



**IMPLEMENTASI BIG DATA ANALYSIS  
MENGUNAKAN ALGORITMA C45 UNTUK SISTEM  
PREDIKSI AKADEMIK MAHASISWA**

**Skripsi**

**Skripsi diajukan sebagai salah satu persyaratan untuk memperoleh gelar  
Sarjana Pendidikan Program Studi Pendidikan Teknik Informatika dan  
Komputer**

**Oleh**

**Willy Pradika**

**NIM.5302414087**

**PENDIDIKAN TEKNIK INFORMATIKA DAN KOMPUTER  
JURUSAN TEKNIK ELEKTRO  
FAKULTAS TEKNIK  
UNIVERSITAS NEGERI SEMARANG**

**2019**

## PERSETUJUAN PEMBIMBING

Nama : Willy Pradika  
NIM : 5302414087  
Program Studi : Pendidikan Teknik Informatika dan Komputer  
Judul : Implementasi Big Data Analysis Menggunakan Algoritma C45  
Untuk Sistem Prediksi Akademik Mahasiswa

Skripsi ini telah disetujui oleh pembimbing untuk diajukan sidang panitia ujian skripsi  
Program Studi Pendidikan Teknik Informatika dan Komputer, Fakultas Teknik  
Universitas Negeri Semarang.

Semarang, 26 Maret 2019

Pembimbing



Arief Arfriandi, S.T., M.Eng.

NIP. 198208242014041001

## HALAMAN PENGESAHAN

Skripsi dengan judul “Implementasi Big Data Analysis Menggunakan Algoritma C45 Untuk Sistem Prediksi Akademik Mahasiswa” telah dipertahankan di depan sidang Panitia Ujian Skripsi Fakultas Teknik UNNES pada 26 Maret tahun 2019.

Oleh

Nama : Willy Pradika

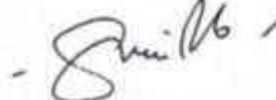
NIM : 5302414087

Program Studi : Pendidikan Teknik Informatika dan Komputer

Panitia :

Ketua Panitia

Sekretaris



Dr.-Ing. Dhidik Prastiyanto, S.T., M.T.,

Ir. Ulfah Mediaty Arief, M.T.

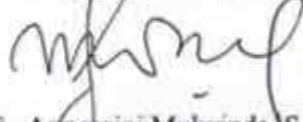
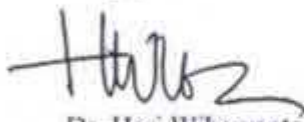
NIP. 197805312005011002

NIP. 196605051998022001

Penguji 1

Penguji 2

Penguji 3/Pembimbing



Dr. Hari Wibawanto, M.T. Ariggraini Mulwinda, S.T., M.Eng.

Arief Arfriandi, S.T., M.Eng.

NIP. 196501071991021001 NIP. 197812262005012001

NIP. 198208242014041001

Mengetahui:

Dekan Fakultas Teknik



Dr. Nur Oudus, M.T

NIP. 196911301994031001

## PERNYATAAN KEASLIAN

Dengan ini saya menyatakan bahwa:

1. Skripsi ini, adalah asli dan belum pernah diajukan untuk mendapatkan gelar akademik (sarjana, magister, dan doktor) baik di Universitas Negeri Semarang (UNNES) maupun perguruan tinggi lain.
2. Karya tulis ini adalah murni gagasan, rumusan, dan penelitian saya sendiri, tanpa bantuan pihak lain, kecuali arahan Pembimbing dan masukan Tim Penguji.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan dalam daftar pustaka.
4. Pernyataan ini saya buat dengan sesungguhnya dan apabila dikemudian hari ditemukan terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya ini, serta sanksi lainnya sesuai dengan norma yang berlaku di perguruan tinggi ini.

Semarang, 26 Maret 2019

Yang membuat pernyataan,



Willy Pradika

NIM. 5302414087

## **MOTTO DAN PERSEMBAHAN**

### **Motto**

- Kita hanya bisa melakukan yang terbaik yang kamu bisa, selebihnya kamu hanya bisa bertawakal serta berdoa kepada Allah SWT. agar diberikan hasil yang terbaik.

### **Persembahan**

Skripsi ini penulis persembahkan kepada :

- Allah SWT yang tak henti-hentinya memberikan kemudahan dan kelancaran dalam penyusunan skripsi ini.
- Kedua orang tua saya, Ibu Endang Sulistyowati dan Bapak Hambali, yang selalu memberikan doa, dukungan, serta semangat.
- Seseorang yang spesial buat saya, Alfida Nur Indah Sari yang selalu memberikan semangat dan keyakinan dalam menyelesaikan skripsi ini.
- Seluruh teman-teman PTIK UNNES angkatan 2014 yang telah dan tengah berjuang bersama-sama menyelesaikan studinya.

## ABSTRAK

Pada kurun waktu antara tahun 1991 hingga 2012, terdapat 801 mahasiswa tahun angkatan 1991 hingga 2005 yang tercatat drop out (DO) atau putus kuliah maupun lulus tidak tepat waktu karena berbagai sebab. Berdasarkan data tersebut bahwa rata-rata kurang lebih 38 mahasiswa tercatat DO atau putus kuliah tiap tahunnya. Mahasiswa DO maupun yang lulus tidak tepat waktu belum dapat diprediksi beberapa waktu sebelumnya. Tujuan dari penelitian ini yaitu membuat model pohon keputusan prediksi akademik menggunakan algoritma C4.5 pada data yang sudah tersimpan di UNNES dan mendapatkan tingkat akurasi hasil prediksi akademik dari model pohon keputusan yang telah terbentuk menggunakan algoritma C4.5.

Proses implementasi pembuatan sistem menggunakan tahapan-tahapan pada model waterfall menggunakan variasi V-Model guna melakukan tahapan uji coba terhadap langkah-langkah yang telah dilalui sebelumnya. Hal ini akan efektif untuk memperkecil kesalahan-kesalahan yang terjadi.

Dari hasil implementasi dan pengujian yang telah dilakukan, diperoleh kesimpulan sebagai berikut. Pada model pohon keputusan yang terbentuk, melalui metode *K-Fold Cross Validation* pada proses uji akurasi prediksi maka model pohon keputusan dengan nilai akurasi tertinggi yaitu 800 dataset terakhir dan sisanya digunakan sebagai data *training* serta mendapatkan atribut berpengaruh pertama yaitu IP Semester 2. Sistem prediksi akademik telah terbentuk dan teruji melalui uji akurasi prediksi berdasarkan model pohon keputusan dari perhitungan algoritma C4.5 dengan tingkat akurasi sebesar 72%.

Kata Kunci: *Algoritma C4.5, Waterfall, Decision Tree*

## **KATA PENGANTAR**

Puji dan syukur penulis ucapkan ke hadirat Allah SWT yang telah melimpahkan rahmat serta karunia-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul “Implementasi Big Data Analysis Menggunakan Algoritma C45 Untuk Sistem Prediksi Akademik Mahasiswa”. Skripsi ini disusun sebagai salah satu persyaratan meraih gelar Sarjana Pendidikan pada Program Studi S1 Pendidikan Teknik Informatika dan Komputer Universitas Negeri Semarang. Penyelesaian skripsi ini tidak lepas dari bantuan berbagai pihak, oleh karena itu penulis menyampaikan ucapan terima kasih kepada :

1. Prof. Dr, Fathur Rokhman, M.Hum., Rektor Universitas Negeri Semarang atas kesempatan yang diberikan kepada penulis untuk menempuh studi di Universitas Negeri Semarang.
2. Dr. Nur Qudus, M.T., Dekan Fakultas Teknik, Dr. Ing. Dhidik Prastiyanto, S.T., M.T., Ketua jurusan Teknik Elektro, Ir. Ulfah Mediaty Arief, M.T., Ketua program studi Pendidikan Teknik Informatika dan Komputer atas fasilitas yang telah disediakan bagi mahasiswa.
3. Bapak Arief Arfriandi, S.T., M.Eng., selaku dosen pembimbing yang telah memberikan bimbingan, arahan, nasehat serta motivasi dalam penulisan karya ini
4. Seluruh dosen Jurusan Teknik Elektro Fakultas Teknik Universitas Negeri Semarang yang telah banyak memberi bekal pengetahuan yang berharga.

5. Teman-teman mahasiswa PTIK Universitas Negeri Semarang angkatan 2014 yang saling memberikan semangat, perhatian, dan penguatan.
6. Berbagai pihak yang telah memberi bantuan untuk penyusunan skripsi ini yang tidak dapat penulis sebutkan satu persatu.

Penulis hanya dapat memanjatkan doa semoga semua pihak yang telah membantu penulis dalam penyusunan skripsi ini mendapatkan pahala dari Allah SWT. Semoga skripsi ini dapat bermanfaat dan memberikan sumbangan yang berarti bagi pihak yang membutuhkan.

Semarang, 26 Maret 2019

Penulis

## DAFTAR ISI

COVER .....	i
PERSETUJUAN PEMBIMBING.....	ii
HALAMAN PENGESAHAN.....	iii
PERNYATAAN KEASLIAN.....	iv
MOTTO DAN PERSEMBAHAN .....	v
ABSTRAK .....	vi
KATA PENGANTAR .....	vii
DAFTAR ISI.....	ix
DAFTAR GAMBAR .....	xii
DAFTAR TABEL.....	xiii
BAB I PENDAHULUAN .....	1
1.1. Latar Belakang .....	1
1.2. Identifikasi Masalah .....	6
1.3. Batasan Masalah.....	7
1.4. Rumusan Masalah .....	7
1.5. Tujuan.....	8
1.6. Manfaat.....	8
BAB II KAJIAN PUSTAKA DAN LANDASAN TEORI .....	10
2.1. Kajian Pustaka .....	10
2.2. Landasan Teori .....	11
BAB III METODE PENELITIAN.....	18
3.1. Waktu dan Tempat Pelaksanaan.....	18

3.2.	Desain Penelitian .....	19
3.3.	Alat dan Bahan Penelitian .....	29
3.5.	Pengujian .....	31
BAB IV HASIL DAN PEMBAHASAN .....		36
4.1	Hasil Penelitian.....	36
BAB V SIMPULAN DAN SARAN .....		56
5.1	Simpulan.....	56
5.2	Saran .....	57
DAFTAR PUSTAKA .....		60
LAMPIRAN.....		62

## DAFTAR GAMBAR

Gambar 3. 1 Waterfall Model .....	<b>Error! Bookmark not defined.</b>	19
Gambar 3. 2 Bagan V-Model .....		20
Gambar 3. 3 Use Case Diagram .....		21
Gambar 3. 4 Flowchart Sistem Prediksi Akademik .....		22
Gambar 3. 5 Halaman Dashboard .....		24
Gambar 3. 6 Halaman Data Record .....		25
Gambar 3. 7 Halaman Model Prediksi .....		25
Gambar 3. 8 Halaman Analisa Data .....		26
Gambar 3. 9 Notasi Pada Flowgraph .....		34
Gambar 3. 10 Konversi Flowchart Menjadi Flowgraph .....		34

## DAFTAR TABEL

Tabel 1. 1 Detail Jumlah Mahasiswa UNNES .....	1
Tabel 3. 1 Jadwal Penelitian.....	18
Tabel 3. 2 Gambaran Tentang Jumlah Kompleksitas .....	35
Tabel 4. 1 Data Perhitungan C4.5 .....	38
Tabel 4. 2 Data Pengelompokkan Uji Validasi.....	44
Tabel 4. 3 Data Hasil Uji Validasi .....	45

# **BAB I**

## **PENDAHULUAN**

### **1.1. Latar Belakang**

Sebagai Perguruan Tinggi Badan Layanan Umum (BLU), pada tahun 2018 jumlah mahasiswa aktif Universitas Negeri Semarang (UNNES) sebanyak kurang lebih 33047 mahasiswa. Mahasiswa aktif tersebar di 8 Fakultas dan Program Pasca Sarjana. Jumlah detil mahasiswa aktif ditunjukkan pada Tabel 1.1.

Tabel 1. 1 Detail Jumlah Mahasiswa UNNES

No	Fakultas/Program	Jumlah Mahasiswa Aktif
1	Fakultas Ilmu Pendidikan	4777
2	Fakultas Bahasa dan Seni	5414
3	Fakultas Ilmu Sosial	2946
4	Fakultas Matematika dan Ilmu Pengetahuan Alam	3695
5	Fakultas Teknik	3627
6	Fakultas Ilmu Keolahragaan	3676
7	Fakultas Ekonomi	4109
8	Fakultas Hukum	1577
9	Program Pasca Sarjana	3226
<b>Total</b>		<b>33047</b>

Pada kurun waktu antara tahun 1991 hingga 2012, terdapat 801 mahasiswa tahun angkatan 1991 hingga 2005 yang tercatat *drop out* (DO) maupun lulus dengan masa studi lebih dari 4 tahun karena berbagai sebab. Berdasarkan data dari tahun 1991 hingga 202 apabila diambil rata-rata tiap tahunnya yaitu rata-rata kurang lebih 38 mahasiswa tercatat DO atau putus kuliah tiap tahunnya. Selain itu, melihat data statistik periode kelulusan mahasiswa pada laman <https://data.unnes.ac.id/index.php/lulusan/masastudi>, terlihat bahwa rata-rata lulusan mahasiswa dengan masa studi diatas 4 tahun pada jenjang S1 dalam periode 2014 sampai 2016 berkisar lebih dari 200 mahasiswa pada tiap-tiap periode kelulusan. Mahasiswa DO maupun yang lulus dengan masa studi lebih dari 4 tahun, yang dapat dikatakan dengan lulus tidak tepat waktu belum dapat diprediksi beberapa waktu sebelumnya, sehingga proses antisipasi yang dilakukan menjadi kurang maksimal. Terkait dengan kelulusan tepat waktu atau dengan masa studi 4 tahun untuk program sarjana menjadi salah satu butir penilaian akreditasi, maka dengan adanya informasi prediksi kelulusan mahasiswa akan menjadikan pengambilan keputusan yang tepat bagi manajemen universitas dalam mengambil langkah strategis guna meningkatkan maupun mempertahankan akreditasinya. Dengan mempertahankan dan meningkatkan akreditasi universitas, maka hal ini akan sejalan dengan proses percepatan UNNES untuk menuju Perguruan Tinggi Badan Layanan Umum yang bereputasi internasional. Bereputasi internasional bermakna universitas yang memiliki citra dan nama baik dalam pergaulan internasional serta menjadi rujukan dalam kegiatan tridarma perguruan tinggi di tingkat internasional. Prediksi kelulusan tepat waktu ini dapat memanfaatkan hasil

pelaksanaan akademik mahasiswa untuk dipelajari karakteristiknya sehingga dapat digunakan untuk membentuk suatu model karakteristik mahasiswa yang lulus tepat waktu maupun tidak tepat waktu.

Dalam menerapkan prediksi terhadap kelulusan mahasiswa terdapat masalah dalam metode yang dilakukan yaitu bahwa prediksi yang dilakukan masih bersifat subyektif dengan melihat perilaku sikap maupun sifat sehari-hari dari mahasiswa yang tampak saja, dengan indikator prediksi yang digunakan tidak sepenuhnya menggunakan data hasil dari pembelajarannya selama ini. Agar dapat lebih terarah dalam melakukan prediksi terhadap kelulusan tepat waktu ini, maka terdapat beberapa metode yang dapat digunakan untuk menunjang tingkat profesionalitas dari prediksi tersebut beberapa diantaranya yaitu *regression*, *neural network*, serta klasifikasi. Berdasarkan penjelasan *Oracle Help Cener* yang terdapat pada website *Oracle Database Online Documentation*, *regression* adalah metode data mining yang digunakan untuk memprediksi angka. Keuntungan, penjualan, tingkat hipotek, nilai rumah, rekaman persegi, suhu, atau jarak dapat diprediksi menggunakan teknik regresi. Misalnya, model regresi dapat digunakan untuk memprediksi nilai rumah berdasarkan lokasi, jumlah kamar, ukuran lot, dan faktor lainnya. Menurut Chauhan, Alok Singh, 2009, *Neural Network* atau Jaringan Saraf Tiruan (JST), sering kali hanya disebut "jaringan saraf" (NN), adalah matematika model atau model komputasi berdasarkan biologis jaringan saraf, dengan kata lain, adalah emulasi dari sistem saraf biologis. Terdiri dari kelompok neuron tiruan yang saling berhubungan dan memproses informasi menggunakan koneksionis pendekatan perhitungan. Dalam kebanyakan kasus, JST adalah sistem adaptif yang mengubah

strukturnya pada informasi eksternal atau internal yang mengalir melalui jaringan selama fase pembelajaran. Pada Untari, Dwi, 2010, menurut Kusriani dan E. T. Luthfi dalam Algoritma Data Mining, Teknik klasifikasi adalah teknik pembelajaran untuk prediksi suatu nilai dari target variabel kategori. Karena dalam prediksi kelulusan mahasiswa ini memiliki dua variabel target yaitu lulus tidak tepat waktu dan dan lulus tepat waktu maka penelitian ini akan menggunakan *data mining* teknik klasifikasi. Terdapat beberapa macam algoritma klasifikasi yang ada diantaranya yaitu metode klasifikasi *nearest neighbour*, *naive bayes* dan *decision tree*.

Dalam penelitiannya mengenai Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian, Sartika Dewi, dkk. 2017 mendapatkan hasil bahwa algoritma klasifikasi *decision tree* merupakan algoritma klasifikasi yang memiliki tingkat akurasi paling tinggi dibandingkan algoritma klasifikasi *naive bayes* dan *nearest neighbour* yaitu mencapai 75.6% pada pengujian yang dilakukan dengan menggunakan mode pengujian *percentage split*. Terdapat beberapa macam metode klasifikasi *decision tree* yaitu ID3, C4.5, dan CART. Menurut penelitian HSSINA, dkk, tentang *A Comparative Study of Decision Tree ID3 and C4.5*, mendapatkan kesimpulan bahwa algoritma C4.5 merupakan algoritma terkuat untuk proses pembuatan *decision tree* jika dibandingkan dengan algoritma lain seperti ID3, C5.0, dan CART. Beberapa penelitian yang meneliti prediksi kelulusan tepat waktu salah satunya Saefulloh, Asep dan Mudjiono, 2013 yaitu Penerapan Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu dan

mendapatkan kesimpulan bahwa implikasi dari temuan penelitian ini mencakup pada dua aspek, yaitu manajerial dan sistem. Pada aspek manajerial dengan memperhatikan hasil pengukuran dan evaluasi maka *Algoritma C4.5* dan *Neural Network* menunjukan algoritma terbaik dalam pengklasifikasian data sehingga metode *Algoritma C4.5* dan *Neural Network* dapat memberikan solusi dalam prediksi kelulusan tepat waktu. Dalam penelitian itu pula disebutkan bahwa algoritma terpilih sebagai algoritma terbaik dalam klasifikasi pemilihan mitra kerja yaitu algoritma C4.5 dan *neural network* yang memiliki tingkat akurasi yang paling tinggi dengan persentase 100% dari kedua algoritma tersebut yang dipergunakan dalam penerapan yaitu algoritma C4.5. Berdasarkan penelitian penelitian diatas maka algoritma C4.5 diputuskan untuk digunakan pada penelitian ini karena algoritma ini lebih kuat dibandingkan algoritma klasifikasi lainnya.

Seiring dengan pesatnya perkembangan teknologi informasi dan komunikasi, mengharuskan penggunaan sistem dalam segala aktifitas di Universitas Negeri Semarang. Seluruh aktifitas akademik maupun identitas mahasiswa UNNES tersimpan dalam database yang terpusat pada Badan Pusat Teknologi Informasi dan Komunikasi yang ada di UNNES. Aktifitas akademik mahasiswa ini secara keseluruhan dikelola menggunakan sistem manajemen akademik seperti SIKADU, SITEDI, MULANG, dan beberapa sistem lainnya. Hal ini sesungguhnya memiliki potensi pada data yang telah tersimpan untuk dapat melakukan prediksi kelulusan dari mahasiswa melihat data aktifitas akademik yang tersimpan ini merupakan data yang sangat banyak sehingga sering disebut dengan istilah *Big Data*. Melihat tingkat efisiensi perhitungan prediksi kelulusan yang dilakukan melibatkan jumlah

data yang sangat banyak ini, maka akan lebih efisien apabila prediksi digunakan dengan melibatkan sebuah sistem informasi. Untuk dapat diakses oleh semua *stack holder* yang terkait menggunakan sistem prediksi kelulusan ini, maka sistem informasi akan dibuat berbasis website. Hal ini juga berkaitan dengan keseluruhan sistem yang terdapat di Universitas Negeri Semarang pada dasarnya menggunakan sistem berbasis website sehingga pengelolaan akses sistem prediksi akademik ini melalui sistem yang telah ada sebelumnya akan dapat lebih mudah. Oleh karena itu, berlatarbelakang dari permasalahan ini penulis mengambil judul penelitian yaitu “Implementasi Algoritma C4.5 pada *Big Data Analysis* untuk Membangun Sistem Prediksi Akademik Mahasiswa”.

## 1.2. Identifikasi Masalah

Berdasarkan latar belakang tersebut, adapun permasalahan yang menjadi dasar pada pembuatan tugas akhir ini antara lain adalah sebagai berikut:

- a. Dari tahun 1991 hingga 2012 terdapat 801 mahasiswa tahun angkatan 1991 hingga 2005 yang tercatat *drop out* (DO) rata-rata kurang lebih 38 mahasiswa,
- b. Aktivitas akademik mahasiswa ini secara keseluruhan dikelola menggunakan sistem manajemen akademik seperti SIKADU, SITEDI, MULANG, dan beberapa sistem lainnya. Hal ini sesungguhnya memiliki potensi pada data yang telah tersimpan untuk dapat melakukan prediksi kelulusan dari mahasiswa melihat data aktivitas akademik yang

tersimpan ini merupakan data yang sangat banyak sehingga sering disebut dengan istilah *Big Data*.

### **1.3. Batasan Masalah**

Agar pembahasan dalam penulisan ini tidak terlalu luas, maka dicantumkan batasan masalah sebagai berikut:

- a. aplikasi yang digunakan untuk implementasi algoritma adalah aplikasi berbasis web,
- b. data mahasiswa berasal dari database yang dikelola oleh BPTIK sebagai badan pusat informasi dan komunikasi di Universitas Negeri Semarang,
- c. jumlah dataset yang digunakan yaitu 4000 data mahasiswa S1 Universitas Negeri Semarang *attribute* yang digunakan adalah ip semester 1, ip semester 2, ip semester 3, Uang Kuliah Tunggal (UKT) serta jumlah sks semester 1, jumlah sks semester 2, jumlah sks semester 3.

### **1.4. Rumusan Masalah**

Bertitik tolak dari latar belakang masalah tersebut di atas, skripsi yang akan dibuat dapat dirumuskan yaitu:

1. Bagaimana membuat model pohon keputusan prediksi akademik menggunakan algoritma C4.5 pada data yang sudah tersimpan di UNNES?

2. Bagaimana tingkat akurasi hasil prediksi akademik dari model pohon keputusan yang telah terbentuk menggunakan algoritma C.45?

### **1.5. Tujuan**

Tujuan dari penelitian implementasi big data analysis menggunakan algoritma C4.5 untuk sistem prediksi akademik mahasiswa yaitu

1. Membuat model pohon keputusan prediksi akademik menggunakan algoritma C4.5 pada data yang sudah tersimpan di UNNES.
2. Mendapatkan tingkat akurasi hasil prediksi akademik dari model pohon keputusan yang telah terbentuk menggunakan algoritma C4.5.

### **1.6. Manfaat**

Beberapa manfaat yang diharapkan dengan adanya sistem prediksi akademik, yaitu:

1. Dengan adanya sistem prediksi akademik ini maka diharapkan sebagai sebuah langkah dalam mempertahankan dan meningkatkan akreditasi Institusi UNNES yang selama ini telah memperoleh akreditasi A dengan melakukan perhatian lebih terhadap kelulusan mahasiswa yang diimplementasikan dalam sistem prediksi akademik mahasiswa.
2. Melalui sistem prediksi akademik ini, maka UNNES dapat memprediksi proses dan ketidaksesuaian akademik untuk mengantisipasi adanya drop out (DO) atau putus kuliah, maupun kelulusan tidak tepat waktu atau dalam kurun waktu lebih dari 4 tahun bagi jenjang S1.

3. Model prediksi akademik ini juga sebagai pendukung pengambilan kebijakan maupun strategi masa depan demi peningkatan kualitas lulusan.

## **BAB II**

### **KAJIAN PUSTAKA DAN LANDASAN TEORI**

#### **2.1. Kajian Pustaka**

Pesatnya pertumbuhan teknologi internet menyebabkan berbagai aktifitas utama di Perguruan Tinggi menggunakan teknologi informasi. Penggunaan teknologi informasi dimulai dari registrasi perkuliahan setiap semester, pembayaran, absensi dan berita acara perkuliahan bahkan proses pembelajaran hingga penilaian. Hal tersebut menghasilkan data yang sangat besar sehingga disebut *big data* tentang pendidikan dan pengajaran, penelitian, pengabdian masyarakat, dan kegiatan penunjang lainnya. *Big data* adalah istilah baru yang digunakan untuk mengidentifikasi dataset yang karena ukurannya yang besar dan sangat kompleks (Fan, Wei). Dengan jumlah mahasiswa yang mencapai puluhan ribu, data-data digital dalam jumlah banyak tersebut meninggalkan tentang apa yang mahasiswa dan akademisi lihat, apa yang mereka baca, keterlibatan dan perilaku mereka, penilaian, maupun tentang kepentingan dan preferensi mereka sehingga menyediakan sejumlah besar data yang dapat ditambang untuk pengalaman pembelajaran (Wagner, 2012).

Beberapa penelitian yang menggunakan algoritma klasifikasi khususnya c4.5, penelitian tersebut antara lain penelitian yang dilakukan oleh Rahman yaitu Algoritma C45 Untuk Menentukan Mahasiswa Penerima Beasiswa (Studi Kasus : Pps Iain Raden Intan Bandar Lampung) menghasilkan suatu kesimpulan yaitu

Bahwa data *mining classification* dengan menggunakan metode pohon keputusan dengan Algoritma C45 untuk menentukan mahasiswa penerima beasiswa pada Program Pascasarjana IAIN Raden Intan Lampung dapat dilakukan. Sejumlah kelebihan dalam penggunaan Algoritma C45 dalam membangun pohon keputusan penerima beasiswa adalah kemampuannya menangani data kontinu maupun data nominal, mengingat bahwa hampir seluruh atribut kriteria penerima beasiswa yang digunakan bertipe data kontinu.

## 2.2. Landasan Teori

### a. Data Mining

Menurut *Gatner Group*, *data mining* adalah suatu proses untuk menemukan informasi yang bermanfaat dari sekumpulan database besar yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik, matematika, kecerdasan buatan, dan *machine learning* (Larose, 2006). *Data mining* dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Larosse, 2006):

#### 1. Deskripsi

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan

kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

## 2. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada penilaian berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh, akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

## 3. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali dalam prediksi nilai dari hasil akan ada di masa mendatang.

Contoh prediksi dalam bisnis dan penelitian adalah:

- a. Prediksi harga beras dalam tiga bulan mendatang
- b. Prediksi presentasi kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikkan

#### 4. Klasifikasi

Di dalam klasifikasi terdapat target variabel kategori. Sebagai contoh penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Kemudian untuk menentukan pendapatan seorang pegawai, dipakai cara klasifikasi dalam data mining.

#### 5. Pengklusteran

Pengklusteran merupakan pengelompokkan *record*, pengamatan atau memperhatikan dan membentuk kelas objek-objek yang mempunyai kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dalam kluster lain. Contoh pengklusteran dalam bisnis dan penelitian adalah:

- a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.
- b. Untuk tujuan audit akuntansi, yaitu melakukan pemisahan terhadap perilaku finansial.
- c. Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar.

## 6. Asosiasi

Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja. Asosiasi mencari kombinasi jenis barang yang akan terjual untuk bulan depan.

### b. Metode *Decision Tree*

Dalam Rismayanti, 2018, Decision tree adalah struktur flowchart yang mempunyai tree (pohon), dimana setiap simpul internal menandakan suatu tes atribut, setiap cabang merepresentasikan hasil tes, dan simpul daun merepresentasikan kelas atau distribusi kelas (Wahyudin, 2009). Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target.

Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang disebut sebagai kriteria dalam pembentukan pohon. Misalkan untuk menentukan main tenis, kriteria yang diperhatikan adalah cuaca, angin, dan suhu. Salah satu atribut merupakan atribut yang menyatakan data solusi per *item* data yang disebut atribut hasil. Banyak algoritma yang dapat dipakai dalam pembentukan pohon keputusan, antara lain ID3, C4.5, CART.

### c. Algoritma C4.5

Menurut Written, 2007, bahwa Algoritma C4.5 merupakan kelompok algoritma pohon keputusan (*decision tree*). Algoritma ini mempunyai input berupa *training samples* dan *samples*. *Training samples* berupa data contoh yang akan digunakan untuk membangun sebuah *tree* yang telah diuji kebenarannya. Sedangkan *samples* merupakan *field-field* data yang nantinya akan kita gunakan sebagai parameter dalam melakukan klasifikasi data, SNATI, 2010. Algoritma C 4.5 adalah salah satu metode untuk membuat *decision tree* berdasarkan training data yang telah disediakan. Algoritma C4.5 merupakan pengembangan dari ID3. Beberapa pengembangan yang dilakukan pada C4.5 adalah sebagai antara lain bisa mengatasi *missing value*, bisa mengatasi *continue data*, dan *praining*. Secara garis besar menurut Written, 2007, langkah-langkah yang dilakukan oleh Algoritma C.45 dalam membentuk pohon keputusan adalah sebagai berikut:

Input : sampel training, label training, atribut

- a. Membuat simpul akar untuk pohon yang dibuat
- b. Jika semua sampel positif, berhenti dengan suatu pohon dengan satu simpul akar, beri tanda (+)
- c. Jika semua sampel negatif, berhenti dengan suatu pohon dengan satu simpul akar, beri tanda (-)

- d. Jika atribut kosong, berhenti dengan suatu bohon dengan suatu simpul akar, dengan label sesuai nilai yang terbanyak yang ada pada label training
  - e. Untuk yang lain, Mulai
    - i. A ---- atribut yang mengklasifikasikan sampel dengan hasil terbaik (berdasarkan *gain ratio*)
    - ii. Atribut keputusan untuk simpul akar ---- A
    - iii. Untuk setiap nilai,  $v_i$ , yang mungkin untuk A
    - iv. Tambahkan cabang di bawah akar yang berhubungan dengan  $A = v_i$
    - v. Tentukan sampel  $S_{v_i}$  sebagai sbset dari sampel yang mempunyai nilai  $v_i$  untuk atribut A
    - vi. Jika sampel  $S_{v_i}$  kosong
      1. Di bawah cabang tambahkan simpul daun dengan label = nilai yang terbanyak yang ada pada label training.
      2. Yang lain tambah cabang baru di bawah cabang yang sekarang C4. (sampel training, label training, atribut – [A].
    - vii. Berhenti
- Mengubah tree yang dihasilkan dalam beberapa *rule*. Jumlah *rule* sama dengan jumlah *path* yang mungkin dapat dibangun dari *root* sampai *leaf node*.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Menurut Ananda, 2014, nilai *gain* dalam algoritma C4.5 adalah perubahan entropi yang terjadi setelah mempartisi data berdasarkan atribut. Entropi adalah pengukuran berdasarkan probabilitas yang digunakan untuk menghitung jumlah ketidakpastian (Ananda, 2014). Entropi dapat digunakan untuk menentukan kemurnian hasil partisi data. Untuk menghitung *gain* digunakan rumus seperti yang tampak pada rumus berikut ini :

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

Keterangan:

$S$  = Himpunan kasus

$A$  = Fitur

$n$  = Jumlah partisi atribut  $A$

$|S_i|$  = Proporsi  $S_i$  terhadap  $S$

$|S|$  = jumlah kasus dalam  $S$

Sementara itu, untuk menghitung nilai entropi dapat digunakan rumus berikut ini :

$$\text{Entropy}(S) = \sum_{i=1}^n -p_i \log_2(p_i) \quad (2)$$

Keterangan:

$S$  = Himpunan kasus

$n$  = Jumlah partisi  $S$

$p_i$  = Proporsi  $S_i$  terhadap  $S$

## **BAB V**

### **SIMPULAN DAN SARAN**

#### **5.1 Simpulan**

Berdasarkan tahapan pengujian yang telah dilakukan, didapatkan sebuah hasil pengujian diantaranya sebagai berikut:

- 1) Pengujian akurasi prediksi menghasilkan nilai akurasi rata-rata yaitu 72%. Pengujian ini menggunakan dataset 4000 data dengan 7 atribut mahasiswa yang digunakan yang meliputi Indeks Prestasi Semester (IP) Semester 1, IP Semester 2, IP Semester 3, Jumlah SKS Semester 1, Jumlah SKS Semester 2, Jumlah SKS Semester 3, serta Uang Kuliah Tunggal (UKT). Melalui metode *K-Fold Cross Validation*, tingkat akurasi tertinggi berdasarkan model prediksi pada tiap-tiap percobaan, maka percobaan dengan k ke 5 mendapatkan nilai akurasi tertinggi yaitu 72% dengan atribut berpengaruh pertama yaitu IP Semester 2.
- 2) Berdasarkan tabel gambaran tingkat kompleksitas menurut Patelia & Vyas, 2014, dapat diambil kesimpulan bahwa berdasarkan nilai kompleksitas yang didapatkan yaitu bernilai 10, menunjukkan bahwa kode program terstruktur dan ditulis dengan baik, tingkat pengujian tinggi, biaya dan upaya sedikit. Pengujian whitebox dilakukan guna mendapati seberapa banyak tahapan yang dilakukan program dalam melakukan perhitungan algoritma, sehingga dapat menindaklanjuti untuk mempersingkat tahapan yang diproses.

Setelah memperoleh hasil dari tahapan pengujian, maka beberapa kesimpulan didapatkan dari penelitian mengenai implementasi *big data analysis* menggunakan algoritma C4.5 pada sistem prediksi akademik, antara lain sebagai berikut:

- a. Pada model pohon keputusan yang terbentuk, melalui metode *K-Fold Cross Validation* pada proses uji akurasi prediksi maka model pohon keputusan dengan nilai akurasi tertinggi yaitu pada model percobaan ke 5 dengan kelompok data validasi yang digunakan yaitu 800 dataset terakhir dan sisanya digunakan sebagai data *training* serta mendapatkan atribut berpengaruh pertama yaitu IP Semester 2.
- b. Sistem prediksi akademik ini dapat digunakan untuk melakukan prediksi terhadap ketepatan kelulusan akademik mahasiswa melalui model pohon keputusan (*decision tree*) yang telah terbentuk dan teruji melalui uji akurasi prediksi berdasarkan model pohon keputusan dari perhitungan algoritma C4.5 dengan tingkat akurasi prediksi yaitu 72%.

## 5.2 Saran

Berdasarkan penelitian yang dilakukan, terdapat beberapa saran yang dapat digunakan untuk penelitian selanjutnya, antara lain:

- a. Pada penelitian selanjutnya diharapkan mampu mengoptimalkan tingkat akurasi prediksi yang dilakukan dengan pemilihan atribut dan kasus yang tepat serta teknik sampling yang tepat guna memperoleh tingkat relevansi data yang tepat mengenai atribut yang sangat mewakili dalam kelulusan mahasiswa.

- b. Pembentukan pohon keputusan pada model prediksi di penelitian ini masih belum optimal sehingga persebaran data atribut yang berpengaruh menjadi sangat banyak. Oleh karena itu, pada penelitian selanjutnya diharapkan mampu mengoptimalkan dan menyederhanakan pohon keputusan model prediksi yang terbentuk.

## DAFTAR PUSTAKA

- Azwanti, Nurul. 2018. *Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada Pt. Capella Dinamik Nusantara Cabang Muka Kuning*, Vol. 13, No. 1.
- Chauhan, Alok Singh, dkk. 2009. *Neural Networks In Data Mining*. India: United Institute of Management, Allahabad, India.
- Defiyanti, Sofi. 2016. *Perbandingan Kinerja Algoritma Id3 Dan C4.5 dalam Klasifikasi Spam-Mail*.
- Fan, Wei. Mining Big Data: Current Status, and Forecast to the Future, *SIGKDD Explorations*, Volume 14, Issue 2.
- Fernandez , Miguel Angel Luque. 2015. *Cross Validation*. London: Cancer Survival Group (LSH&TM).
- HSSINA, B., dkk., 2014, *A Comparative Study of Decision Tree ID3 and C4.5*, Sultan Moulay Slimane University, Morocco.
- Larose D, T., 2006, *Data Mining Methods and Models*, Jhon Wiley & Sons, Inc. Hoboken New Jersey.
- Listiana, Mila, dkk. 2015. *Perbandingan Algoritma Decision Tree (C4.5) Dan Naïve Bayes Pada Data Mining Untuk Identifikasi Tumbuh Kembang Anak Balita (Studi Kasus Puskesmas Kartasura)*. Surakarta: Universitas Muhammadiyah Surakarta.

Nurlifa, Alfian, dkk. 2014. *Analisis Pengaruh User Interface Terhadap Kemudahan Penggunaan Sistem Pendukung Keputusan Seorang Dokter*. ISBN: 978-602-1180-04-4.

Oracle Help Center. 2019. *Regression* di [https://docs.oracle.com/cd/B28359\\_01/datamine.111/b28129/regress.htm#DMCON005](https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/regress.htm#DMCON005) (di akses 24 Mei).

Pressman, Roger S. 2010. *Software engineering : a practitioner's approach*. ISBN 978-0-07-337597-7 — ISBN 0-07-337597-7.

Rahman, Muhammad Arif. 2015. Algoritma C45 Untuk Menentukan Mahasiswa Penerima Beasiswa (Studi Kasus : Pps Iain Raden Intan Bandar Lampung). Jurnal TIM Darmajaya Vol. 01 No. 02.

Rismayanti, 2018. Decision Tree Penentuan Masa Studi Mahasiswa Prodi Teknik Informatika (Studi Kasus: Fakultas Teknik dan Komputer Universitas Harapan Medan). Jurnal Sistem Informasi Volume: 02, Number: 01, April 2018 ISSN 2579-5341 (online).

Saefulloh, Asep dan Mudjiono. 2013. Penerapan Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu. InfoSys Journal, Vol.2 No.1.

Sartika Dewi, dkk. 2017. Perbandingan Algoritma Klasifikasi *Naive Bayes*, *Nearest Neighbour*, dan *Decision Tree* pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian, Vol. 1 No. 2.

Suryana. 2010. *Metodologi Penelitian Model Praktis Penelitian Kuantitatif dan Kualitatif*. Bandung: Universitas Pendidikan Indonesia.

Untari, Dwi. 2010. *Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan Metode Decision Tree C4.5*. Semarang: UDINUS.

Wagner, Ellen; Ice, Phil. 2012. *Data Changes Everything : Delivering on the Promise of Learning Analytics in Higher Education*. Educase Review. [www.educase.edu/ero](http://www.educase.edu/ero).