TESIS

PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN ALGORITMA C4.5 DAN K-MEANS

Disusun oleh:

Nama : **Arif Budiman**
NIM : **17.51.1054**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2019

TESIS

PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN ALGORITMA C4.5 DAN K-MEANS

STUDENT COMPLETION PREDICTION USING C4.5 ALGORITHM AND K-MEANS

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Arif Budiman**
NIM : **17.51.1054**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2019

HALAMAN PENGESAHAN

PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN ALGORITMA C4.5 DAN K-MEANS

STUDENT COMPLETION PREDICTION USING C4.5 ALGORITHM AND K-MEANS

Dipersiapkan dan Disusun oleh

Arif Budiman

17.51.1054

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Magister Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari selasa, 3 Desember 2019

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 3 Desember 2019
Ketua Universitas AMIKOM Yogyakarta

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN ALGORITMA C4.5 DAN K-MEANS

STUDENT COMPLETION PREDICTION USING C4.5 ALGORITHM AND K-MEANS

Dipersiapkan dan Disusun oleh

Arif Budiman

17.51.1054

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Magister Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari selasa, 3 Desember 2019

Pembimbing Utama

Dr. Arief Setyanto, S.Si., M.T.
NIK. 190302036

Anggota Tim Penguji

Dr. Arief Setyanto, S.Si., M.T.
NIK. 190302036

Dr. Kusrini, M.Kom.
NIK. 190302106

Pembimbing Pendamping

Ferry Wahyu Wibowo, S.Si., M.Cs.
NIK. 190302235

Dr. Andi Sunyoto, M.Kom.
NIK. 190302052

Tesis ini telah diterima sebagai salah satu persyaratan
Untuk memperoleh gelar Magister Komputer
Yogyakarta, 3 Desember 2019
Direktur Program Pascasarjana

Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Arif Budiman
NIM : 17.51.1054
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 Dan K-Means

Dosen Pembimbing Utama : Dr. Arief Setyanto, S.Si., M.T.
Dosen Pembimbing Pendamping : Ferry Wahyu Wibowo, S.Si., M.Cs.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 3 Desember 2019
Yang Menyatakan,

Arif Budiman

HALAMAN PERSEMBAHAN

Segala puji bagi Allah atas segala kekuatan, nikmat, dan karunia yang telah diberikan sehingga penulis dapat menyelesaikan tesis yang berjudul "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 dan K-Means". Tidak lupa shalawat serta salam untuk Baginda Rasullullah Muhammad Shalallahu'alahi wassalam yang telah memberikan teladan sebaik baiknya teladan.

Karya tulis ini dengan bangga penulis persembahkan untuk :

1. Kedua orang tua saya, Bapak Sarmadi dan Ibu Noryani yang telah memberikan segala doa, dukungan, serta semangat yang tak terhingga kepada penulis.
2. Kekasih saya tercinta Wulan yang telah memberikan segala doa, dukungan, serta semangat dan motivasi sehingga penulis mampu menyelesaikan tesis
3. Adik-adik saya tercinta Tia, Nisa, dan Ayu
4. Kedua pembimbing saya, Bapak Arief Setyanto dan Bapak Ferry Wahyu Wibowo yang telah membimbing penulis dalam penyusunan tesis.

HALAMAN MOTTO

"Talk more do more"

" Hidup itu seperti mengendarai sepeda. Untuk menjaga keseimbangan, kamu harus terus berjalan maju" – Albert Einstein

"Jangan terlalu khawatir terhadap segala sesuatu yang mengganggumu, dan lihatlah sisi positif yang bisa dilakukan"

"Banyak kegagalan hidup terjadi karena orang-orang tidak menyadari Betapa dekatnya kesuksesan ketika mereka menyerah" – Thomas alfa edison

KATA PENGANTAR

Segala puji bagi Allah yang telah memberikan rahmat, kekuatan, kesempatan, sehingga penulis dapat menyelesaikan tesis dengan judul "Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 dan K-Means".

Penyusunan tesis ini untuk memenuhi persyaratan kelulusan Program Magister di Universitas AMIKOM Yogyakarta pada Jurusan Teknik Informatika.

Dalam penyusunan tesis ini tidak lepas dari dukungan berbagai pihak, oleh karena itu pada kesempatan ini penulis mengucapkan terimakasih kepada :

1. Bapak Prof. Dr. M. Suyanto, MM selaku Rektor Universitas "AMIKOM" Yogyakarta.
2. Ibu Dr. Kusrini, M.kom selaku direktur program pascasarjana Universitas "AMIKOM" Yogyakarta.
3. Bapak Dr. Arief Setyanto, S.Si., M.T. selaku dosen pembimbing pertama yang telah meluangkan waktu untuk membimbing, mengarahkan, dan memberikan ilmu saat penyusunan tesis.
4. Bapak Ferry Wahyu Wibowo, S.Si., M.Cs. selaku dosen pembimbing kedua yang telah meluangkan waktu untuk membimbing, mengarahkan, dan memberikan ilmu saat penyusunan tesis.
5. Kedua orang tua penulis, untuk doa dan dukungan yang terus mengalir.
6. Bapak Ibu dosen dan pegawai Universitas "AMIKOM" Yogyakarta.
7. Semua pihak telah membantu dalam proses penyusunan tesis yang terlibat langsung maupun tidak yang tidak dapat penulis sebutkan satu persatu.

Yogyakarta, 3 Desember 2019

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Mahasiswa merupakan aset penting yang dimiliki oleh institusi penyelenggara pendidikan. lama masa studi mahasiswa untuk jenjang sarjana semestinya dapat diselesaikan dalam 4 tahun namun kenyatannya masih banyak mahasiswa yang tidak dapat meyelesaikan kuliahnya selama 4 tahun karena berbagai macam faktor antara lain IPK sem 1, IPK sem 2, IPK sem 3, IPK sem 4, Total Sks, Presensi sem 1, Presensi sem 2, Presensi sem 3, Presensi sem 4, Mengulang, Remedi, Asal, JenisKelamin, Nilai TOEFL dan SLTA. Sehingga hal itu mempengaruhi indikator keberhasilan dalam proses penyelenggaraan pendidikan di perguruan tinggi. Oleh karena itu perlu adanya pemantauan selama 4 semester berjalan untuk mengetahui ketepatan kelulusan mahasiswa tersebut.

Penelitian ini menggunakan K-Means untuk membagi data atribut mahasiswa berdasarkan jumlah k terbaik dan algoritma C4.5 untuk menghasilkan pengetahuan untuk prediksi kelulusan mahasiswa. model yang dihasilkan dievaluasi menggunakan Confusin matrix dan Roc Curve.

Penelitian ini menggunakan algoritma C4.5 dan K-means menghasilkan akurasi 88,56% dan nilai AUC 0,823 yang termasuk kedalam Good Classification

Kata kunci: Kelulusan Mahasiswa, Algoritma C4.5, K-Means

ABSTRACT

Students are important assets owned by collage. the length of study for undergraduate students should be completed in 4 years but in reality there are still many students who cannot complete their studies for 4 years due to various factors such as Ipk sem 1, Ipks sem 2, Ipk sem 3, Ipk sem 4, Total Sks, Presensi sem 1, Presensi sem 2, Presensi sem 3, Presensi sem 4, Mengulang, Remedi Asal, JenisKelamin, Nilai TOEFL dan SLTA. So that it affects the indicators of success on University. Therefore there is a need for monitoring for 4 semesters in order to determine the accuracy of the graduation of these students.

This research using K-Means to divide student attribute data based on the best number of k and C4.5 algorithm to obtain knowledge for predicting student graduation. the result of this model is evaluated using Confusion matrix and Roc Curve.

This research using C4.5 algorithm and K-Means obtain accuracy 88.56% and AUC 0.823 classified as Good Classification.

Keywords: Student Completion, C4.5 Algorithm, K-Means

BAB I

PENDAHULUAN

1.1. Latar Belakang

Mahasiswa merupakan aset penting yang dimiliki oleh institusi penyelenggara pendidikan. Berdasarkan peraturan akademik untuk jenjang sarjana lama masa studi mahasiswa semestinya dapat diselesaikan dalam 4 tahun namun kenyatannya masih banyak mahasiswa yang tidak dapat meyelesaikan kuliahnya selama 4 tahun karena berbagai macam factor. Sehingga hal itu mempengaruhi indikator keberhasilan dalam proses penyelenggaraan pendidikan di perguruan tinggi. Oleh karena itu perlu adanya pemantauan serta evaluasi terhadap ketepatan dalam kelulusan mahasiswa.

Data Mining adalah teknik yang memanfaatkan data dalam jumlah yang besar yang tersedia didalam *database* untuk menguraikan penemuan pengetahuan (Turban, Aronson, & Liang, 2005). Terdapat banyak metode dalam *data mining* diantaranya seperti *Decision Trees, Bayesian, Artificial Neural Networks, Nearest Neighbor, Support Vector Machines* dan lainnya (Suyanto, 2017).

Penelitian (Purwanto & Darmadi, 2018) menyatakan bahwa Algoritma C4.5 memiliki nilai *precision* dan *accuracy* yang paling baik dibandingkan algoritma *K-Nearest Neighbor, Naïve Bayes, Rule Induction* dan *Deep Learning*. Namun algoritma konvensional C4.5 masih memiliki nilai akurasi yang lebih rendah dibandingkan SVM(Daud, Aljohani, & Abbasi, 2017). Oleh karena itu ini dapat menjadi peluang untuk meningkatkan akurasi algoritma C4.5 agar menjadi

lebih baik. Metode yang digunakan untuk meningkatkan akurasi algorima klasifikasi adalah K-Means(Umam, Zulfahmi, & Nababan, 2017)(Khan & Mohamudally, 2016).

Berdasarkan uraian latar belakang masalah diatas maka penulis bermaksud melakukan penelitian tentang prediksi kelulusan mahasiswa menggunakan Algorima C4.5 dan K-Means.

1.2. Rumusan Masalah

Berdasarkan latar belakang masalah yang telah diuraikan sebelumnya masalah yang dirumuskan dalam penelitian ini yaitu :

1. Apa faktor yang paling mempengaruhi kelulusan mahasiswa dengan menggunakan algoritma C4.5 dan K-Means?
2. Berapa akurasi prediksi kelulusan mahasiswa dengan menggunakan algoritma C4.5 dan K-Means?
3. Berapa nilai AUC prediksi kelulusan mahasiswa dengan menggunakan algoritma C4.5 dan K-Means?

1.3. Batasan Masalah

Batasan masalah dalam penelitian ini yaitu :

1. Sumber data dalam penelitian ini adalah data mahasiswa jurusan informatika tahun 2011, 2012, 2013, 2014.
2. Metode penambangan data yang digunakan dalam penelitian ini adalah algoritma C4.5.
3. Metode seleksi atribut yang digunakan dalam penelitian ini adalah Koefisien Kontingensi.
4. Metode pengelompokan atribut yang digunakan dalam penelitian ini adalah *K means*.
5. Kandidat atribut yang akan digunakan dalam memprediksi kelulusan mahasiswa antara lain NPM, Jenis Kelamin, Asal, Angkatan, Tahun Lulus, SKS Total, IPK sem 1, IPK sem 2, IPK sem 3, IPK sem 4, IPK sem 5, IPK sem 6, IPK sem 7, IPK sem 8, Presensi sem 1, Presensi sem 2, Presensi sem 3, Presensi sem 4, Konsentrasi, Nilai UN, Nilai Toefl, SLTA, Makul Mengulang dan Makul Remedi, Pekerjaan Ayah, Pekerjaan Ibu, Penghasilan Ayah, Penghasilan Ibu, Beasiswa Bidikmisi.
6. Sumber Data yang digunakan adalah data mahasiswa S1 jalur reguler bukan mengulang atau transfer.
7. Hasil klasifikasi dari prediksi adalah mahasiswa lulus yang Tepat Waktu dan Terlambat.
8. Tool yang digunakan adalah Microsoft excel 2010 dan Rapidminer versi 9.
9. Metode Pengujian model menggunakan *Confusion matrix* dan *ROC Curve*.

1.4. Tujuan Penelitian

Tujuan yang ingin dicapai dalam penelitian ini yaitu :

1. Mengetahui akurasi algoritma algoritma C4.5 dan K-Means dalam memprediksi kelulusan mahasiswa Universitas AMIKOM Yogyakarta.
2. Mengetahui faktor-faktor yang mempengaruhi kelulusan mahasiswa di Universitas Amikom Yogyakarta.
3. Mengetahui parameter terbaik untuk memprediksi kelulusan mahasiswa dengan menggunakan algoritma C4.5 dan K-Means.

1.5. Manfaat Penelitian

Manfaat yang ingin dicapai dalam penelitian ini yaitu :

1. Dapat membantu pihak yang terkait dengan lembaga kemahasiswaan dalam mengambil keputusan.
2. Dapat membantu dalam membuat kebijakan di kampus agar mahasiswa lulus tepat waktu sehingga visi dan misi Universitas AMIKOM dapat tercapai.
3. Dapat menerapkan ilmu yang didapat selama kuliah di MTI Universitas Amikom Yogyakarta.

1.6. Keaslian Penelitian

Tabel 1.1 Matriks literatur review dan posisi penelitian

Prediksi kelulusan mahasiswa menggunakan Algoritma C4.5 dan K-Means

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Classification of Province Based on Dropout Rate.	Pertiwi, A. G., Widyaningtyas, T., & Pujianto, U, International Conference on Sustainable Information Engineering and Technology (SIET), 2017	Mengklasifikasikan tingkat *drop out* berdasarkan provinsi di indonesia menggunakan algoritma C4.5	Dalam penelitian menggunakan 128 record yang terdiri dari 7 atribut antara lain jumlah penduduk miskin, jumlah pembiayaan daerah, jumlah pendapatan daerah, jumlah belanja daerah, tpak, rasio gini, dan tingkat putus sd. Pengujian yang dilakukan menggunakan *confussion matrix* dan menghasilkan tingkat akurasi sebesar 71.2%.	Penelitian hanya menggunakan 1 metode, akurasi dapat dikembangkan dengan menambah jumlah data, atribut, dan menggunakan algoritma optimasi	Penelitian yang dilakukan menggunakan Algoritma C4.5 dan K-Means untuk memprediksi kelulusan mahasiswa, dengan mengelompokan data mahasiswa menjadi 3 klaster menggunakan K-Means dan seleksi atribut menggunakan Koefisien Kontingensi
2	Student Performance Analysis Using Clustering Algorithm	Singh, I., Sabitha, A. S. A. I., Bansal, A., & Cse, A. 6th International Conference - Cloud System and Big	Memprediksi kinerja mahasiswa menggunakan K-means	Dalam penelitian menggunakan atribut antara lain Xth, XIIth, B.Tech marks, projects, internships, and skills set. kemampuan mahasiswa dan dibagi menjadi 3 cluster. dari	Dapat dilakukan perangkingan obyek dalam klaster dan menambahkan algorima	Penelitian yang dilakukan menggunakan Algoritma C4.5 dan K-Means untuk memprediksi kelulusan mahasiswa, dengan

Tabel 1.1 (Lanjutan)

No	Judul	Pencliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		Data Engineering (Confluence), 2016		hasil pengujian diperoleh mahasiswa yang memiliki kinerja terendah pada cluster 0	klasifikasi untuk peningkatan analisis kinerja mahasiswa	mengelompokan data mahasiswa menjadi 3 klaster menggunakan K-Means dan seleksi atribut menggunakan Koefisien Kontingensi
3	Predicting Student Performance using Advanced Learning Analytics	Daud, A., Aljohani, N. R., & Abbasi, R. A, International World Wide Web Conference Committee (IW3C2), 2017	Memprediksi kinerja mahasiswa menggunakan beberapa algoritma klasifikasi yang berbeda	Data yang digunakan diperoleh dari beberapa universitas yang berbeda. Hasil pengujian diketahui bahwa SVM memiliki akurasi yang paling tinggi dibandingkan C4.5, CART, Bayes network dan Naive Bayes.	Dalam penelitian hanya menggunakan algoritma klasifikasi konvensional sehingga dapat menggunakan algoritma optimasi untuk meningkatkan kinerja model	Penelitian yang dilakukan menggunakan Algoritma C4.5 dan K-Means untuk memprediksi kelulusan mahasiswa, dengan mengelompokan data mahasiswa menjadi 3 klaster menggunakan K-Means dan seleksi atribut menggunakan Koefisien Kontingensi
4	Perbandingan Kinerja Algoritma C4.5 Dan Naive	Supriyanti, W., Kusrini, & Amborowati, A, Jurnal INFORMA	Memprediksi ketepatan pemilihan konsentrasi mahasiswa menggunakan	Dalam penilinitian diperoleh tingkat akurasi C4.5 sebesar 84,43% dan *Naive bayes* sebesar 78,47% . namun setelah	Menggunakan algoritma klasifikasi yang menggunakan	Penelitian yang dilakukan menggunakan Algoritma C4.5 dan K-Means untuk

Tabel 1.1 (Lanjutan)

No	Judul	Penciliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa	Politeknik Indonusa Surakarta Vol. 1 Nomor 3, 2016	algoritma C4.5 dan *Naive bayes* yang dikombinasikan dengan metode *Forward selection*	penambahan metode untuk seleksi fitur menunjukan peningkatan akurasi C4.5 sebesar 0.40% dan 3.54%	metode optimasi selain foward selection untuk memperoleh akurasi yang lebih baik	memprediksi kelulusan mahasiswa, dengan mengelompokan data mahasiswa menjadi 3 klaster menggunakan K-Means dan seleksi atribut menggunakan Koefisien Kontingensi
5	An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm.	Khan, D. M., & Mohamudally, N, Journal Of Computing, Volume 3, Issue 12, (March), 2016	Memprediksi penderita kanker payudara dengan kombinasi K-means dengana lgoritma ID3	Dalam penelitian rules yang lebih sedikit yang berdampak pada waktu eksukusi yang lebih singkat sehingga model yang dihasilkan efisien dalam melakukan ekstrasi rule menggunakan ID3 dan K-Means	Menggunakan dataset yang berbeda dan mengkombinasikan algoritma klasifikasi untuk menghasilkan model yang lebih baik	Penelitian yang dilakukan menggunakan Algoritma C4.5 dan K-Means untuk memprediksi kelulusan mahasiswa, dengan mengelompokan data mahasiswa menjadi 3 klaster menggunakan K-Means dan seleksi atribut menggunakan Koefisien Kontingensi

Tabel 1.1 (Lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
6	Accuracy Improvement of C4.5 using K-Means Clustering	Rajeshinigo, D., & Jebamalar, J. P. A, International Journal of Science and Research (IJSR), 2017	Meningkatkan akurasi algoritma C4.5 dengan melakukan pengelompokan data interval menjadi 2 klaster menggunakan K-means	Dalam penelitian menggunakan data mahasiswa yang terdiri dari 8 atribut antara lain IAT, AS(Attendance Status), CA(Current Arrears), HTL(Hostel), PC(Project Completion), AC(Assignment Completion), PW(Parents Work) dan AD(Academic Detention). Hasil Penggabungan C4.5 menggunakan K-means menghasilkan akurasi 92%	Perlu adanya pengujian data dengan jumlah data dan atribut yang lebih banyak. Selain akurasi perlu juga menghitung nilai AUC untuk evaluasi model klasifikasi	Penelitian yang dilakukan menggunakan Algoritma C4.5 dan K-Means untuk memprediksi kelulusan mahasiswa, dengan mengelompokan data mahasiswa menjadi 3 klaster menggunakan K-Means dan seleksi atribut menggunakan Koefisien Kontingensi

1.7. Hipotesis

Berdasarkan perumusan masalah yang telah diungkapkan sebelumnya, maka kesimpulan sementara (hipotesa) yang dapat ditarik pada penelitian prediksi kelulusan mahasiswa dengan menggunakan algoritma C4.5 dan K-Means adalah metode C4.5 dan K-Means memiliki akurasi yang baik.

1.8. Metode Penelitian

1.8.1. Jenis, Sifat dan Pendekatan Penelitian

Penelitian ini menggunakan jenis penelitian *eksperimen*, dengan tahapan penelitian sebagai berikut :

1. Pengumpulan Data

   Pengumpulan data merupakan langkah awal pada suatu penelitian. Data yang digunakan pada penelitian ini adalah data kelulusan mahasiswa informatika

2. Pengolahan Data Awal

   Pengolahan awal (*Preprocessing*) merupakan tahap untuk mempersiapkan data yang telah diperoleh dari tahap pengumpulan data, yang akan digunakan pada tahap selanjutnya.

3. Desain Eksperimen

   Tahapan ini akan membahas desain eksperimen yang digunakan pada penelitian

4. Eksperimen dan Pengujian

   Tahapan ini akan membahas tahapan penelitian dan teknik pengujian yang akan digunakan.

5. Evaluasi Penelitian

   Tahapan ini akan membahas hasil evaluasi dari eksperimen yang telah digunakan.

1.8.2. Metode Pengumpulan Data

Metode pengumpulan data yang dilakukan pada penelitian ini yaitu :

1. Observasi

   Observasi dilakukan dengan melakukan pengamatan secara langsung mengenai kondisi jurusan informatika Universitas Amikom Yogyakarta.

2. Dokumentasi

   Pengambilan data melalui dokumen tertulis maupun elektronik dari lembaga atau institusi sesuai dengan yang diangkat dalam tesis ini yaitu di Universitas Amikom Yogyakarta. Dokumen diperlukan untuk mendukung keperluan data yang lain.

3. Studi pustaka

   Mengetahui informasi secara teori mengenai pokok permasalahan dan teori pendukung yang digunakan sebagai dasar pemikiran dalam membahas permasalahan yang ada, membaca referensi yang berhubungan dengan metode Algoritma C4.5 dan K-Means.

1.8.3. Metode Analisis Data

Metode untuk menganalisis data dalam penerapan data mining ini adalah *Knowledge Discovery in Database* (KDD) yang terdiri dari beberapa tahapan yaitu *Cleaning and Integration, Selection and Transformation, Data mining,* dan *Evaluation and Interpretation* (Han, Kamber, & Pei, 2012) :

1. *Cleaning and Integration*

   Tahap ini merupakan proses pembersihan terhadap data yang dilakukan untuk memastikan data yang diperoleh sebelumnya dapat digunakan serta bebas dari duplikasi, kesalahan dan *validation rules* sudah sesuai.

2. *Selection and Transformation*

   Tahap ini merupakan proses untuk melakukan perubahan bentuk tabel terhadap data yang telah dipilih, sehingga data yang dipilih sesuai dengan proses yang dilakukan.

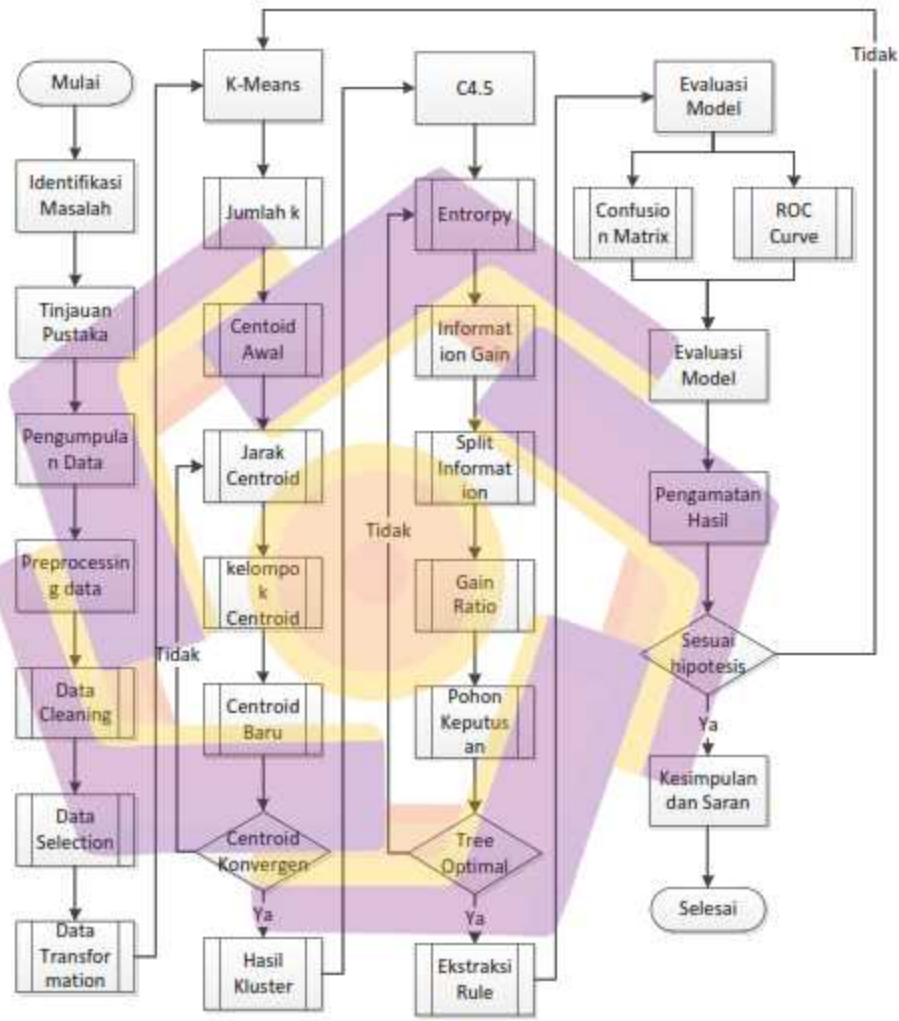
3. *Data mining*

   Tahapan yang secara langsung melibatkan teknik data mining. Teknik pemilihan data mining pada Prediksi Kelulusan Mahasiswa Informatika yaitu adalah algoritma C4.5 dan K-Means

4. *Evaluation and Interpretation*

   Tahap ini merupakan proses untuk menguji dan memvalidasi data yang telah dimodelkan sebelumnya.

1.8.4. Alur Penelitian



Gambar 1.1 Alur Penelitian

1.9. Sistematika Penulisan

Sistematika penulisan dibuat untuk mempermudah dalam penyusunan tesis yang memuat uraian secara garis besar isi dari tesis yaitu :

BAB I PENDAHULUAN

Bab ini berisi uraian latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, dan manfaat penelitian, beserta hipotesis

BAB II TINJAUAN PUSTAKA

Bab ini berisi tinjauan pustaka, keaslian penelitian, dan landasan teori. Tinjauan pustaka merupakan uraian hasil-hasil penelitian sebelumnya yang menjadi landasan penelitian, sedangkan landasan teori berisi teori-teori atau konsep untuk menyusun solusi pada penelitian yang akan dilakukan.

BAB III METODE PENELITIAN

Bab ini berisi jenis, sifat, dan pendekatan penelitian, metode pengumpulan data, metode analisis data, dan alur penelitian.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini menjelaskan hasil penelitian dan pembahasan dari penelitian yang telah dilakukan.

BAB V PENUTUP

Bab ini berisi kesimpulan dan saran yang diharapkan bermanfaat untuk penelitian selanjutnya.

BAB II

LANDASAN TEORI

2.1. Tinjauan Pustaka

Beberapa penelitian terdahulu yang digunakan sebagai referensi dari penelitian yang akan dibuat adalah sebagai berikut :

Penelitian (Pertiwi, Widyaningtyas, & Pujianto, 2017) bertujuan untuk mengklasifikasikan tingkat *drop out* berdasarkan provinsi di indonesia menjadi dua bagian yaitu Tinggi dan Rendah. Dataset yang dikumpulkan berjumlah 128 baris yang terdiri dari 7 atribut antara lain jumlah penduduk miskin, jumlah pembiayaan daerah, jumlah pendapatan daerah, jumlah belanja daerah, tpak, rasio gini, dan tingkat putus sd. Pengujian yang dilakukan menggunakan *confussion matrix* dan menghasilkan tingkat akurasi sebesar 71.2%.

Penelitian (Purwanto & Darmadi, 2018) menggunakan algoritma klasifikasi dalam menentukan minat siswa menggunakan 5 algoritma klasifikasi yang berbeda antara lain *K-Nearest Neighbor, Naïve Bayes, Rule Induction* dan *Deep Learning*. Dari hasil pengujian diketahui bahwa algoritma C4.5 memiliki nilai *precision* dan *accuracy* yang paling baik dibandingkan algoritma *K-Nearest Neighbor, Naïve Bayes, Rule Induction* dan *Deep Learning*.

Penelitian (Haris, Abdullah, & Hasim, 2016) bertujuan untuk mempelajari prediksi pendaftaran mahasiswa di perguruan tinggi menggunakan *data mining*. Beberapa metode dikombinasikan untuk menghasilkan prediksi dengan hasil terbaik berdasarkan model yang dibentuk.

Penelitian (Singh, Sabitha, Bansal, & Cse, 2016) bertujuan untuk memprediksi kinerja mahasiswa menggunakan K-means. Atribut yang digunakan dalam penelitian ini antara lain Xth, XIIth, B.Tech marks, projects, internships, and skills set. kemampuan mahasiswa dan dibagi menjadi 3 cluster. dari hasil pengujian diperoleh mahasiswa yang memiliki kinerja terendah pada cluster 0.

Penelitian (Purba, Tamba, & Saragih, 2018) bertujuan untuk mengelompokan mahasiswa yang berpotensi untuk drop out menggunakan K-Means. Data mahasiswa berjumlah 36 record yang terdiri dari beberapa atribut antara dari No, No Nim Name Courses Program, Credits Sks, Quality Total, Gpa yang kemudian dikelompokan menjadi 3 klaster mahasiswa yaitu sangat berpotensi drop out, berpotensi drop out dan tidak berpotensi drop out. Dari hasil penelitian diketahui bahwa mahasiswa lebih berpotensi drop out terdapat pada cluster 1 karena memiliki Credit Total System, Quality Total, and Grade Point Average (GPA) terendah dibandingkan cluster 2 dan 3.

Penelitian (Khan & Mohamudally, 2016) bertujuan untuk meprediksi penderita kanker payudara dengan kombinasi teknik clustering menggunakan K-means dengan klasifikasi algoritma ID3. Dari hasil pengujian menggunakan ID3 dan K-means menghasilkan rules yang lebih sedikit yang berdampak pada waktu eksukusi yang lebih singkat sehingga model yang dihasilkan efisien dalam melakukan ekstrasi rule menggunakan ID3 dan K-means.

Penelitian (Umam et al., 2017) bertujuan untuk melalukan kombinasi antara KNN dan K-means. Dari hasil pengujian diperoleh akurasi tertinggi sebesar 71,43% disaat jumlah k adalah 1 dan 2 sedangkan akurasi terendah sebesar

60,71% disaat nilai k berjumlah 4, 6, 8, 9 dan 10.

Penelitian (Anam & Santoso, 2018) membandingkan tingkat akurasi antara algoritma C4.5 dan dan *Naive Bayes* dalam klasifikasi mahasiswa penerima beasiswa. Hasil pengujian menunjukan algoritma C4.5 memiliki tingkat akurasi sebesar 96.40% lebih baik dari tingkat akurasi algoritma *Naive Bayes* sebesar 95.11%. Pengujian divalidasi dengan 10-fold cross validation dan evalusi menggunakan *confusion matrix* sekaligus evaluasi kinerja model menggunakan tool RapidMiner.

Penelitian (Supriyanti, Kusrini, & Amborowati, 2016) bertujuan untuk memprediksi ketepatan pemilihan konsentrasi mahasiswa menggunakan algoritma C4.5 dan *Naive bayes* yang dikombinasikan dengan metode *Forward selection* untuk menghasilkan akurasi yang lebih baik. Dari hasil pengujian diperoleh tingkat akurasi C4.5 sebesar 84,43% dan *Naive bayes* sebesar 78,47% . namun setelah penambahan metode untuk seleksi fitur menunjukan peningkatan akurasi C4.5 sebesar 0.40% dan 3.54%.

Penelitian (Daud et al., 2017) bertujuan memprediksi kinerja mahasiswa menggunakan beberapa algoritma klasifikasi yang berbeda. Data yang digunakan diperoleh dari beberapa universitas yang berbeda. Hasil pengujian diketahui bahwa SVM memiliki akurasi yang paling tinggi dibandingkan C4.5, CART, Bayes network dan Naive Bayes.

Penelitian (Rajeshinigo & Jebamalar, 2017) bertujuan untuk meningkatkan akurasi algoritma C4.5 dengan melakukan pengelompokan data interval menjadi 2 klaster menggunakan K-means, data yang digunakan dalam

penelitian ini merupakan data mahasiswa yang terdiri dari 8 atribut antara lain IAT(Internal Assessment Test), AS(Attendance Status), CA(Current Arrears), HTL(Hostel), PC(Project Completion), AC(Assignment Completion), PW(Parents Work) dan AD(Academic Detention). Hasil Penggabungan C4.5 menggunakan K-means menghasilkan akurasi sebesar 92% meningkat 19% dibandingkan yang hanya menggunakan C4.5.

2.2. Landasan Teori

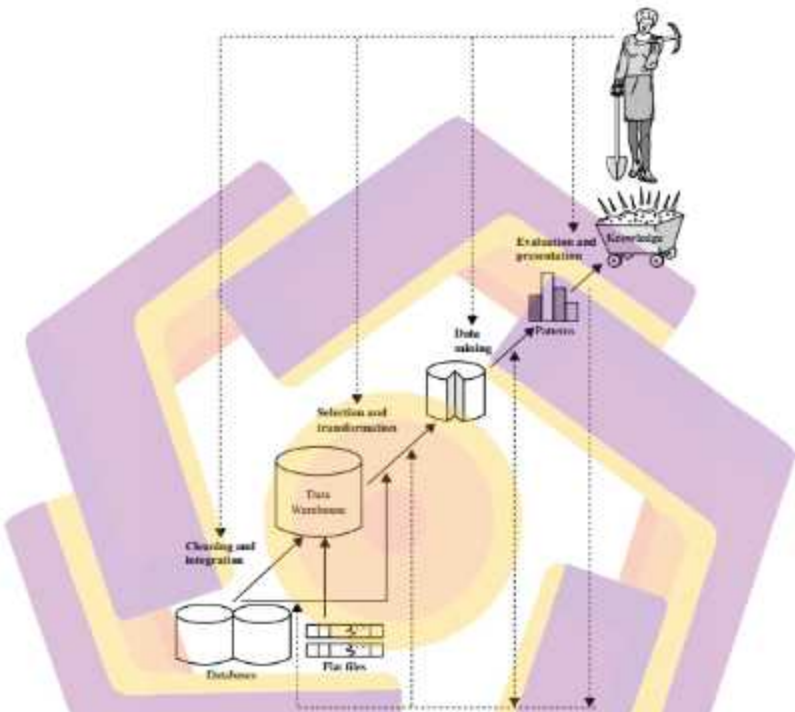
2.2.1. *Data Mining*

Data mining merupakan proses untuk menemukan sebuah pengetahuan dengan mencari pola tersembunyi dalam database yang menggunakan menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat (Turban et al., 2005).

Data mining merupakan serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data yang ada yang biasanya berupa data yang sangat besar dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga (Kusrini & Luthfi, 2009).

Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan secara otomatis (Hermawati, 2013).

2.2.2. Proses *Data Mining* Untuk *Knowledge Discovery*



Gambar 2.1 *Data mining* untuk *knowledge discovery* (Han, Kamber, & Pei, 2012)

Berikut ini merupakan rangkaian proses dalam data mining untuk menghasilkan sebuah pengetahuan seperti yang ditampilkan pada gambar 2.1 :

1. *Data Cleaning*

   Tahapan menghilangkan data noise dan data yang tidak konsisten atau relevan dengan tujuan akhir dari proses *data mining*.

2. *Data Integration*

   Tahapan menggabungkan atau mengkombinasikan data dari berbagai macam sumber.

3. *Data Selection*

   Tahapan memilih atau menyeleksi data apa saja yang relevan dan diperlukan dari database.

4. *Data Transformation*

   Tahapan mentransformasikan data ke dalam bentuk yang lebih sesuai untuk di *mining*.

5. *Data Mining*

   Tahapan penerapan suatu metode untuk menghasilkan pengetahuan berdasarkan suatu pola-pola tertentu.

6. *Pattern Evaluation*

   Tahan identifikasi untuk mengevaluasi apakah pola yang ditemukan dapat mewakili pengetahuan yang dihasilkan berdasarkan perhitungan tertentu.

7. *Knowledge Presentation*

   Tahapan mempresentasikan knowledge yang sudah didapatkan dari user

2.2.3. Analisis Korelasi

2.2.3.1. *Koefisien Kontingensi*

Koefisien Kontingensi adalah merupakan metode yang digunakan untuk menghitung hubungan antar 2 variabel untuk data yang bertipe nominal(Sugiyono, 2015). Misalnya terdapat 2 variabel A dan B yang masing-masing memiliki kategori dimana kategori A antara lain A1, A2, A3,..., Ak dan kategori B antara lain B1, B2, B3,..., Br. (Siegel, 1994).

2.2.4. Klusterisasi

Klusterisasi adalah pengelompokkan sebuah *record*, pengamatan dan membentuk kelas kedalam sebuah objek yang mempunyai kemiripan. Algoritma klaster melakukan pembagian keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (*homogen*), yang mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal. Salah satu metode yang dapat digunakan untuk klasterisasi adalah *K-mean*.

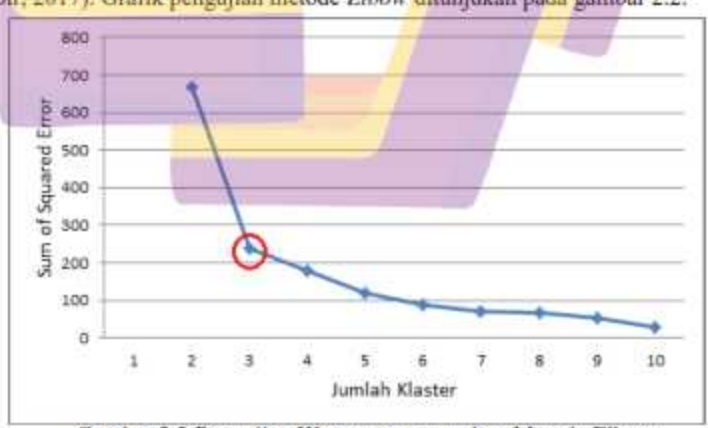
2.2.4.1. Algoritma K-Means

K-means merupakan metode klastering yang membagi data menjadi suatu klaster. Klastering dilakukan untuk mengelompokan data kedalam suatu klaster yang memiliki karakteristik yang sama. Dalam *k-means* pemisahan data dilakukan dengan melakukan perhitungan secara terus-menerus sampai tidak ada perubahan data dalam setiap klaster.

Parameter yang digunakan dalam algoritma *K-means* adalah nilai K. Nilai k yang digunakan biasanya didasarkan pada informasi yang di ketahui sebelumnya tentang sebenarnya berapa banyak *cluster* data yang muncul dalam x, berapa banyak *cluster* yang dibutuhkan untuk penerapannya, atau jumlah k paling ideal dapat ditentukan dengan melakukan pengujian salah satu metode yang dapat digunakan untuk mencari nilai k terbaik adalah Metode *Elbow*.

2.2.5. Metode *Elbow*

Metode *Elbow* merupakan metode yang digunakan untuk mencari jumlah klaster yang paling ideal berdasarkan perbandingan nilai *Sum of Squared Error* antara jumlah cluster yang nantinya akan membentuk sebuah siku pada suatu titik. Hasil perbandingan nila *Sum of Squared Error* pada setiap klaster dapat ditunjukkan dengan menggunakan grafik. Apabila dalam grafik menunjukan penurunan paling besar maka jumlah klaster tersebut merupakan klaster terbaik (Alatubir, 2017). Grafik pengujian metode *Elbow* ditunjukan pada gambar 2.2.



Gambar 2.2 Pengujian Klaster menggunakan Metode Elbow

2.2.6. Klasifikasi

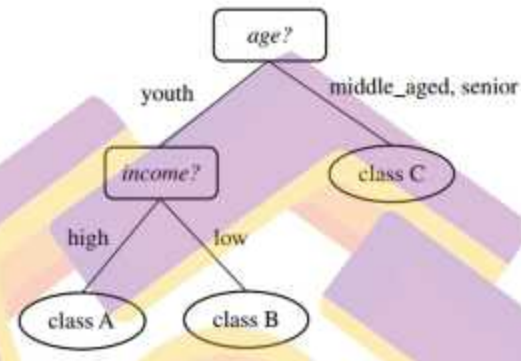
Klasifikasi adalah tindakan untuk memberikan kelompok pada setiap keadaan. Setiap keadaan berisi sekelompok atribut, salah satunya adalah class attribute. Metode ini butuh untuk menemukan sebuah model yang dapat menjelaskan *class attribute* itu sebagai fungsi dari *input attribute*. Misalnya untuk mengklasifikasikan suhu dalam tiga kelas, yaitu suhu panas, suhu sejuk, suhu dingin. Salah satu model yang dapat digunakan untuk klasifikasi adalah model pohon keputusan (Kusrini & Luthfi, 2009).

2.2.6.1. *Decision Tree* (Pohon Keputusan)

Pohon keputusan merupakan metode klasifikasi dan prediksi yang kuat dan terkenal, metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang dapat merepresentasi aturan yang akan dihasilkan. Pohon keputusan beguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel dan variabel target. Salah satu algoritma yang dapat dipakai dalam pembentukan pohon keputusan yaitu C4.5 (Kusrini & Luthfi, 2009).

Pohon *(tree)* adalah sebuah struktur data yang terdiri dari simpul *(node)* dan rusuk *(edge)*. Simpul pada sebuah pohon dibedakan menjadi tiga, yaitu simpul akar *(root node)*, simpul percabangan/internal *(branch/ internal node)* dan simpul daun *(leaf node)*. Pohon keputusan merupakan representasi sederhana dari teknik klasifikasi untuk sejumlah kelas berhingga, dimana simpul internal maupun simpul akar ditandai dengan nama atribut, rusuk-rusuknya diberi label nilai atribut

yang mungkin dan simpul daun ditandai dengan kelas-kelas yang berbeda (Hermawati, 2013), Contoh pohon keputusan seperti yang ditampilkan pada gambar 2.2.



Gambar 2.2 Contoh Pohon Keputusan (Han, Kamber, & Pei, 2012)

Pada pohon keputusan klasifikasi keputusannya adalah Class A, Class B atau Class C. Apabila Age = Youth & Income = hign (Class A), Age = Youth & Income = Low (Class B), Age = Middle_aged, Senior (Class C).

2.2.6.1.1.Algoritma C4.5

Algoritma C4.5 merupakan salah satu algortima yang dapat digunakan untuk membentuk pohon keputusan yang ditemukan oleh John Ross Quinlan. Algoritma C4.5. Algoritma C4.5 menggunakan *Gain Ratio* dalam pemilihan atribut. *Gain Ratio* digunakan untuk mengatasi atribut yang memiliki nilai yang sangat bervariasi dan dihitung berdasarkan *Split Information* (Suyanto, 2017).

Entropy adalah parameter yang digunakan untuk mengukur keberagaman dalam suatu himpunan data. Apabila semakin besar tingkat keberagaman dalam suatu himpunan data maka akan semakin besar juga nilai *entropy*-nya. *Information Gain* digunakan untuk mengukur efektivitas suatu atribut dalam mengklasifikasian data. *Split Information* digunakan untuk mencari nilai Gain Ratio. *Gain Ratio* digunakan untuk mengatasi atribut yang memiliki nilai yang sangat bervariasi yang tidak bisa diselesaikan menggunakan Information Gain(Suyanto, 2017)(Kusrini & Luthfi, 2009). Berikut adalah langkah-langkah dalam membangun pohon keputusan dengan menggunakan algoritma C4.5 yaitu :

1. Pilih atribut sebagai *node* akar
2. Buat cabang untuk tiap-tiap nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

2.2.7. Evaluasi Model

2.2.7.1. *Confusion Matrix*

Confusion matrix adalah alat yang digunakan sebagai evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah *matrix* dari prediksi yang akan dibandingkan dengan kelas sebenarnya atau dengan kata lain berisi informasi nilai sebenarnya dan prediksi pada klasifikasi (Gonureschu, 2011) tabel *confusion matrix* ditampilkan pada tabel 2.1.

Tabel 2.1 *Confusion Matrix*

Classification	*Predicted Class*	
	Class = Yes	*Class = No*
Class = Yes	TP(*True Positive*)	FN(*False Negative*)
Class = No	FP(*False Positive*)	TN(*True Negative*)

Pada tabel *confusion matrix* di atas, *true positive* (TP) adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positive* (FP) adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* (FN) adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* (TN) adalah jumlah *record* negatif yang diklasifikasikan sebagai negatif. Setelah data uji diklasifikasikan maka akan didapatkan *confusion matrix* sehingga dapat dihitung jumlah *akurasi, presisi,* dan *recall* (Gonureschu, 2011).

2.2.7.2. *Reciever Operating Curve*(ROC)

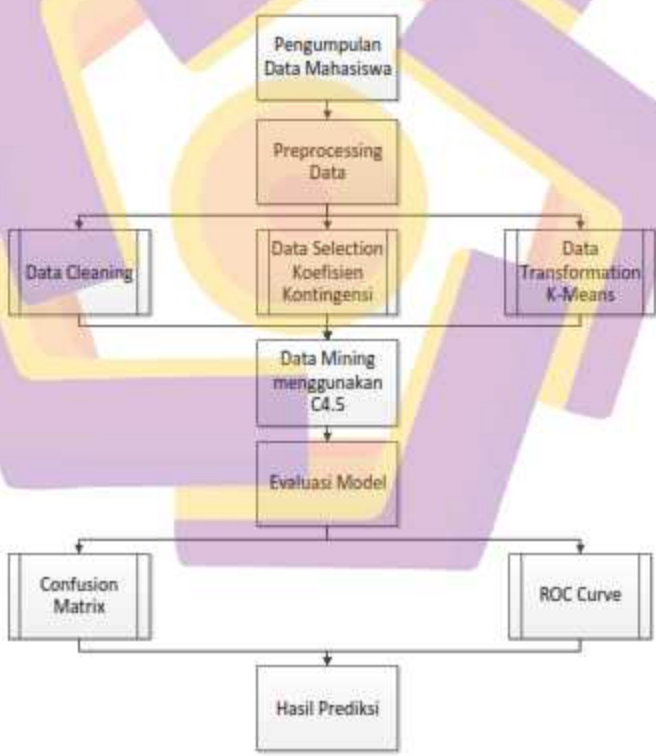
Reciever Operating Curve(ROC) menunjukan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan confusion matrix. *ROC* adalah grafik dua dimensi dengan *false positives* sebagai garis horizontal dan *true positives* untuk mengukur perbedaan performasi metode yang digunakan. (Gonureschu, 2011).

BAB III

METODOLOGI PENELITIAN

3.1. Gambaran Umum Penelitian

Pada penelitian ini bertujuan untuk memprediksi ketepatan kelulusan mahasiswa dan mengetahui faktor yang paling mempengaruhi tingkat kelulusan mahasiswa menggunakan algoritma C4.5 dan menggunakan *K-means* untuk discretize datanya. Berikut adalah skema sistem ditampilkan pada gambar 3.1.



Gambar 3.1 Skema penelitian

Tahapan berdasarkan gambar 3.1 yaitu, pertama kali data yang berkaitan dengan kelulusan mahasiswa dikumpulkan kemudian dataset tersebut di *cleaning*, setelah data tersebut bersih kemudian dilakukan seleksi atribut menggunakan *koefisien kontingensi* untuk menghetahui hubungan korelasi antara atribut dengan kelulusan mahasiswa selanjutnya data di *descretize* menggunakan *K-means* untuk merubah data interval menjadi data kategori. Setelah itu dilakukan pengujian menggunakan metode elbow untuk mencari nilai k paling optimal. Kemudian dilakukan proses transformasi data sejumlah nilai k yang paling optimal. Dataset yang dihasilkan kemudian dimodelkan menggunakan algoritma C4.5 untuk memprediksi kelulusan mahasiswa yang berupa sebuah aturan dari pohon keputusan, tahapan selanjutnya yaitu evaluasi model untuk mengukur kinerja model yang sudah dihasilkan menggunakan *Confuson Matrix* dan *ROC Curve*.

3.2. Pengumpulan Data

Dalam penelitian ini menggunakan dataset mahasiswa yang telah lulus pada tahun 2011-2014 yang dapat digunakan untuk memprediksi kelulusan mahasiswa 2015-Sekarang. Data tersebut diperoleh dari Innovation Centre AMIKOM yang berupa sebuah dokumen elektronik yang berisi data kelulusan mahasiswa.

3.3. *Preprocessing Data*

3.3.1. *Data Cleaning*

Data yang telah dipeoleh kemudian dilakukan cleaning untuk membuang data yang tidak konsisten, dan memperbaiki kesalahan penulisan pada data. Dalam pembersihan data dilakukan penambahan nilai pada attribut yang tidak memiliki nilai dan menghapus duplikasi data.

3.3.2. *Data Selection*

Proses *data selection* bertujuan untuk memilih data yang relevan yang digunakan untuk penelitian. Proses seleksi atribut menggunakan metode *koefisien kontingensi* untuk mengetahui korelasi antara 2 variabel yang betipe nominal. Varibel yang diuji merupakan atribut mahasiswa dan kelulusan mahasiswa. Berikut adalah kontingensi antar 2 varibel ditunjukan pada tabel 3.1.

Tabel 3.1 Kontingensi

Baris	Kolom				Total
	A_1	A_2	...	A_k	
B_1	(A_1B_1)	(A_2B_1)	...	(A_kB_1)	
B_2	(A_1B_2)	(A_2B_2)	...	(A_kB_2)	
...	...	...	...	...	
B_r	(A_1B_r)	(A_2B_r)		(A_kB_r)	
Total					

Dari tabel diatas baris merupakan variabel *dependen*(kelulusan mahasiswa) dan kolom merupakan variabel *independen*(atribut mahasiswa). selanjutnya dilakukan perhitungan chi kuadrat (X^2) kuadrat menggunakan persamaan 1.

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(OP_{ij} - E_{ij})^2}{EP_{ij}} \quad (1)$$

Kemudian nilai X^2 digunakan untuk menghitung *koefisien korelasi* menggunakan persamaan 2

$$C = \sqrt{\frac{x^2}{N + x^2}} \quad (2)$$

3.3.3. *Data Transformation*

3.3.3.1. *Data Discretization* menggunakan *K-Means*

Algoritma *K-Means* digunakan untuk mengelompakan data yang betipe interval menjadi kategori sejumlah nilai k. Nilai k yang digunakan untuk *cluster* pada penelitian ini berjumlah 3. Berikut adalah proses klastering menggunakan *k-means* antara lain :

1. Tentukan nilai k sebagai jumlah *cluster* yang ingin dibentuk.
2. Bangkitkan k *centroid* (titik pusat cluster) awal.
3. Hitung jarak setiap data ke masing-masing centroid menggunakan rumus *Euclidean Distance* seperti pada persamaan 3

$$D(X,Y) = \sum (X_i - Y_i)^2 \quad (3)$$

Keterangan :

D(X,Y) : Jarak objek antara Xi dan Yi

Xi : Koordinat dari objek Xi pada dimensi i

Yi : Koordinat dari objek Yi pada dimensi i

4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroid*nya.
5. Tentukan *centroid* baru dengan cara menghitung nilai rata-rata dari setiap nilai yang ada pada masing-masing *centroid* menggunakan persamaan (4).

$$Ci = \left(\frac{\sum x}{n} \right) \quad (4)$$

6. Apabila *centroid* baru dengan *centroid* lama tidak sama, proses diulang kembali dari langkah 3 hingga menghasilkan *centroid konvergen*.

3.3.3.2. Metode *Elbow*

Metode *Elbow* digunakan untuk menguji jumlah klaster paling optimal yang digunakan dalam penelitian. Pengujian dilakukan dengan membandingkan nilai SSE (*Sum of Square Error*) dari masing-masing nilai cluster. Berikut cara menghitung SSE pada *K-Means* menggunakan persamaan 5.

$$SSE = \sum_{k=1}^{k} \sum_{i \in S_k} \left\| x_i - c_k \right\|^2 \quad (5)$$

Setelah melakukan perhitungan *Sum of Squared Error* pada masing-masing klaster kemudian dilakukan pengamatan untuk mencari nilai k yang mengalami penurunan paling besar dan diikuti dengan penurunan secara perlahan-lahan sampai hasil dari nilai k tersebut stabil. Nilai k yang memiliki penurunan paling besar dan membentuk sebuah siku merupakan jumlah klaster paling ideal, misalnya nilai cluster K=2 ke K=3, kemudian dari K=3 ke K=4, terlihat

penurunan drastis membentuk siku pada titik K=3 maka nilai *cluster* k yang ideal adalah K=3.

3.4. *Data Mining*

3.4.1. Algoritma C4.5

Algoritma C4.5 digunakan untuk memodelkan data kelulusan mahasiswa yang terdiri dari klasfikasi mahasiswa tepat waktu dan mahasiswa terlambat yang digunakan untuk memprediksi kelulusan mahasiswa dimasa mendatang. Langkah-langkah dalam modeling data menggunakan algoritma C4.5 antara lain :

1. Melakukan perhitungan nilai *entropy* menggunakan persamaan 6

$$Entropy(S) = \sum_{i}^{c} - P_i Log_2 P_i \qquad (6)$$

Keterangan :

S : himpuanan kasus

c : jumlah nilai yang terdapat pada atribut target (jumlah kelas)

pi : rasio antar jumlah sampel dikelas i dengan jumlah semua sampel pada himpunan data

2. Melakukan perhitungan nilai *information gain* menggunakan persamaan 7.

$$Gain(S, A) = Entropy(S) - \frac{|S_v|}{|S|} Entropy(S_v) \qquad (7)$$

Keterangan :

A : atribut

V : menyatakan suatu nilai yang mungkin untuk atribut A

Value(A) : himpunan nilai-nilai yang mungkin untuk atribut A

$|S_v|$: jumlah sampel untuk nilai v

$|S|$: jumlah seluruh sampel data

Entropy (S_v) : entropy untuk sampel-sampel yang memiliki nilai v

3. Melakukan perhitungan *split information* menggunakan persamaan 8

$$SplitInformation(S, A) = \sum_{i=1}^{c} -\frac{|S_i|}{|S|} Log_2 \frac{|S_i|}{|S|} \quad (8)$$

Keterangan :

S : himpunan sampel data

S_1 – Sc: sub himpunan sampel data yang terbagi berdasarkan jumlah variasi nilai pada atribut A

4. Melakukan perhitungan *gain ratio* menggunakan persamaan 9

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInformation(S, A)} \quad (9)$$

5. Proses tersebut diulang pada tiap atribut yang memiliki nilai sehingga diperoleh nilai *gain ratio* tertinggi

3.5. Evaluasi Model

3.5.1. *Confusion Matrix*

Dalam evaluasi menggunakan *confusion matrix* menggunakan data klasifikasi tepat waktu dan terlambat. *true positive* (TP) adalah jumlah klasifikasi tepat waktu positif yang diklasifikasikan sebagai positif, *false positive* (FP) adalah jumlah jumlah klasifikasi tepat waktu negatif yang diklasifikasikan sebagai positif, *false negatives* (FN) adalah jumlah klasifikasi terlambat positif yang diklasifikasikan sebagai negatif, *true negatives* (TN) adalah jumlah klasifikasi terlambat negatif yang diklasifikasikan sebagai negatif yang ditampilkan pada tabel 3.2.

Tabel 3.2 Nilai *Confusion Matrix*

Actual	Predicted	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	TP(*True Positive*)	FP(*False Negative*)
True Terlambat	FP(*False Positive*)	TN(*True Negative*)

Setelah data uji diklasifikasikan maka akan didapatkan *confusion matrix* sehingga dapat dihitung jumlah *akurasi, presisi,* dan *recall*. Berikut rumus untuk menghitung *akurasi, presisi,* dan *recall* pada *confusion matrix* ditunjukan menggunakan persamaan 9, 10 dan 11 :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (9)$$

$$\text{Pr}esisi = \frac{TP}{TP + FP} \times 100\% \quad (10)$$

$$\text{Re}call = \frac{TP}{TP + FN} \times 100\% \quad (11)$$

3.5.2. *Reciever Operating Curve*(ROC)

Untuk menguji kenerja klasifikasi digunakan *ROC curve* dengan cara menghitung *TPR/Sensitivity, Specificity,* dan *FPR* berdasarkan *confusion matrixnya* untuk tiap treshold. Berikut rumus untuk menghitung *TPR/Sensitivity, Specificity,* dan *FPR* pada *confusion matrix* ditunjukan menggunakan persamaan 12, 13 dan 14 :

$$TPR\,/\,Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

$$Specificit\,y = \frac{TN}{TN + FP} \times 100\% \quad (13)$$

$$FPR = 1 - Specificit\,y$$

$$FPR = \frac{TP}{TP + FN} \times 100\% \quad (14)$$

Dari perhitungan perbandingan TPR dan FPR kemudian digambarkan dalam *ROC Curve. True Positive Rate* (TPR) adalah jumlah klasifikasi tepat waktu positif yang diklasifikasikan sebagai positif, *False positive Rate* (FPR) adalah jumlah jumlah klasifikasi tepat waktu negatif yang diklasifikasikan sebagai positif untuk menghasilkan nilai *Area Under Curve*(AUC). Nilai AUC dapat diklasifikasikan menjadi lima kelompok yaitu :

0.90 – 1.00 = *Excellent Classification*

0.80 – 0.90 = *Good Classification*

0.70 – 0.80 = *Fair Classification*

0.60 – 0.70 = *Poor Classification*

0.50 – 0.60 = *Failure*

BAB IV

HASIL DAN PEMBAHASAN

4.1. Pendahuluan

Pengujian yang dilakukan menggunakan data berjumlah 1941 *record* yang terdiri dari data kelulusan mahasiswa tahun 2011-2014, penerapan metode algoritma C4.5 dan *K-Means* digunakan untuk mendapatkan akurasi terbaik dari model yang digunakan memprediksi kelulusan mahasiswa. tahapan dalam penelitian yaitu proses Pengumpulan data, *Data Cleaning, Data Selection, Discretize data* menggunakan *K-Means,* Pengujian nilai k menggunakan Metode *Elbow* dan *Modeling Data* Menggunakan Algoritma C4.5. untuk memudahkan pengujian digunakan beberapa software yaitu *microsoft excel* dan *rapidminer.*

4.2. Pengumpulan data

Pengumpulan data dilakukan untuk mendapatkan data yang berkaitan dengan kelulusan mahasiswa. Data yang diperoleh berasal dari Innovation Center AMIKOM adalah data mahasiswa tahun ajaran 2011-2014 dengan jumlah data sebanyak 1941 *record* dengan rincian seperti ditunjukan pada tabel 4.1.

Tabel 4.1 *Dataset* Mahasiswa

Angkatan	Jumlah
2011	704
2012	693
2013	478
2014	66
Total	1941

Sedangkan data atribut mahasiswa terdiri dari faktor internal antara lain NPM, Jenis Kelamin, Asal, Angkatan, Tahun Lulus, SKS Total. IPK sem 1, IPK sem 2, IPK sem 3, IPK sem 4, IPK sem 5, IPK sem 6, IPK sem 7, IPK sem 8, Presensi sem 1, Presensi sem 2, Presensi sem 3, Presensi sem 4, Konsentrasi, Nilai UN, Nilai TOEFL, SLTA, Makul Mengulang dan Makul Remedi dan faktor eksternal antara lain Pekerjaan Ayah, Pekerjaan Ibu, Penghasilan Ayah, Penghasilan Ibu, Beasiswa dengan rincian seperti ditunjukan pada tabel 4.2.

Tabel 4.2 Atribut Data Mahasiswa

No	Atribut	Keterangan
1	NPM	Merupakan no induk mahasiswa
2	Jenis Kelamin	Merupakan jenis kelamin mahasiswa yang terdiri dari laki-laki atau perempuan
3	Asal	Merupakan alamat tempat tinggal mahasiswa
4	Angkatan	Merupakan angkatan masuk mahasiswa
5	Tahun Lulus	Merupakan tahun kelulusan mahasiswa
6	Total Sks	Merupakan total sks yang diambil mahasiswa Semester 1-4
7	IPK sem 1	Merupakan index prestasi kumulatif mahasiswa semester 1
8	IPK sem 2	Merupakan index prestasi kumulatif mahasiswa semester 2
9	IPK sem 3	Merupakan index prestasi kumulatif mahasiswa semester 3
10	IPK sem 4	Merupakan index prestasi kumulatif mahasiswa semester 4
11	IPK sem 5	Merupakan index prestasi kumulatif mahasiswa semester 5
12	IPK sem 6	Merupakan index prestasi kumulatif mahasiswa semester 6
13	IPK sem 7	Merupakan index prestasi kumulatif mahasiswa semester 7
14	IPK sem 8	Merupakan index prestasi kumulatif mahasiswa semester 8
15	Presensi sem 1	Merupakan jumlah Presensi mahasiswa di semester 1
16	Presensi sem 2	Merupakan jumlah Presensi mahasiswa di semester 2
17	Presensi sem 3	Merupakan jumlah Presensi mahasiswa di semester 3
18	Presensi sem 4	Merupakan jumlah Presensi mahasiswa di semester 4
19	Konsentrasi	Merupakan konsentrasi yang diambil mahasiswa
20	Pekerjaan Ayah	Merupakan pekerjaan ayah mahasiswa
21	Pekerjaan Ibu	Merupakan pekerjaan ibu maahasiswa
22	Penghasilan Ayah	Merupakan jumlah penghasilan ayah mahasiswa
23	Penghasilan Ibu	Merupakan jumlah penghasilan ibu mahasiswa
24	Beasiswa	Merupakan keterangan penerima beasiswa
25	Nilai UN	Merupakan nilai ujian nasional mahasiswa

Tabel 4.2 (Lanjutan)

No	Atribut	Keterangan
26	Nilai Toefl	Merupakan nilai toefl mahasiswa
27	SLTA	Merupakan asal stla mahasiswa
28	Makul Mengulang	Merupakan mata kuliah mengulang yang diambil mahasiswa pada semester 1-4
29	Makul Remedi	Merupakan mata kuliah remedi yang diambil mahasiswa pada semester 1-4

Berikut adalah contoh dataset kelulusan mahasiswa yang telah diperoleh ditampilkan pada tabel 4.3.

Tabel 4.3 Dataset Mentah Mahasiswa

Npm	Jenis Kelamin	SksTotal	NilaiUN	IPS_1	IPS_2	IPS_3	IPS_4	n
11.11.4617	L	148	0	3,17	3,08	3,17	3,15	...
11.11.4618	L	144	7,25	3,67	3	2,23	2,23	...
11.11.4620	L	144	8,53	3,83	3,83	3,86	3,83	...
11.11.4621	L	144	7,68	3,33	3,38	3,32	3,26	...
11.11.4623	L	148	0	3,5	3,5	3,58	3,67	...

4.3. *Preprocessing Data*

4.3.1. *Data Cleaning*

Pembersihan data dilakukan dengan mengisi nilai atribut yang kosong dan menghapus data ganda dari dataset. Data yang bernilai kosong diisi dengan rata-rata jumlah data dari keseluruhan data untuk data bertipe numerical sedangkan data bertipe string diisi dengan nilai yang paling sering ada pada masing-masing atribut. Berikut adalah contoh dataset mahasiswa setelah proses *cleaning* ditampilkan pada tabel 4.4.

Tabel 4.4 Dataset Mahasiswa setelah *Cleaning*

Npm	Jenis Kelamin	SksTotal	NilaiUN	IPS_1	IPS_2	IPS_3	IPS_4	n
11.11.4617	L	148	7,83	3,17	3,08	3,17	3,15	...
11.11.4618	L	144	7,25	3,67	3	2,23	2,23	...
11.11.4620	L	144	8,53	3,83	3,83	3,86	3,83	...
11.11.4621	L	144	7,68	3,33	3,38	3,32	3,26	...
11.11.4623	L	148	7,83	3,5	3,5	3,58	3,67	...

4.3.2. *Data Selection*

Proses seleksi atribut menggunakan *koefisien kontingensi* untuk mengetahui korelasi antara variabel Jenis Kelamin, Asal, Nilai TOEFL, Nilai UN dan SLTA dengan kelulusan mahasiswa. Berikut ini adalah contoh pengujian *korelasi* antara atribut SLTA dengan kelulusan mahasiswa yang ditunjukan pada tabel 4.5.

Tabel 4.5 Kontingensi SLTA

Kelulusan	SLTA			Jumlah
	SMA	SMK	MA	
Tepat waktu	311	1291	30	1632
Terlambat	108	187	14	309
Jumlah	419	1478	44	1941

Berdasarkan tabel 4.3 kemudian dihitung berapa persen dari masing-masing sampel yang lulus tepat waktu dan terlambat untuk menghitung nilai frekuensi yang diharapkan (fh) :

1. Perhitungan sampel yang lulus tepat waktu yaitu :

$$\frac{311+1291+30}{1941}=\frac{1632}{1941}=0{,}841$$

2. Perhitungan sampel yang lulus terlambat yaitu :

$$\frac{108+187+14}{1941}=\frac{309}{1941}=0{,}159$$

Setelah ditentukan *persentase* dari sampel mahasiswa yang lulus tepat waktu dan terlambat kemudian dilakukan perhitungan frekuensi yang diharapkan(fh) kelompok kelulusan mahasiswa.

1. lulus tepat waktu

F_h SMA = 0,841 x 419 = 352,297

F_h SMK = 0,841 x 1478 = 1242,708

F_h MA = 0,841 x 44 = 36,995

= 1632

2. lulus terlambat

F_h SMA = 0,159 x 419 = 66,703

F_h SMK = 0,841 x 1478 = 235,292

F_h MA = 0,159 x 44 = 7,005

= 309

Dari hasil perhitungnan berikut selanjutnya dimasukan kedalam tabel perhitungan kontingensi SLTA yang ditunjukan pada tabel 4.6.

Tabel 4.6 Perhitungan Kontingensi SLTA

Kelulusan	SMA		SMK		MA		Jumlah
	f_o	f_h	f_o	f_h	f_o	f_h	
Tepat Waktu	311	352,297	1291	1242,708	30	36,995	1632
Terlambat	108	66,703	187	235,292	14	7,005	309
Jumlah	419		1478		44		1941

selanjutnya dilakukan perhitungan chi kuadrat menggunakan permasaan 1

$$X^2 = \frac{(311-352{,}297)^2}{352{,}297} + \frac{(1291-1242{,}708)^2}{1242{,}708} + \frac{(30-36{,}995)^2}{36{,}995} + \ldots$$

$$= \frac{(108-66{,}703)^2}{66{,}703} + \frac{(187-235{,}292)^2}{235{,}292} + \frac{(14-7{,}005)^2}{7{,}005}$$

$$X^2 = 4{,}841 + 1{,}877 + 25{,}567 + 9{,}912 + 1{,}323 + 6{,}986$$

$$X^2 = 50{,}505$$

Perhitungan chi kuadrat diperoleh nilai chi kuadrat hitung = 50,505 selanjutnya untuk menghitung *koefisien kontingensi c* maka nilai tersebut dimaasukan dalam persamaan 2.

$$C = \sqrt{\frac{50{,}505}{1941 + 50{,}505}}$$

$$C = 7{,}108$$

Jadi besarnya koefisien korelasi antara atribut SLTA dan Kelususan mahasiswa adalah 7,108. Pengujian korelasi dan signifikansi *koefisien c* membandingkan nilai chi kuadrat hitung dengan chi kuadrat tabel, berdasarkan perhitungan diatas ternyata kuadrat chi hitung lebih besar dari chi tabel (Chi-Square tabel pada DF 2 dan signifikansi 0,05 = 5.991). Hasil pengujian korelasi ditujukan pada tabel 4.7.

Tabel 4.7 Pengujian Korelasi

No	Atribut	Perbandingan X^2 dan X^2_a	Keterangan
1	SLTA	50,505 > 5,991	Berkorelasi dengan Kelulusan
2	Jenis Kelamin	31,019 > 3,841	Berkorelasi dengan Kelulusan
3	Alamat	0,238 < 5,991	Tidak Berkorelasi dengan Kelulusan
4	Nilai UN	0,920 < 5,991	Tidak Berkorelasi dengan Kelulusan
5	Nilai TOEFL	4,465 < 5,991	Tidak Berkorelasi dengan Kelulusan

Berdasarkan perhitungan pada tabel 4.7 Diketahui bahwa atribut Asal, Nilai UN dan Nilai TOEFL tidak memiliki korelasi atau hubungan signifikan dengan kelulusan sehingga dihapus. Daftar atribut yang dihapus ditunjukan di tabel 4.8.

Tabel 4.8 Daftar Atribut Yang Dihapus

No	Atribut	Keterangan
1	NPM	Data Tidak Relevan Karena merupakan ID mahasiswa
2	Asal	Data Tidak Relevan Karena tidak berkorelasi dengan kelulusan
3	IPK sem 5	Data Tidak Relevan karena menggunakan prediksi 4 semester diawal untuk menentukan kelulusan 4 semester selanjutnya
4	IPK sem 6	Data Tidak Relevan karena menggunakan prediksi 4 semester diawal untuk menentukan kelulusan 4 semester selanjutnya
5	IPK sem 7	Data Tidak Relevan karena menggunakan prediksi 4 semester diawal untuk menentukan kelulusan 4 semester selanjutnya
6	IPK sem 8	Data Tidak Relevan karena menggunakan prediksi 4 semester diawal untuk menentukan kelulusan 4 semester selanjutnya
7	Konsentrasi	Sebagian Besar Data Tidak Ada
8	Pekerjaan Ayah	Sebagian Besar Data Tidak Ada
9	Pekerjaan Ibu	Sebagian Besar Data Tidak Ada
10	Penghasilan Ayah	Sebagian Besar Data Tidak Ada
11	Penghasilan Ibu	Sebagian Besar Data Tidak Ada
12	Beasiswa	Sebagian Besar Data Tidak Ada
13	Nilai UN	Data Tidak Relevan Karena tidak berkorelasi dengan kelulusan
14	Nilai TOEFL	Data Tidak Relevan Karena tidak berkorelasi dengan kelulusan

Sedangkan perhitungan tabel 4.7 Diketahui bahwa atribut Jenis Kelamin dan memiliki korelasi atau hubungan signifikan dengan kelulusan sehingga atribut Jenis Kelamin dan SLTA akan digunakan. Daftar atribut yang digunakan ditunjukan pada tabel 4.9.

Tabel 4.9 Hasil Seleksi Atribut

No	Atribut	Keterangan
1	IPK sem 1	Data Relevan karena menggunakan prediksi 4 semester diawal untuk menentukan kelulusan 4 semester selanjutnya
2	IPK sem 2	Data Relevan karena menggunakan prediksi 4 semester diawal untuk menentukan kelulusan 4 semester selanjutnya
3	IPK sem 3	Data Relevan karena menggunakan prediksi 4 semester diawal untuk menentukan kelulusan 4 semester selanjutnya
4	IPK sem 4	Data Relevan karena menggunakan prediksi 4 semester diawal untuk menentukan kelulusan 4 semester selanjutnya
5	Total Sks sem 1-4	Data Relevan karena memiliki hubungan dengan kelulusan
6	Presensi Sem 1	Data Relevan karena memiliki hubungan dengan kelulusan
7	Presensi Sem 2	Data Relevan karena memiliki hubungan dengan kelulusan
8	Presensi Sem 3	Data Relevan karena memiliki hubungan dengan kelulusan
9	Presensi Sem 4	Data Relevan karena memiliki hubungan dengan kelulusan
10	Makul Mengulang	Data Relevan karena memiliki hubungan dengan kelulusan
11	Makul Remedi	Data Relevan karena memiliki hubungan dengan kelulusan
12	Jenis Kelamin	Data Relevan karena berkorelasi dengan kelulusan
13	SLTA	Data Relevan karena berkorelasi dengan kelulusan
14	Klasifikasi	Data Relevan karena merupakan klasifikasi kelulusan mahasiswa

Atribut Angkatan dan Tahun Lulus digunakan sebagai Label Klasifikasi untuk menentukan mahasiswa yang lulus tepat waktu atau terlambat. Berikut adalah contoh dataset mahasiswa setelah proses *selection* ditampilkan pada tabel 4.10.

Tabel 4.10 Dataset Mahasiswa setelah *Selection*

Npm	Jenis Kelamin	SksTotal	SLTA	IPS_1	IPS_2	IPS_3	IPS_4	n
11.11.4617	L	148	SMK	3,17	3,08	3,17	3,15	...
11.11.4618	L	144	SMA	3,67	3	2,23	2,23	...
11.11.4620	L	144	SMK	3,83	3,83	3,86	3,83	...
11.11.4621	L	144	SMK	3,33	3,38	3,32	3,26	...
11.11.4623	L	148	SMK	3,5	3,5	3,58	3,67	...

4.3.3. *Data Transformation*

4.3.3.1. *Data Discretization* menggunakan *K-means*

Setelah semua data dibersihkan dari proses *data cleaning* selanjutnya dilakukan proses *discretization* data untuk merubah data interval menjadi kategori pada atribut IPK sem 1, IPK sem 2, IPK sem 3, IPK sem 4, Total Sks, Presensi Sem 1, Presensi Sem 2, Presensi Sem 3, Presensi Sem 4, Makul Mengulang dan Makul Remedi menggunakan algoritma *K-Means Clustering* agar data dapat digunakan untuk dimodelkan. Proses pengelompokan beberapa data menjadi beberapa cluster yaitu :

1. Tentukan jumlah cluster yang diinginkan. Dalam penelitian ini data tiap atribut dikelompokan menjadi 3 cluster .
2. Tentukan titik pusat awal dari setiap cluster . Dalam penelitian ini titik pusat awal ditentukan secara random dan didapat titik pusat dari setiap cluster dapat dilihat pada tabel 4.11 dan contoh data sample yang digunakan dapat dilihat pada tabel 4.12.

Tabel 4.11 Pengelompokan Data

Titik Pusat Awal	Data IPS 1
Cluster 1	2.33
Cluster 2	3.50
Cluster 3	2.92

Tabel 4.12 Data IPK mahasiswa Semester 1

IPS 1
3.17
3.67
3.83
3.33
3.50
3.50
3.25
3.75
3.75
3,08
3.50
3.25
...n

Setelah diketahui nilai k dan pusat cluster awal selanjutnya mengukur jarak antara pusat *cluster* menggunakan *Euclidian Distance* pada persamaan 3 untuk mendapatkan matriks jarak yaitu C1, C2, dan C3 sebagai berikut :

1. Perhitungan jarak data pertama dengan pusat *cluster* pertama

   $D_{11} = \sqrt{(3{,}17 - 2{,}33)^2} = 0{,}84$

2. Perhitungan jarak data pertama dengan pusat *cluster* kedua

   $D_{12} = \sqrt{(3.17 - 3.50)^2} = 0.33$

3. Perhitungan jarak data pertama dengan pusat *cluster* ketiga

   $D_{13} = \sqrt{(3{,}17 - 2.92)^2} = 0.25$

Berdasarkan perhitungan yang dilakukan kemudian disatukan untuk mencari jarak terpendek yang ditunjukan pada tabel 4.13.

Tabel 4.13 Perhitungan Jarak Cluster

IPS_1	C1	C2	C3	Jarak Terpendek
3.17	0,84	0,33	0,25	0,25
3.67	1,34	0,17	0,75	0,17
3.83	1,5	0,33	0,91	0,33
3.33	1	0,17	0,41	0,17
3.50	1,17	0	0,58	0
3.50	1,17	0	0,58	0
3.25	0,92	0,25	0,33	0,25
3.75	1,42	0,25	0,83	0,25
3.75	1,42	0,25	0,83	0,25
3.08	0,75	0,42	0,16	0,16
3.50	1,17	0	0,58	0
3.25	0,92	0,25	0,33	0,25

Jarak hasil perhitungan akan dilakukan perbandingan dan dipilih jarak terdekat antara data dengan pusat cluster, jarak ini menunjukkan bahwa data tersebut berada dalam satu kelompok dengan pusat cluster terdekat. Dengan cara membandingkah hasil cluster dan diambil yang paling kecil. Berikut ini akan ditampilkan data matriks pengelompokan group, nilai 1 berarti data tersebut berada dalam group (kelompok data). Hasil pengelompokan data ditampilkan pada tabel 4.14.

Tabel 4.14 Pengelompokan Data *Cluster*

C1	C2	C3
		1
	1	
	1	
	1	
	1	
	1	

Tabel 4.14 (Lanjutan)

C1	C2	C3
	1	
	1	
	1	
		1
	1	
	1	

Setelah diketahui anggota tiap-tiap cluster kemudian dihitung pusat cluster baru menggunakan persamaan 4.

$$C_1 = \frac{2{,}17 + 2{,}50 + 2{,}58 + 2{,}58 + 2{,}42 + 2{,}58 + 2{,}50 + 2{,}50 + ..n}{49} = \frac{117{,}58}{49} = 2{,}40$$

$$C_2 = \frac{3{,}67 + 3{,}83 + 3{,}33 + 3{,}50 + 3{,}50 + 3{,}25 + 3{,}75 + 3{,}75 + ..n}{1525} = \frac{5424{,}56}{1525} = 3{,}56$$

$$C_3 = \frac{3{,}17 + 3{,}08 + 3{,}08 + 3{,}00 + 3{,}17 + 3{,}00 + 3{,}00 + 3{,}17 + ..n}{367} = \frac{1100{,}94}{367} = 3{,}00$$

Berdasarkan perhitungan dihasilkan nilai *centroid* baru. Proses ini terus dilakukan sampai tidak ada perubahan data pada *centroid* baru dan lama. Berikut adalah pengelompakan data menggunakan *K-Means* dengan nilai k ditunjukan pada tabel 4.15.

Tabel 4.15 Klaster Data IPS_1 menggunakan *K-Means*

No	Data IPS_1	Klaster
1	3,17	cluster_3
2	3,67	cluster_2
3	3,83	cluster_2
4	3,33	cluster_2
5	2,17	cluster_1
6	2,50	cluster_1
7	3,08	cluster_3
8	2,58	cluster_1

Tabel 4.15 (Lanjutan)

No	Data IPS_1	Klaster
9	3,50	cluster_2
10	3,50	cluster_2
11	2,58	cluster_1
12	3,08	cluster_3
13	2,42	cluster_1
14	3,00	cluster_3
15	3,17	cluster_3
...	...	...
1941	2,17	cluster_1

4.3.3.2. Metode Elbow

Pengujian Elbow digunakan untuk menuntukan jumlah k paling optimal dengan melakukan perbandingan dari hasil nilai *Sum of Squared Error* pada masing-masing klaster sehingga ditemukan jumlah penurunan tertinggi pada nilai *Sum of Squared Error*. Pengujian dilakukan menggunakan data IPK sem 1 untuk mengetahui jumlah k yang paling optimal yang akan digunakan sebelum proses klasifikasi. Hasil pengujian dengan menghitung nilai *Sum of Squared Error* pada masing-masing klaster ditunjukan pada tabel 4.16.

Tabel 4.16 Pengujian menggunakan Metode Elbow

Klaster	Hasil Sum of Squared Error				
	IPS_1	TPS1	TSKS	MMS	MRS
K2	79,04	43678,08	592,24	1318,53	544,93
K3	39,51	22721,32	233,72	578,66	229,94
K4	24,52	11701,65	92,82	350,57	106,22
K5	15,69	7939,25	45,27	202,03	45,96
K6	11,90	6164,72	19,66	136,71	20,50
K7	8,68	4311,63	2,66	79,13	9,91
K8	6,37	3728,61	0	40,54	3,70

Berdasarkan tabel 4.16 diketahui nilai k terbaik adalah K3 karena memiliki tingkat penurunan terbesar, selanjutnya berdasarkan tabel 4.16 kemudian digambarkan grafik informasi *Sum of Squared Error* pada tiap klaster yang ditunjukan pada gambar 4.1.

Gambar 4.1 Grafik Pengujian Metode *Elbow*

Dari gambar 4.1 menunjukan pengujian menggunakan Metode *Elbow* pada atribut IPS_1 dan diperoleh nilai *Sum of Square Error* yang memiliki penurunan tertinggi pada klaster berjumlah 3 sehingga K3 merupakan klaster yang terbaik.

4.3.3.3. Transformasi Data

Dari proses klasteriasi yang dilakukan menggunakan *K-Means* dan dari hasil pengujian menunjukan jumlah k terbaik adalah 3 kemudian pengelompokan data untuk di *Tranformasi* kan seperti ditampilkan pada tabel 4.17.

Tabel 4.17 Kategori Nilai Atribut

Atribut	Nilai Atribut	Kategori	Tipe Data
IPK Sem 1	2,00-3,08	1	Polynominal
	3,17-3,50	2	
	3,58-4,00	3	
IPK Sem 2	1,63-2,96	1	Polynominal
	3,00-3,42	2	
	3,46-4,00	3	
IPK Sem 3	1,53-2,89	1	Polynominal
	2,90-3,39	2	
	3,41-4,00	3	
IPK Sem 4	1,27-2,84	1	Polynominal
	2,85-3,40	2	
	3,41-4,00	3	
Presensi Sem 1	48-90	1	Polynominal
	91-104	2	
	105-125	3	
Presensi Sem 2	21-86	1	Polynominal
	87-112	2	
	113-126	3	
Presensi Sem 3	6-83	1	Polynominal
	84-114	2	
	115-151	3	
Presensi Sem 4	1-75	1	Polynominal
	76-109	2	
	110-142	3	
Total Sks	66-78	1	Polynominal
	84-92	2	
	94-96	3	
Makul Mengulang	0-1	1	Polynominal
	2-5	2	
	6-19	3	
Makul Remedi	0	1	Polynominal
	1-3	2	
	4-9	3	
Jenis Kelamin	L	1	Polynominal
	P	2	
SLTA	SMA	1	Polynominal
	SMK	2	
	MA	3	
Klasifikasi	<=4 tahun	Tepat waktu	Polynominal (Label)
	>4 tahun	Terlambat	

Sebelum data diklasifikasi menggunakan algoritma C4.5 perlu dilakukan perubahan data interval menjadi kategori menggunakan K-Means karena algoritma C4.5 hanya dapat mengolah data yang bersifat kategori, selain itu dengan melakukan discretize menghasilkan data yang lebih efisien dan akurat. Berikut contoh dataset mahasiswa yang belum diklaster menggunakan *K-Means* pada tabel 4.18.

Tabel 4.18 Dataset Mahasiswa sebelum *Transformasi*

IPS_1	IPS_2	IPS_3	IPS_4	TPS1	TPS2	TPS3	TPS4	TSKS	MMS	MRS
3,17	3,08	3,17	3,15	102	109	130	124	96	0	3
3,67	3	2,23	2,23	102	109	67	50	96	1	2
3,83	3,83	3,86	3,83	102	109	128	120	96	0	0
3,33	3,38	3,32	3,26	102	109	123	106	96	0	0
3,5	3,5	3,58	3,67	102	109	119	118	96	0	0

Berikut contoh dataset mahasiswa yang setelah diklaster menggunakan *K-Means* dengan nilai k paling ideal pada tabel 4.19.

Tabel 4.19 Dataset Mahasiswa setelah *Transformasi*

IPS_1	IPS_2	IPS_3	IPS_4	TPS1	TPS2	TPS3	TPS4	TSKS	MMS	MRS
2	2	2	2	2	2	3	3	3	1	2
3	2	1	1	2	2	1	1	3	1	2
3	3	3	3	2	2	3	3	3	1	1
2	2	2	2	2	2	3	2	3	1	1
2	3	3	3	2	2	3	3	3	1	1

4.4. *Data Mining*

4.4.1. Analisa Algoritma C4.5

Algoritma C4.5 digunakan untuk mengklasifikasi data mahasiswa mahasiswa uang lulus tepat waktu atau terlambat, ada beberapa tahapan yang harus dilalui untuk membentuk sebuah pohon keputusan yaitu :

1. Melakukan perhitungan untuk mencari Entropy dari jumlah kasus berdasarkan jumlah data Tepat Waktu dan Terlambat. Perhitungan tersebut dihitung dengan persamaan 6.

$$Entropy(Total) = \left(-\left(\frac{1632}{1941}\right)Log_2\left(\frac{1632}{1941}\right)\right)+\left(-\left(\frac{309}{1941}\right)Log_2\left(\frac{309}{1941}\right)\right)$$

$$Entropy(Total) = 0{,}632383$$

Setelah ditemukan nilai *Entropy* Total kemudian dilakukan perhitungan *Entropy* dari setiap kategori yaitu IPK_1, IPK_2, IPK_3, IPK_4, TPS2, TPS3, TPS4, TSKS, MMS dan MRS. Berikut contoh perhitungan untuk atribut IPS_1.

$$EntropyIPS\_1(1) = \left(-\left(\frac{232}{327}\right)Log_2\left(\frac{232}{327}\right)\right)+\left(-\left(\frac{93}{327}\right)Log_2\left(\frac{95}{327}\right)\right)$$

$$EntropyIPS\_1(1) = 0{,}869392$$

$$EntropyIPS\_1(2) = \left(-\left(\frac{696}{829}\right)Log_2\left(\frac{696}{829}\right)\right)+\left(-\left(\frac{133}{829}\right)Log_2\left(\frac{133}{829}\right)\right)$$

$$EntropyIPS\_1(2) = 0{,}635347$$

$$EntropyIPS\_1(3) = \left(-\left(\frac{704}{785}\right)Log_2\left(\frac{704}{785}\right)\right)+\left(-\left(\frac{81}{785}\right)Log_2\left(\frac{81}{785}\right)\right)$$

$EntropyIPS\_1(2) = 0{,}47901$

2. Dilakukan perhitungan untuk mencari nilai *Information Gain* dengan menggunakan persamaan 7.

$$Gain(Total, IPS\_1) = 0{,}632383\left(\left(\frac{327}{1941}\right)*0{,}869392\right)-\left(\left(\frac{829}{1941}\right)*0{,}635347\right)-\ldots$$

$$=\left(\left(\frac{785}{1941}\right)*0{,}47901\right)$$

$Gain(Total, IPS\_1) = 0{,}020834$

3. Dilakukan perhitungan nilai dari *Split Information* dengan menggunakan persamaan 8.

$$SplitInfo(Total, IPS\_1) = -\left(\left(\left(\frac{327}{1941}\right)Log_2\left(\frac{327}{1941}\right)\right)+\left(\left(\frac{829}{1941}\right)Log_2\left(\frac{829}{1941}\right)\right)+\ldots\right)$$

$$=\left(\left(\frac{785}{1941}\right)Log_2\left(\frac{785}{1941}\right)\right)$$

$SplitInfo(Total, IPS\_1) = 1{,}485277$

4. Dilakukan Perhitungan *Gain Ratio* dengan membagi *Information Gain* dengan *Split Information* menggunakan persamaan 9.

$$GainRatio(Total, IPS\_1) = \frac{0{,}020834}{1{,}485277}$$

$$= 0{,}014027$$

5. Berdasarkan persamaan diatas kemudian dilakukan perhitungan pada semua atribut untuk mendapatkan *Gain Ratio* tertinggi yang akan digunakan sebagai node akar seperti ditunjukan pada tabel 4.20.

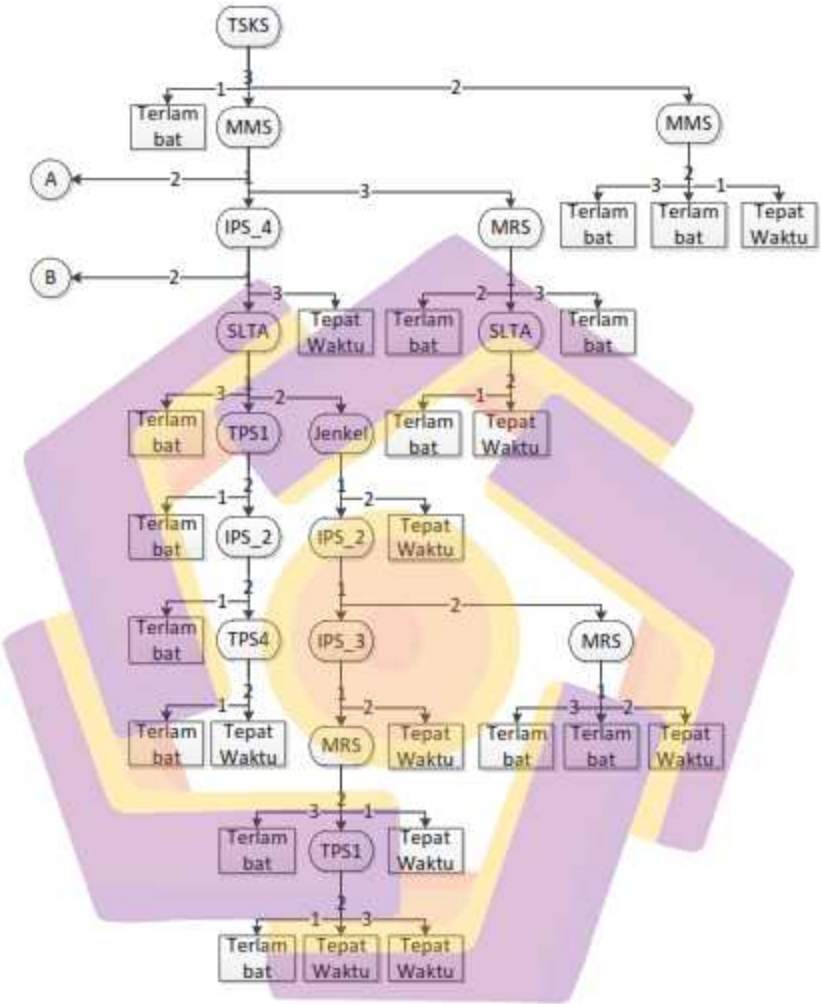
Tabel 4.20 Perhitungan node Akar

Attribut	Nilai	Sum	Tepat waktu	Terlambat	Entropy	Information Gain	Split Info	Gain Ratio
Total		1941	1632	309	0,632383			
IPS 1						0,020834	1,485277	0,014027
	1	327	232	95	0,869392			
	2	829	696	133	0,635347			
	3	785	704	81	0,47901			
IPS 2						0,04902	1,376567	0,0356104
	1	201	116	85	0,982773			
	2	863	708	155	0,679214			
	3	877	808	69	0,397501			
IPS 3						0,053187	1,375986	0,0386534
	1	201	115	86	0,984932			
	2	843	687	156	0,691008			
	3	897	830	67	0,383199			
IPS 4						0,06643	1,360807	0,0488168
	1	187	96	91	0,999484			
	2	864	707	157	0,68381			
	3	890	829	61	0,360449			
TPS1						0,013991	1,11567	0,0125404
	1	98	77	21	0,749595			
	2	1305	1061	244	0,695096			
	3	538	494	44	0,408436			
						0,019399	1,047277	0,0185237
TPS2	1	56	44	12	0,749595			
	2	1315	1059	256	0,711153			
	3	570	529	41	0,373084			
TPS3						0,044784	1,172765	0,0381864
	1	127	70	57	0,992428			
	2	545	417	128	0,786392			
	3	1269	1145	124	0,461709			
						0,055919	1,100567	0,0508094
TPS4	1	117	54	63	0,995727			
	2	468	356	112	0,793901			
	3	1356	1222	134	0,465244			
TSKS						0,008868	0,080605	0,1100229
	1	2	1	1	1			
	2	16	5	11	0,896038			
	3	1923	1626	297	0,620856			

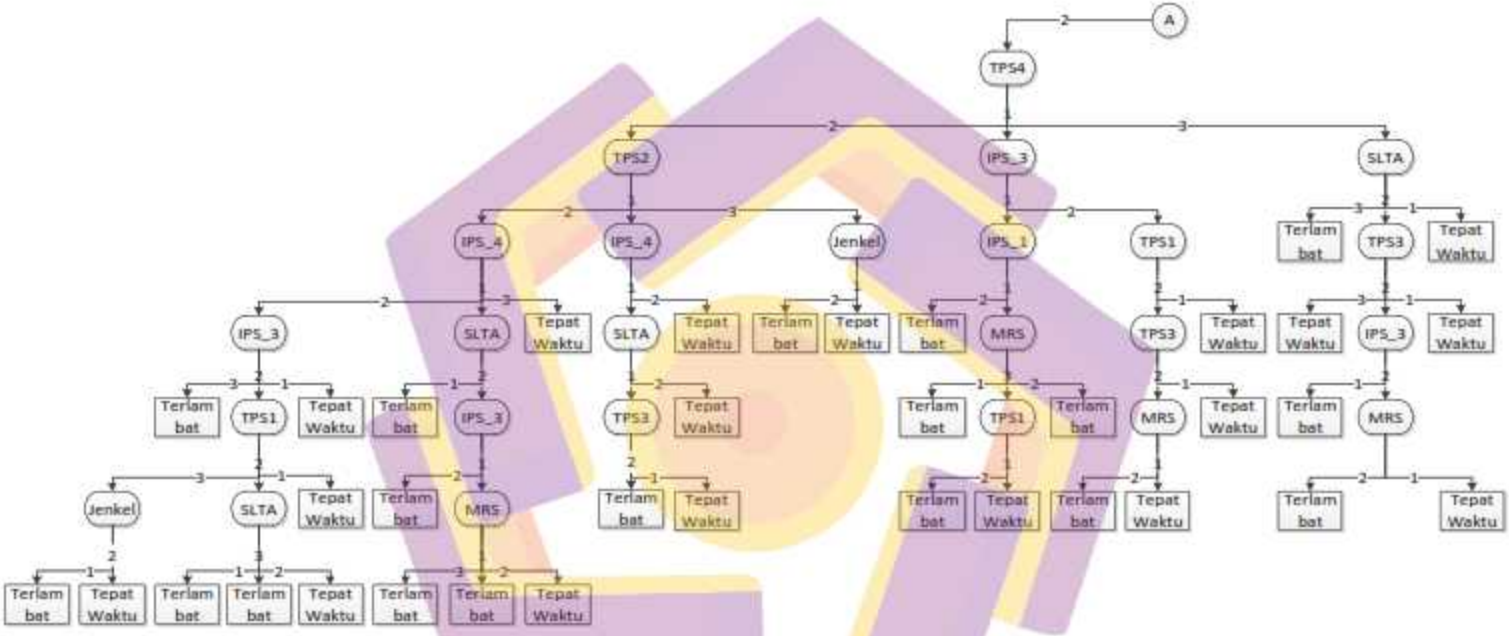
Tabel 4.20 (Lanjutan)

Attribut	Nilai	Sum	Tepat waktu	Terlambat	Entropy	Information Gain	Split Info	Gain Ratio
MMS						0,042071	0,555006	0,0758022
	1	1738	1516	222	0,551188			
	2	166	106	60	0,943877			
	3	37	10	27	0,841852			
MRS						0,031121	0,726295	0,0428484
	1	1634	1422	212	0,556733			
	2	261	194	67	0,821733			
	3	46	16	30	0,932112			
Jenis Kelamin						0,014318	0,62866	0,0227761
	1	1635	1342	293	0,678335			
	2	306	290	16	0,296036			
SLTA						0,017047	0,900667	0,018927
	1	419	311	108	0,823342			
	2	1478	1291	187	0,547823			
	3	44	30	14	0,902393			

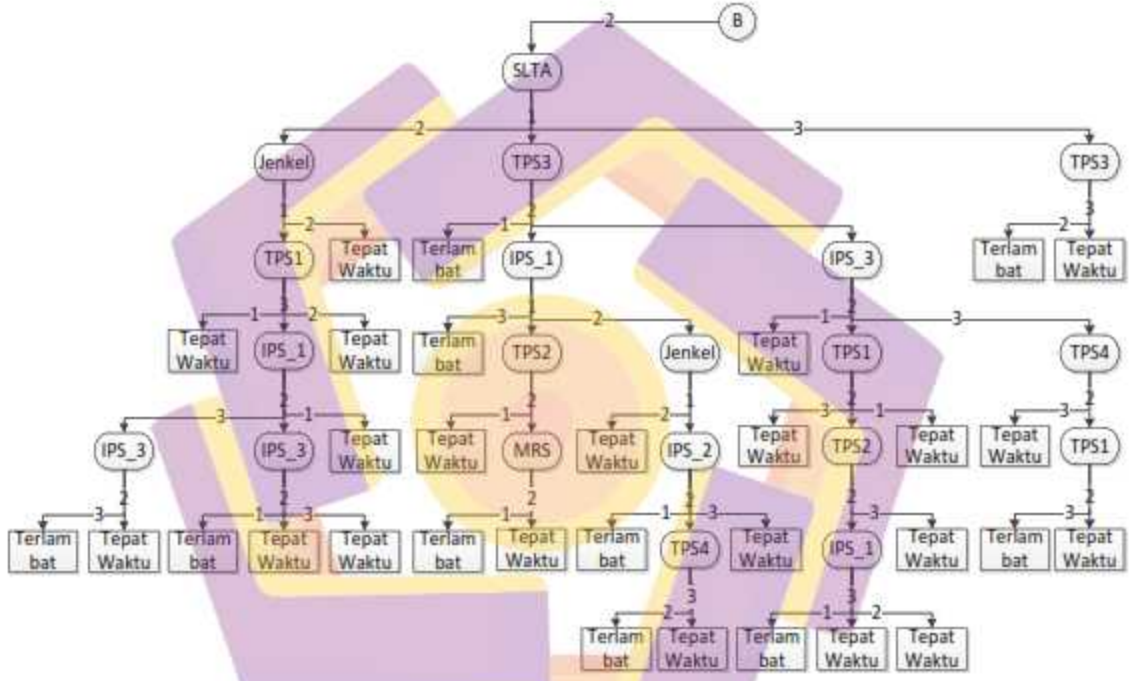
Berdasarkan tabel 4.20 diketahui bahwa atribut TSKS memiliki nilai gain ratio tertinggi sehingga TSKS digunakan sebagai node akar, Pada tahapan selanjutnya dilakukan perhitungan kembali menggunakan cabang TSKS terhadap masing-masing nilai atribut yaitu 1, 2 dan 3 hingga menghasilkan sebuah klasifikasi lulus tepat waktu atau terlambat. Dari hasil analisa menggunakan algoritma C4.5 terbentuk sebuah pohon keputusan yang ditunjukan pada gambar 4.2.



Gambar 4.2 Hasil Pohon Keputusan



Gambar 4.2 (Lanjutan)



Gambar 4.2 (Lanjutan)

Dari hasil pembentukan pohon keputusan pada gambar 4.3 diperoleh 92 *rule* yang dapat digunakan untuk memprediksi kelulusan mahasiswa antara lain:

1. Jika TSKS = 1 Maka Terlambat
2. Jika TSKS = 2 dan MMS = 1 Maka Tepat Waktu
3. Jika TSKS = 2 dan MMS = 2 Maka Terlambat
4. Jika TSKS = 2 dan MMS = 3 Maka Terlambat
5. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 3 Maka Terlambat
6. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 1 dan TPS1 = 1 Maka Terlambat
7. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 1 dan TPS1 = 2 dan IPS_2 = 1 Maka Terlambat
8. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 1 dan TPS1 = 2 dan IPS_2 = 2 dan TPS4 = 1 Maka Terlambat
9. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 1 dan TPS1 = 2 dan IPS_2 = 2 dan TPS4 = 2 Maka Tepat Waktu
10. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 1 dan TPS1 = 3 Maka Tepat Waktu
11. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 1 dan IPS_3 = 1 dan MRS = 1 Maka Tepat Waktu
12. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 1 dan IPS_3 = 1 dan MRS = 2 dan TPS1 = 1 Maka Terlambat
13. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 1 dan IPS_3 = 1 dan MRS = 2 dan TPS1 = 2 Maka Tepat Waktu
14. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 1 dan IPS_3 = 1 dan MRS = 2 dan TPS1 = 3 Maka Tepat Waktu
15. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 1 dan IPS_3 = 1 dan MRS = 3 Maka Terlambat
16. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 1 dan IPS_3 = 2 Maka Tepat Waktu
17. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 2 dan MRS = 1 Maka Terlambat
18. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 2 dan MRS = 2 Maka Tepat Waktu
19. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 1 dan IPS_2 = 2 dan MRS = 3 Maka Terlambat

20. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 1 dan SLTA = 2 dan JenKel = 2 Maka Tepat Waktu
21. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 3 dan IPS_1 = 1 Maka Tepat Waktu
22. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 3 dan IPS_1 = 2 Maka Tepat Waktu
23. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 3 dan IPS_1 = 3 Maka Terlambat
24. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 1 Maka Terlambat
25. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 1 dan TPS2 = 1 Maka Tepat Waktu
26. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 1 dan TPS2 = 2 dan MRS = 1 Maka Terlambat
27. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 1 dan TPS2 = 2 dan MRS = 2 Maka Tepat Waktu
28. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 2 dan JenKel = 1 dan IPS_2 = 1 Maka Terlambat
29. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 2 dan JenKel = 1 dan IPS_2 = 2 dan TPS4 = 2 Maka Terlambat
30. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 2 dan JenKel = 1 dan IPS_2 = 2 dan TPS4 = 3 Maka Tepat Waktu
31. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 2 dan JenKel = 1 dan IPS_2 = 3 Maka Tepat Waktu
32. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 2 dan JenKel = 2 Maka Tepat Waktu
33. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 2 dan IPS_1 = 3 Maka Terlambat
34. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 1 Maka Tepat Waktu
35. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 2 dan TPS1 = 1 Maka Tepat Waktu
36. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 2 dan TPS1 = 2 dan TPS2 = 2 dan IPS_1 = 1 Maka Terlambat
37. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 2 dan TPS1 = 2 dan TPS2 = 2 dan IPS_1 = 2 Maka Tepat Waktu

38. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 2 dan TPS1 = 2 dan TPS2 = 2 dan IPS_1 = 3 Maka Tepat Waktu
39. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 2 dan TPS1 = 2 dan TPS2 = 3 Maka Tepat Waktu
40. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 2 dan TPS1 = 3 Maka Tepat Waktu
41. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 3 dan TPS4 = 2 dan TPS1 = 2 Maka Tepat Waktu
42. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 3 dan TPS4 = 2 dan TPS1 = 3 Maka Terlambat
43. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 1 dan TPS3 = 3 dan IPS_3 = 3 dan TPS4 = 3 Maka Tepat Waktu
44. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 1 dan TPS1 = 1 Maka Tepat Waktu
45. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 1 dan TPS1 = 2 Maka Tepat Waktu
46. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 1 dan TPS1 = 3 dan IPS_1 = 1 Maka Tepat Waktu
47. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 1 dan TPS1 = 3 dan IPS_1 = 2 dan IPS_3 = 1 Maka Terlambat
48. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 1 dan TPS1 = 3 dan IPS_1 = 2 dan IPS_3 = 2 Maka Tepat Waktu
49. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 1 dan TPS1 = 3 dan IPS_1 = 2 dan IPS_3 = 3 Maka Tepat Waktu
50. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 1 dan TPS1 = 3 dan IPS_1 = 3 dan IPS_3 = 2 Maka Tepat Waktu
51. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 1 dan TPS1 = 3 dan IPS_1 = 3 dan IPS_3 = 3 Maka Terlambat
52. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 2 dan JenKel = 2 Maka Tepat Waktu
53. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 2 dan SLTA = 3 Maka Tepat Waktu
54. Jika TSKS = 3 dan MMS = 1 dan IPS_4 = 3 Maka Tepat Waktu
55. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 1 dan IPS_1 = 1 dan MRS = 1 Maka Terlambat
56. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 1 dan IPS_1 = 1 dan MRS = 2 Maka Terlambat

57. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 1 dan IPS_1 = 1 dan MRS = 3 dan TPS1 = 1 Maka Tepat Waktu
58. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 1 dan IPS_1 = 1 dan MRS = 3 dan TPS1 = 2 Maka Terlambat
59. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 1 dan IPS_1 = 2 Maka Terlambat
60. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 2 dan TPS1 = 1 Maka Tepat Waktu
61. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 2 dan TPS1 = 2 dan TPS3 = 1 Maka Tepat Waktu
62. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 2 dan TPS1 = 2 dan TPS3 = 2 dan MRS = 1 Maka Tepat Waktu
63. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 1 dan IPS_3 = 2 dan TPS1 = 2 dan TPS3 = 2 dan MRS = 2 Maka Terlambat
64. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 1 dan IPS_4 = 1 dan SLTA = 1 dan TPS3 = 1 Maka Tepat Waktu
65. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 1 dan IPS_4 = 1 dan SLTA = 1 dan TPS3 = 2 Maka Terlambat
66. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 1 dan IPS_4 = 1 dan SLTA = 2 Maka Tepat Waktu
67. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 1 dan IPS_4 = 2 Maka Tepat Waktu
68. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 1 dan SLTA = 1 Maka Terlambat
69. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 1 dan SLTA = 2 dan IPS_3 = 1 dan MRS = 1 Maka Terlambat
70. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 1 dan SLTA = 2 dan IPS_3 = 1 dan MRS = 2 Maka Tepat Waktu
71. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 1 dan SLTA = 2 dan IPS_3 = 1 dan MRS = 3 Maka Terlambat
72. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 1 dan SLTA = 2 dan IPS_3 = 2 Maka Terlambat
73. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 2 dan IPS_3 = 1 Maka Tepat Waktu
74. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 2 dan IPS_3 = 2 dan TPS1 = 1 Maka Tepat Waktu
75. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 2 dan IPS_3 = 2 dan TPS1 = 2 Maka Tepat Waktu

76. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 2 dan IPS_3 = 2 dan TPS1 = 3 dan JenKel = 1 Maka Terlambat
77. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 2 dan IPS_3 = 2 dan TPS1 = 3 dan JenKel = 2 Maka Tepat Waktu
78. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 2 dan IPS_3 = 3 Maka Terlambat
79. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 2 dan IPS_4 = 3 Maka Tepat Waktu
80. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 3 dan JenKel = 1 Maka Tepat Waktu
81. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 2 dan TPS2 = 3 dan JenKel = 2 Maka Terlambat
82. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 3 dan SLTA = 3 Maka Terlambat
83. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 3 dan SLTA = 1 Maka Tepat Waktu
84. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 3 dan SLTA = 2 dan TPS3 = 1 Maka Tepat Waktu
85. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 3 dan SLTA = 2 dan TPS3 = 2 dan IPS_3 = 1 Maka Terlambat
86. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 3 dan SLTA = 2 dan TPS3 = 2 dan IPS_3 = 2 dan MRS = 1 Maka Tepat Waktu
87. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 3 dan SLTA = 2 dan TPS3 = 2 dan IPS_3 = 2 dan MRS = 2 Maka Terlambat
88. Jika TSKS = 3 dan MMS = 2 dan TPS4 = 3 dan SLTA = 2 dan TPS3 = 3 Maka Tepat Waktu
89. Jika TSKS = 3 dan MMS = 3 dan MRS = 1 dan SLTA = 1 Maka Terlambat
90. Jika TSKS = 3 dan MMS = 3 dan MRS = 1 dan SLTA = 2 Maka Tepat Waktu
91. Jika TSKS = 3 dan MMS = 3 dan MRS = 2 Maka Terlambat
92. Jika TSKS = 3 dan MMS = 3 dan MRS = 3 Maka Terlambat

4.5. Evaluasi Model

4.5.1. *Confusion Matrix*

Dari hasil *Preprocessing Data* menggunakan *K-Means* dan dimodelkan menggunakan algoritma C4.5 selanjutnya dievaluasi menggunakan *Confusion Matrix*. Berikut adalah perhitungan akurasi C4.5 dan *K-Means* ditampilkan pada tabel 4.21.

Tabel 4.21 *Confusion Matrix* Algorimta C4.5 dan *K-Means*

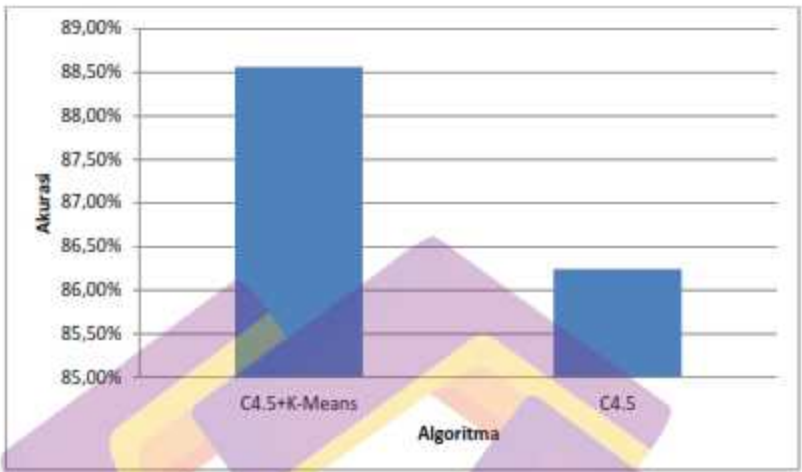
Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1584	48
True Terlambat	174	135

Dari tabel 4.21 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,56%. Selanjutnya dilakukan perhitungan *Confusion Matrix* untuk mengetahui tingkat akurasi C4.5 tanpa proses diskretisasi data menggunakan *K-Means*. Berikut adalah hasil *Confusion matrix* yang terbentuk ditampilkan pada tabel 4.22.

Tabel 4.22 *Confusion Matrix* Algoritma C4.5

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1617	15
True Terlambat	252	57

Dari Tabel 4.22 diketahui bahwa nilai akurasi yang dihasilkan sebesar 86,24%. Berikut adalah perbandingan nilai akurasi antar skenario pengujian ditampilkan pada gambar 4.3.



Gambar 4.3 Perbandingan Akurasi Model

Dari gambar 4.3 diketahui bahwa algoritma C4.5 dan *K-Means* merupakan model terbaik karena memiliki akurasi tertinggi. Untuk mengetahui faktor yang paling mempemgaruhi kelulusan mahasiswa menggunakan algoritma C4.5 dan *K-Means* maka dilakukan pengujian menggunakan beberapa skenario antara lain :

1. Skenario Pengujian 1

   Pengujian dilakukan dengan menghapus atribut IPS_1 dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut IPS_1 ditampilkan pada tabel 4.23.

Tabel 4.23 *Confusion Matrix* Tanpa IPS_1

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1589	43
True Terlambat	183	126

Dari tabel 4.23 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,36%.

2. Skenario Pengujian 2

Pengujian dilakukan dengan menghapus atribut IPS_2 dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut IPS_2 ditampilkan pada tabel 4.24.

Tabel 4.24 *Confusion Matrix* Tanpa IPS_2

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1590	42
True Terlambat	183	126

Dari tabel 4.24 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,41%.

3. Skenario Pengujian 3

Pengujian dilakukan dengan menghapus atribut IPS_3 dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut IPS_3 ditampilkan pada tabel 4.25.

Tabel 4.25 *Confusion Matrix* Tanpa IPS_3

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1587	45
True Terlambat	178	131

Dari tabel 4.25 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,51%.

4. Skenario Pengujian 4

Pengujian dilakukan dengan menghapus atribut IPS_4 dan mengamati *confusion matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya.

Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut IPS_4 ditampilkan pada tabel 4.26.

Tabel 4.26 *Confusion Matrix* Tanpa IPS_4

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1579	53
True Terlambat	177	132

Dari tabel 4.26 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,15%.

5. Skenario Pengujian 5

Pengujian dilakukan dengan menghapus atribut TPS1 dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut TPS1 ditampilkan pada tabel 4.27.

Tabel 4.27 *Confusion Matrix* Tanpa TPS1

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1582	50
True Terlambat	172	137

Dari tabel 4.27 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,56%.

6. Skenario Pengujian 6

Pengujian dilakukan dengan menghapus atribut TPS2 dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut TPS2 ditampilkan pada tabel 4.28.

Tabel 4.28 *Confusion Matrix* Tanpa TPS2

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1586	46
True Terlambat	177	132

Dari tabel 4.28 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,51%.

7. Skenario Pengujian 7

Pengujian dilakukan dengan menghapus atribut TPS3 dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut TPS3 ditampilkan pada tabel 4.29.

Tabel 4.29 *Confusion Matrix* Tanpa TPS3

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1594	38
True Terlambat	188	121

Dari tabel 4.29 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,36%.

8. Skenario Pengujian 8

Pengujian dilakukan dengan menghapus atribut TPS4 dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut TPS4 ditampilkan pada tabel 4.30.

Tabel 4.30 *Confusion Matrix* Tanpa TPS4

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1577	55
True Terlambat	178	131

Dari tabel 4.30 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,00%.

9. Skenario Pengujian 9

Pengujian dilakukan dengan menghapus atribut TSKS dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut TSKS ditampilkan pada tabel 4.31.

Tabel 4.31 *Confusion Matrix* Tanpa TSKS

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1582	50
True Terlambat	175	134

Dari tabel 4.31 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,41%.

10. Skenario Pengujian 10

Pengujian dilakukan dengan menghapus atribut MMS dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut MMS ditampilkan pada tabel 4.32.

Tabel 4.32 *Confusion Matrix* Tanpa MMS

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1582	50
True Terlambat	196	113

Dari tabel 4.32 diketahui bahwa nilai akurasi yang dihasilkan sebesar 87,33%.

11. Skenario Pengujian 11

Pengujian dilakukan dengan menghapus atribut MRS dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya.

Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut MRS ditampilkan pada tabel 4.33.

Tabel 4.33 *Confusion Matrix* Tanpa MRS

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1586	46
True Terlambat	186	123

Dari tabel 4.33 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,05%.

12. Skenario Pengujian 12

Pengujian dilakukan dengan menghapus atribut JenisKelamin dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut JenisKelamin ditampilkan pada tabel 4.34.

Tabel 4.34 *Confusion Matrix* Tanpa JenisKelamin

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1594	38
True Terlambat	186	123

Dari tabel 4.34 diketahui bahwa nilai akurasi yang dihasilkan sebesar 88,46%.

13. Skenario Pengujian 13

Pengujian dilakukan dengan menghapus atribut SLTA dan mengamati *Confusion Matrix* yang terbentuk sehingga dapat hitung tingkat akurasinya. Berikut adalah hasil *Confusion Matrix* tanpa menggunakan atribut SLTA ditampilkan pada tabel 4.35.

Tabel 4.35 *Confusion Matrix* Tanpa SLTA

Aktual	Prediksi	
	Pred. Tepat Waktu	Pred. Terlambat
True Tepat Waktu	1598	34
True Terlambat	220	89

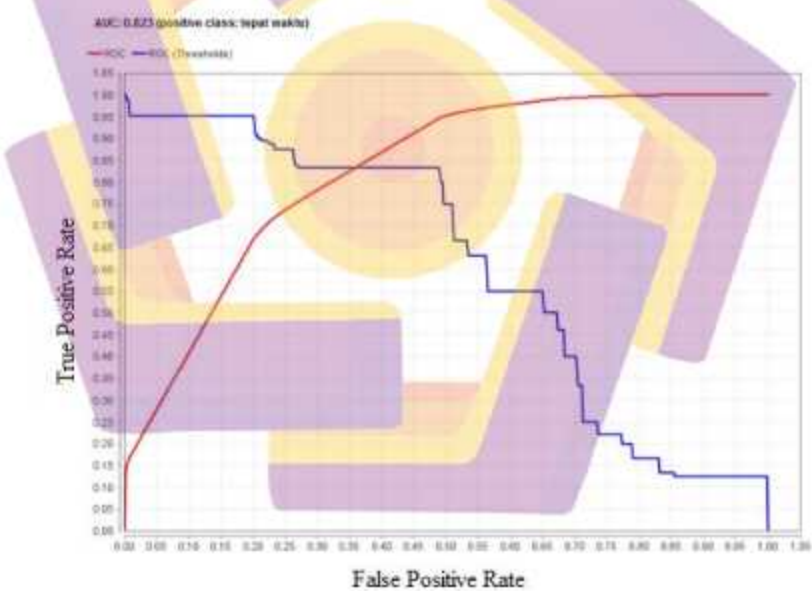
Dari tabel 4.35 diketahui bahwa nilai akurasi yang dihasilkan sebesar 86,91%. Berikut adalah perbandingan nilai akurasi antar skenario pengujian ditampilkan pada gambar 4.4.

Gambar 4.4 Perbandingan Akurasi Pengujian

Dari gambar 4.4 diketahui bahwa faktor yang paling mempengaruhi kelulusan mahasiswa menggunakan algorima C4.5 dan *K-Means* adalah SLTA. Dalam pengujian skenario 13 diperoleh tingkat akurasi terkecil sehingga SLTA memberikan dampak yang paling besar terhadap akurasi.

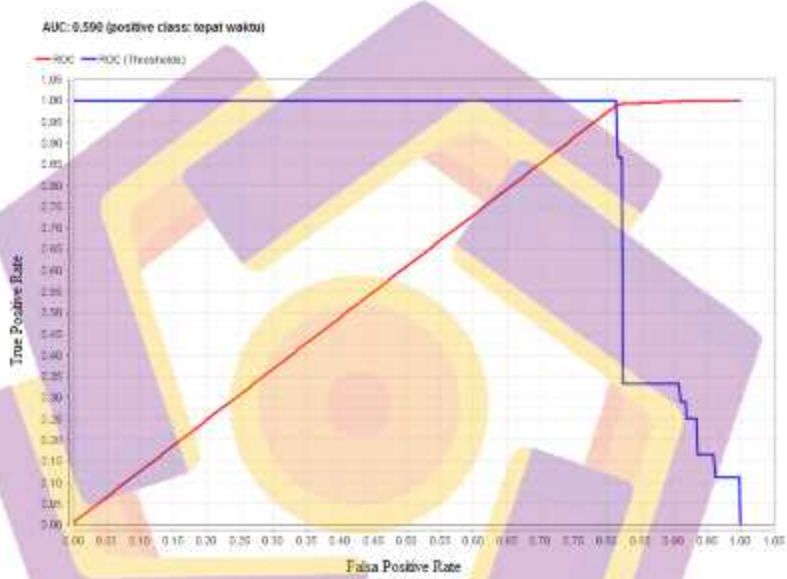
4.5.2. Kurva ROC

Kurva ROC digunakan untuk menunjukan untuk mengenali data positif secara akurat dan tingkat dimana model tersebut salah mengenali data negatif sebagai positif. Dalam Kurva ROC sumbu *vertical* menyatakan *True Positive* sedangkan sumbu *horizontal* menyatakan *False Positive*. Untuk mengukur ketelitian dari suatu model digunakan nilai *Area Under Curve*(AUC) yang menggambarkan perbandingan performa metode yang digunakan. Berikut adalah hasil perhitungan Kurva ROC menggunakan algoritma C4.5 dan *K-Means* ditunjukan pada gambar 4.5.



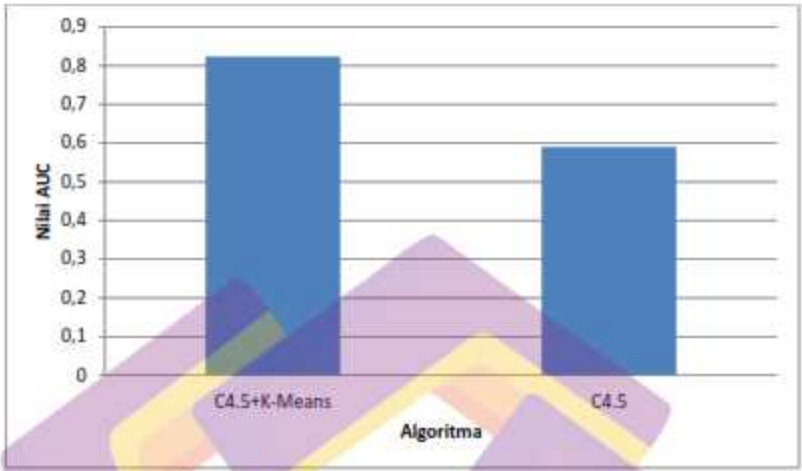
Gambar 4.5 Hasil Kurva ROC Algoritma C4.5 dan *K-Means*

Dari gambar 4.5 menunjukan grafik ROC C4.5 dan *K-Means* memiliki nilai AUC (*Area Under Curve*) sebesar 0,823 yang tergolong *Good Classification*. Selanjutnya dilakukan perhitungan nilai AUC algorima C4.5 tanpa menggunakan *K-Means* yang ditunjukan pada gambar 4.6.



Gambar 4.6 Hasil Kurva ROC Algoritma C4.5

Berikut adalah perbandingan nilai AUC (*Area Under Curve*) antara algoritma C4.5 dan *K-Means* dengan algoritma C4.5 tanpa menggunakan *K-Means* ditampilkan pada gambar 4.7.



Gambar 4.7 Pebandingan Nilai AUC Model

Akurasi kurva ROC yang bernilai sempurna adalah apabila nilai AUC yang mencapai 1,000 sedangkan akurasi bernilai buruk apabila nilai Kurva ROC nya bernilai dibawah 0,500. Dari gambar 4.7 diketahui bahwa algoritma C4.5 dan *K-Means* menghasilkan nilai AUC (*Area Under Curve*) sebesar 0,823 yang tergolong *Good Classification*, sedangan algoritma C4.5 tanpa menggunakan *K-Means* menghasilkan nilai AUC (*Area Under Curve*) sebesar 0,590 yang tergolong *Failure*.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan maka dapat diambil beberapa kesimpulan, antara lain :

1. Faktor yang paling mempengaruhi kelulusan mahasiswa dengan menggunakan algoritma C4.5 dan *K-Means* adalah atribut Pendidikan Menengah (SLTA).
2. Akurasi yang dihasilkan berdasarkan *confusion matrix* dalam memprediksi kelulusan mahasiswa menggunakan algoritma C4.5 dan *K-Means* adalah 88,56%.
3. Nilai *Area Under Curve(AUC)* yang dihasilkan berdasarkan Kurva *Reciever Operating Curve(ROC)* dalam memprediksi kelulusan mahasiswa menggunakan algoritma C4.5 dan *K-Means* adalah 0,823 yang termasuk kedalam *Good Classification*.

5.2. Saran

Berdasakan hasil penelitian yang dilakukan maka saran untuk mengembangkan penelitian adalah dengan melakukan pengujian dengan menambahkan beberapa artibut dan jumlah data yang lebih banyak dan melakukan perbandingan dengan model algoritma modeling yang berbeda untuk mendapatkan nilai akurasi yang lebih baik.

DAFTAR PUSTAKA

PUSTAKA BUKU

Alatubir D. A. V. (2017). Penerapan Algoritma K-Means Untuk Pengelompokan Sekolah Menengah Atas Di Provinsi Daerah Istimewa Yogyakarta Berdasarkan Nilai Daya Serap Ujian Nasional Bahasa Indonesia. Yogyakarta: Universitas Sanata Darma Yogyakarta.

Gonureschu, F. (2011). *Data Mining : Concepts, Models and Tecniques*. New York: Springer - Verlag Berlin Heidelberg.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining Concepts and Techniques 3nd Edition*. Waltham: The Morgan Kaufmann Publisher.

Hermawati, F. A. (2013). *Data Mining*. Yogyakarta: Penerbit Andi.

Kusrini, & Luthfi, E. T. (2009). *Agoritma Data Mining*. Yogyakarta: Penerbit Andi.

Suyanto. (2017). Data Mining Untuk Klasifikasi dan Klasterisasi Data. Bandung: Penerbit Informatika.

Siegel, S. (1994). *STATISTIK NONPARAMETRIK UNTUK ILMU-ILMU SOSIAL*. Jakarta: Gramedia Jakarta.

Sugiyono, P. D. (2015). *STATISTIK NON PARAMETRIS Untuk Penelitian*. Bandung: ALFABETA Bandung.

Turban, E., Aronson, J. E., & Liang, T.-P. (2005). *Decision Support Systems and Intelligent Systems* (7th ed.). USA: Pearson Education.

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Anam, C., & Santoso, H. B. (2018). Perbandingan Kinerja Algoritma C4 . 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa. *Jurnal Ilmiah Ilmu-Ilmu*

Teknik Vol.8 No.1 Edisi Mei 201, *8*(1), 13–19.

Daud, A., Aljohani, N. R., & Abbasi, R. A. (2017). *Predicting Student Performance using Advanced Learning Analytics*. (October). https://doi.org/10.1145/3041021.3054164

Haris, N. A., Abdullah, M., & Hasim, N. (2016). *A Study on Students Enrollment Prediction using Data Mining A Study on Students Enrollment Prediction using Data Mining*. (March). https://doi.org/10.1145/2857546.2857592

Khan, D. M., & Mohamudally, N. (2016). An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm. *Journal Of Computing, Volume 3, Issue 12*, (March).

Pertiwi, A. G., Widyaningtyas, T., & Pujianto, U. (2017). Classification of Province Based on Dropout Rate. *2017 International Conference on Sustainable Information Engineering and Technology (SIET) Classification*, (5), 5–8. https://doi.org/10.1109/SIET.2017.8304173

Purba, W., Tamba, S., & Saragih, J. (2018). *The effect of mining data k-means clustering toward students profile model drop out potential*.

Rajeshinigo, D., & Jebamalar, J. P. A. (2017). *Accuracy Improvement of C4 . 5 using K means Clustering*. *6*(6), 2755–2758.

Singh, I., Sabitha, A. S. A. I., Bansal, A., & Cse, A. (2016). *STUDENT PERFORMANCE ANALYSIS USING CLUSTERING ALGORITHM*. 294–299.

Supriyanti, W., Kusrini, & Amborowati, A. (2016). Perbandingan kinerja algoritma c4.5 dan naive bayes untuk ketepatan pemilihan konsentrasi mahasiswa. *Jurnal INFORMA Politeknik Indonusa Surakarta Vol. 1 Nomor 3 Tahun 2016*, *1*(2012).

Umam, K., Zulfahmi, M., & Nababan, A. A. (2017). PERBANDINGAN RAPID CENTROID ESTIMATION (RCE) — K NEAREST NEIGHBOR (K-NN) DENGAN K MEANS — K NEAREST NEIGHBOR (K-NN). *InfoTekJar (Jurnal Nasional Informatika Dan Teknologi Jaringan)*, *2*(1), 79–89.