study of Attribute Selection using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, hal. 271–277, 2010.