KAJIAN *DATA MINING* UNTUK MEMPREDIKSI KELULUSAN MAHASISWA DENGAN METODE KLASIFIKASI

(Studi Kasus : STMIK LIKMI BANDUNG)

TESIS

Disusun sebagai salah satu syarat untuk
memperoleh gelar Magister Komputer
dari Sekolah Tinggi Manajemen Informatika dan Komputer LIKMI

Oleh:

ALVIANUS DIAN PRADYTYA

NPM : 2015210107

PROGRAM STUDI PASCASARJANA
MAGISTER SISTEM INFORMASI
SEKOLAH TINGGI MANAJEMEN INFORMATIKA & KOMPUTER
LIKMI
BANDUNG

2018

KAJIAN *DATA MINING* UNTUK MEMPREDIKSI KELULUSAN MAHASISWA DENGAN METODE KLASIFIKASI

(Studi Kasus : STMIK LIKMI BANDUNG)

Oleh:

ALVIANUS DIAN PRADYTYA

NPM : 2015210107

Bandung, 20 Januari 2018
Menyetujui,

Dr. Sani Susanto, Ir., M.Sc.
Pembimbing

PROGRAM STUDI PASCASARJANA
MAGISTER SISTEM INFORMASI
SEKOLAH TINGGI MANAJEMEN INFORMATIKA & KOMPUTER LIKMI BANDUNG
2018

"Bersyukurlah kepada TUHAN, sebab Ia baik!
Bahwasanya untuk selama-lamanya kasih setia-Nya" ...
(Mazmur 107:1)

Tesis ini ku persembahkan untuk:

Kedua orang tua-ku tercinta, Ed. Mugiyana, S.Pd dan M. Sri Murwani;
Mas Peter Eko Doki P.H, S.Kep., Ners;
Mbok Djamini;
juga teristimewa Meriana Candra P, S.Kom., M.Kom

ABSTRAK

KAJIAN *DATA MINING* UNTUK MEMPREDIKSI KELULUSAN MAHASISWA DENGAN METODE KLASIFIKASI
(Studi Kasus: STMIK LIKMI BANDUNG)

Oleh:

ALVIANUS DIAN PRADYTYA
NPM: 2015210107

Sekolah Tinggi Manajemen Informatika & Komputer (STMIK) LIKMI merupakan salah satu perguruan tinggi swasta yang berada di Kota Bandung. Di tengah persaingan antar perguruan tinggi yang ada, terjadi penurunan jumlah penerimaan mahasiswa di mana banyaknya alternatif pilihan perguruan tinggi dengan nilai akreditasi yang cukup baik untuk melanjutkan studi jenjang sarjana. Kualitas lulusan mahasiswa merupakan salah satu kriteria yang ingin dicapai oleh para calon mahasiswa pada saat awal mendaftarkan diri di sebuah perguruan tinggi. Kajian *data mining* dengan mengadopsi metodologi CRISP-DM (*Cross Industry Standard Process for Data Mining*) dinilai sangat sesuai sebagai solusi untuk mendapatkan pola dari mahasiswa yang berpotensi hanya mendapatkan predikat lulus Memuaskan (rata-rata).

Dalam penelitian ini dibandingkan tiga model dalam metode klasifikasi yaitu *Discriminant Analysis, K-Nearest Neighbor,* dan *Decision Tree*. Berdasarkan model prediksi yang dihasilkan, dipilih model *Decision Tree* dengan algoritma CART (*Classification and Regression Trees*) sebagai model prediksi yang dinilai paling baik yakni tingkat akurasi model sebesar 80,11% dengan akurasi memprediksi mahasiswa yang lulus dengan predikat hanya memuaskan (rata-rata) sebesar 80,68%. Subjek (mata kuliah) yang sangat mempengaruhi kelulusan mahasiswa dalam kasus ini secara berurutan adalah: Pengantar Teknologi Informasi, Matematika Bisnis, Pengantar Akuntansi, Pengembangan Sistem Informasi I, Kecapakan Antar Personal, Character Building, dan Statistik I. Pengetahuan yang dihasilkan dari model ini dapat dijadikan *rule* sebagai masukan kepada dosen wali dalam membimbing mahasiswa semester awal (semester I – IV) untuk mengevaluasi prestasi belajar mahasiswa se-dini mungkin agar nantinya mahasiswa tersebut mampu memperoleh predikat lulus *Cumlaude* (Dengan Pujian) yang pada akhirnya mampu mendukung peningkatan nilai akreditasi STMIK LIKMI.

Kata Kunci : *Data Mining*, CRISP-DM, *Discriminant Analysis*, *K-Nearest Neighbor*, *Decision Tree*

ABSTRACT

DATA MINING STUDY FOR PREDICTING STUDENT'S GRADUATE WITH CLASSIFICATION METHOD
(Case Study: STMIK LIKMI BANDUNG)

By:

ALVIANUS DIAN PRADYTYA
NPM: 2015210107

Sekolah Tinggi Manajemen Informatika & Komputer (STMIK) LIKMI is one of the private universities in Bandung. In the midst of competition among the existing universities, there is a decrease in the number of student admissions where the number of alternative choice of universities with accreditation value is good enough to continue the undergraduate study. The quality of graduate students is one of the criteria to be achieved by prospective students at the beginning of enrolling in a college. The data mining study by adopting CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is considered to be very suitable as a solution to get patterns from students who have the potential to only get satisfactory passing predicate (on average).

In this study compared three models in the classification method of Discriminant Analysis, K-Nearest Neighbor, and Decision Tree. Based on the generated prediction model, we selected Decision Tree model with CART (Classification and Regression Trees) algorithm as the best prediction model that is 80.11% model accuracy accuracy predicting graduated students with only satisfactory predicate (average) of 80.68%. The subjects (subjects) that greatly affect the graduation of students in this case are sequentially: Introduction to Information Technology, Business Mathematics, Introduction to Accounting, Information System Development I, Interpersonal Appreciation, Character Building, and Statistics I. Knowledge generated from this model can made the rule as an input to the guardian lecturer in guiding the students of the first semester (semester I - IV) to evaluate the student's early learning achievement so that later the student is able to obtain the predicate to pass Cumlaude (With Praise) which in turn can support the accreditation value increase STMIK LIKMI.

Key words : Data Mining, CRISP-DM, Discriminant Analysis, K-Nearest Neighbor, Decision Tree

KATA PENGANTAR

Puji dan Syukur penulis panjatkan kepada Allah Yang Maha Kuasa atas Rahmat dan Kasih-Nya sehingga penulis dapat menyelesaikan tesis yang berjudul *"**KAJIAN DATA MINING UNTUK MEMPREDIKSI KELULUSAN MAHASISWA DENGAN METODE KLASIFIKASI*** (Studi Kasus: STMIK LIKMI BANDUNG)" yang merupakan salah satu syarat kelulusan yang diajukan dalam program pendidikan pasca sarjana Sistem Informasi di Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) LIKMI Bandung. Pada kesempatan ini penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada berbagai pihak yang telah turut andil dalam penyusunan tesis ini. Selama penyusunan penelitian ini penulis memperoleh banyak dorongan, semangat, bantuan, bimbingan dan nasehat sampai terselesaikannya tugas akhir ini, terutama terimakasih secara khusus kepada :

1. Bapak Dr. Sani Susanto, Ir., M.Sc selaku dosen pembimbing, atas luangan waktu yang disediakan untuk memberikan masukan-masukan serta motivasi yang berguna kepada penulis.
2. Bapak Dr. H. Budi Permana, S.E., Ak., M.Sc selaku Ketua STMIK LIKMI, Bapak Dr. Djajasukma Tjahjahdi, S.E,. M.T selaku Wakil Ketua I STMIK LIKMI, dan Bapak Dr. Hery Heryanto, S.Kom., M.Kom selaku Kepala BAPSI, yang telah memberikan kesempatan, izin serta bantuan kepada penulis untuk melakukan pengambilan data guna keperluan penelitian tesis ini.
3. Rekan-rekan seperjuangan angkatan 2015 program pasca sarjana yang selalu memotivasi penulis selama penelitian hingga penylesaian penyusunan tesis ini.
4. Sanak saudara yang selalu memberikan semangat, doa, harapan dan dorongan kepada penulis untuk dapat menyelesaikan penyusunan tesis ini.
5. Kedua orang tua tersayang, Bapak Eduardus Mugiyana S.Pd dan Ibu Margaretha Sri Murwani, Kakak tercinta Peter Eko Doki Pratiwa Hadi S.Kep., Ners, juga Mbok Djamini tersayang, yang selalu memberikan dukungan doa, materi dan moril kepada penulis selama poses penyusunan tesis ini.

6. Teristimewa untuk Meriana Candra Permatasari, S.Kom., M.Kom yang selalu mendampingi penulis dalam setiap keluh kesah selama pendidikan pasca sarjana hingga terselesaikannya penyusunan tesis ini.
7. Dan seluruh pihak yang tidak dapat penulis sebutkan namanya satu per satu.

Penulis menyadari bahwa masih terdapat banyak kekurangan dalam penulisan ini karena berbagai keterbatasan yang dimiliki penulis. Oleh karena itu dengan rendah hati penulis siap menerima setiap kritik dan saran yang lebih membangun guna menambah pengetahuan dan pengalaman bagi penulis. Besar harapan penulis agar laporan tesis ini dapat bermanfaat dan berguna bagi para pembaca sekalian. Terima kasih, Tuhan memberkati.

Bandung, 20 Januari 2018

Penulis

DAFTAR ISI

ABSTRAK 2
ABSTRACT v
KATA PENGANTAR vi
DAFTAR ISI viii
DAFTAR GAMBAR xi
DAFTAR TABEL xiii
DAFTAR RUMUS xiv
DAFTAR LAMPIRAN xv
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Tujuan Penelitian 3
1.4 Ruang Lingkup Penelitian 4
1.5 Sistematika Penulisan 4
BAB II TINJAUAN PUSTAKA 6
2.1 *Data Mining* 6
2.2 Langkah-langkah *Data Mining* 7
2.3 Fungsi-fungsi dalam *Data Mining* 9
2.4 Fungsi Klasifikasi 10
2.4.1 Model *Discriminant Analysis* (Analisis Diskriminan) 12
2.4.2 Model *K-Nearest Neighbor* 14
2.4.3 Model *Decision Tree* 18
2.5 Evaluasi Performa Model Klasifikasi 22
2.6 Penelitian-penelitian Terdahulu 23
BAB III OBJEK DAN METODOLOGI PENELITIAN 26
3.1 Gambaran Umum Objek Penelitian (STMIK LIKMI) 26
3.2 Metodologi Penelitian 27

3.2.1 Menentukan Topik Penelitian 27
3.2.2 Menentukan Objek Penelitian 27
3.2.3 Menentukan Metode atas Topik Terpilih 27
3.2.4 Studi Literatur 27
3.2.5 Merumuskan Masalah 29
3.2.6 Fase-fase CRISP-DM 29
3.2.7 Kesimpulan dan Saran 29
BAB IV ANALISIS DAN PEMBAHASAN HASIL 30
4.1 Analisis Proses *Data Mining* 30
4.1.1 Fase *Business Understanding* 30
4.1.1.1 Tujuan bisnis 30
4.1.1.2 Penilaian kondisi (*Assess Situation*) 31
4.1.1.3 Tujuan *data mining* 32
4.1.2 Fase *Data Understanding* 32
4.1.2.1 Pengumpulan data awal 32
4.1.2.2 Gambaran data yang dibutuhkan 33
4.1.2.3 *Explore data* 33
4.1.2.4 Verifikasi kualitas data 34
4.1.3 Fase *Data Preparation* 34
4.1.3.1 Pemilihan data dan penentuan variabel 34
4.1.3.2 *Clean data* 35
4.1.4 Fase *Modeling* 36
4.1.4.1 Model *Discriminant Analysis* 36
4.1.4.2 Model *K-Nearest Neighbor* 40
4.1.4.3 Model *Decision Tree* 43
4.1.5 Fase *Evaluation* 47
4.1.5.1 Model *Discriminant Analysis* 47
4.1.5.2 Model *K-Nearest Neighbor* 48
4.1.5.3 Model *Decision Tree* 49

4.1.6 Fase *Deployment* 50
4.2 Pembahasan Hasil Pemodelan 51
4.2.1 Model *Discriminant Analysis* 51
4.2.2 Model *K-Nearest Neighbor* 52
4.2.3 Model *Decision Tree* 54
4.2.4 Perbandingan Hasil Pemodelan 55
BAB V KESIMPULAN DAN SARAN 56
5.1 Kesimpulan 56
5.2 Saran 58
DAFTAR PUSTAKA 59

DAFTAR GAMBAR

Gambar 1. 1 Data Perguruan Tinggi Swasta Wilayah Kopertis-IV 2

Gambar 1. 2 Data Penerimaan Mahasiswa STMIK LIKMI Periode 2012 – 2017 2

Gambar 2. 1 Fase dari CRISP-DM 7

Gambar 2. 2 Algoritma Klasifikasi 11

Gambar 2. 3 Ilustrasi Model *K-Nearest Neigbor* 14

Gambar 2. 4 Langkah-langkah Algoritma *K-Nearest Neighbor* 15

Gambar 2. 5 Visualisasi proses *V-Fold Cross-validation* 16

Gambar 2. 6 Contoh *Decision Tree* dalam Klasifikasi Mamalia 19

Gambar 2. 7 Contoh Model CART 20

Gambar 2. 8 Dataset yang siap diproses 20

Gambar 2. 9 Model yang dihasilkan oleh *software* IBM SPSS Modeler 18.0 22

Gambar 3. 1 Metodologi Penelitian 28

Gambar 4. 1 Urutan Proses Pengerjaan *Data Mining* 30

Gambar 4. 2 Proses akhir dalam pemrosesan model *Discriminant Analysis* 36

Gambar 4. 3 Variabel prediktor terpilih untuk model analisis diskriminan 37

Gambar 4. 4 Kriteria fungsi diskriminan yang terbentuk 38

Gambar 4. 5 *Functions at Group Centroids* 39

Gambar 4. 6 Proses akhir dalam pemrosesan model *K-Nearest Neighbor* 40

Gambar 4. 7 Hasil *V-Fold Cross-validation* 41

Gambar 4. 8 Hasil *euclidean distance* 42

Gambar 4. 9 Diagram *peers* 43

Gambar 4. 10 Variabel prediktor terpilih untuk memodelkan algoritma CART 44

Gambar 4. 11 Proses akhir dalam pemrosesan algoritma CART 44

Gambar 4. 12 Model prediksi kelulusan mahasiswa algoritma CART 46

Gambar 4. 13 Hasil klasifikasi model *Discriminant Analysis* 48

Gambar 4. 14 Hasil klasifikasi model *K-Nearest Neighbor* 49

Gambar 4. 15 Hasil klasifikasi model *Decision Tree* (CART) 49

Gambar 4. 16 *Deployment* model prediksi status kelulusan mahasiswa 50

Gambar 4. 17 Variabel prediktor K-NN 53

DAFTAR TABEL

Tabel 2. 1 Kelebihan dan Kekurangan Model Klasifikasi 11
Tabel 2. 2 Urutan Data Hasil Uji Berdasarkan Jarak Terpendek 17
Tabel 2. 3 Data Latihan Status Kelulusan Mahasiswa 17
Tabel 2. 4 Data Hasil Prediksi K-NN Status Kelulusan Abdul Kodir 18
Tabel 2. 5 *Confusion Matrix* 22
Tabel 2. 6 Klasifikasi Akurasi 23
Tabel 2. 7 Penelitian Terdahulu 23

Tabel 4. 1 Daftar *field* yang dibutuhkan 33
Tabel 4. 2 Variabel Prediktor dan Variabel Target 35
Tabel 4. 3 *Rule* algoritma CART 47
Tabel 4. 4 Daftar mata kuliah berpengaruh pada model *Discriminant Analysis* 51
Tabel 4. 5 Daftar mata kuliah berpengaruh pada algoritma CART 54
Tabel 4. 6 Perbandingan hasil model prediksi 55

DAFTAR RUMUS

Rumus Analisis Diskriminan (2.1) 13

Rumus *Euclidean Distance* (2.2) 16

Rumus Algoritma CART (2.3) 21

DAFTAR LAMPIRAN

Lampiran 1 Persiapan data

Lampiran 2 Tahapan model *Discriminant Analysis*

Lampiran 3 Tahapan model *K-Nearest Neighbor*

Lampiran 4 Tahapan model *Decision Tree* (CART)

BAB I

PENDAHULUAN

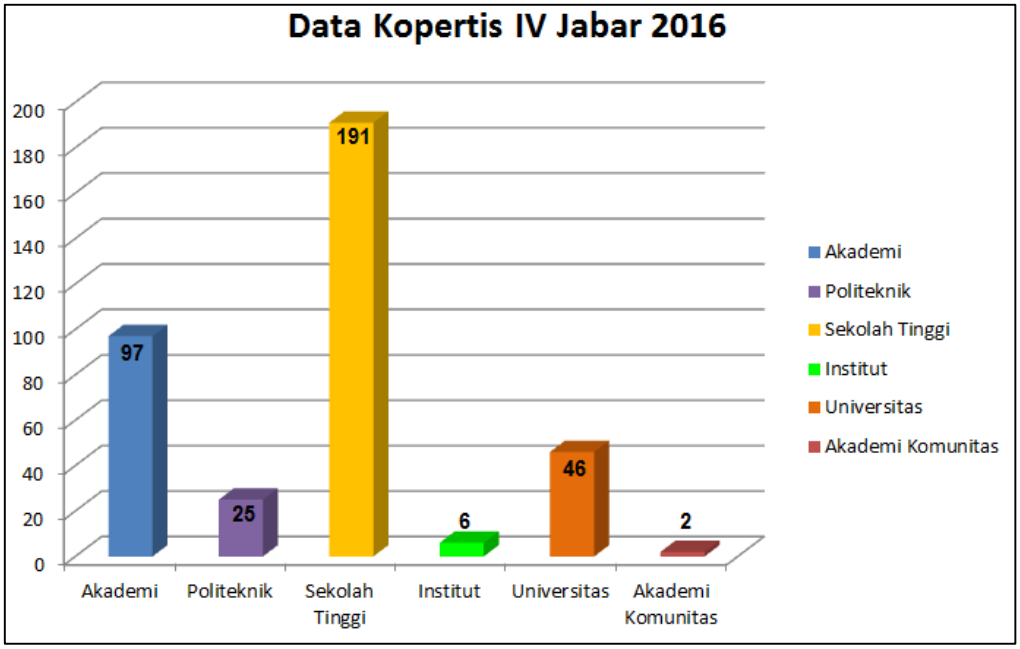
1.1 Latar Belakang

Salah satu kriteria penilaian akreditasi perguruan tinggi yaitu kelulusan mahasiswa, apakah tepat waktu atau tidak (BAN-PT, 2011). Status nilai akreditasi yang diperoleh sebuah institusi perguruan tinggi tersebut menjadi tolak ukur sebagian besar perusahaan (khususnya di Indonesia) untuk menilai kualitas mahasiswa lulusan institusi tersebut, yang pada akhirnya dapat diterima atau ditolak untuk dapat bekerja di perusahaan tersebut. Oleh karena itu, setiap institusi perguruan tinggi terus berjuang membenahi manajemennya untuk dapat meningkatkan status nilai akreditasi.

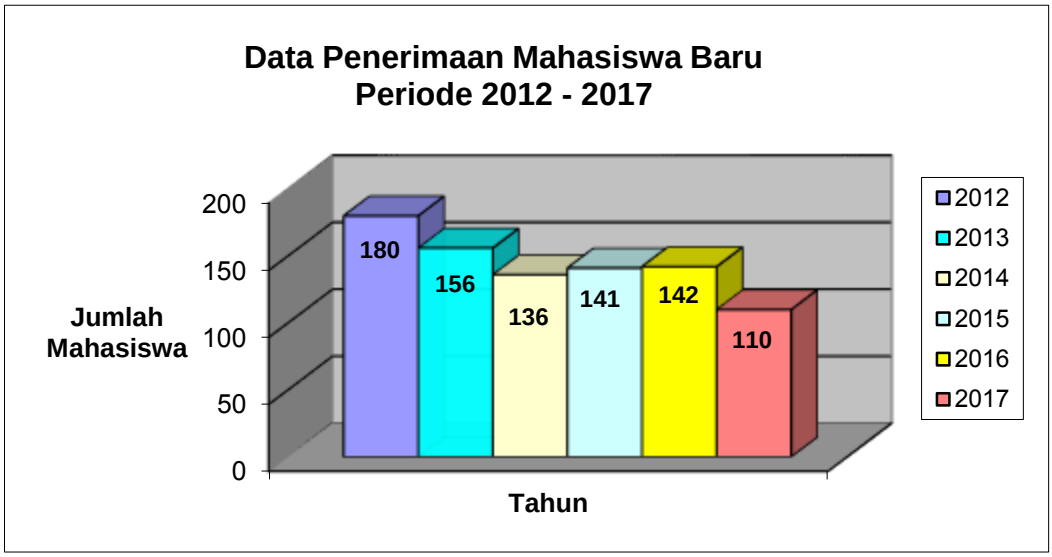
Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) LIKMI, salah satu perguruan tinggi swasta dengan konsep kampus digital, mengemban visi – 2020: menjadi perguruan tinggi terkemuka dalam bidang bisnis dan informatika dengan lulusan yang berkarakter unggul dan dapat memenuhi kebutuhan masyarakat. Hasil penilaian akreditasi BAN-PT yang diperoleh STMIK LIKMI atas penyelenggaraan pendidikan secara menyeluruh pada tahun 2014 yakni terakreditasi B (jenjang Diploma III, Sarjana, dan Pasca Sarjana). Semakin tingginya persaingan dalam dunia kerja, STMIK LIKMI dirasa masih perlu untuk meningkatkan nilai akreditasi sehingga para mahasiswa lulusannya dapat berkompetisi secara unggul untuk memenuhi kebutuhan masyarakat.

Bermunculnya berbagai insitusi perguruan tinggi swasta khususnya yang ada di wilayah Kopertis-IV memberikan alternatif pilihan bagi para calon mahasiswa/i untuk melanjutkan studi ke jenjang yang lebih tinggi. Salah satu dampak langsung terhadap STMIK LIKMI adalah penurunan jumlah pendaftar calon mahasiswa/i dari tahun ke tahun. Untuk mengantisipasi penurunan jumlah pendaftar calon mahasiswa/i ini maka diperlukannya suatu solusi yakni dengan menerapkan penggunaan konsep *data mining* dalam institusi perguruan tinggi (STMIK LIKMI) guna mempertahankan eksistensinya di tengah persaingan yang semakin ketat. Dengan meningkatkan kualitas lulusan mahasiswa STMIK LIKMI setiap waktunya, maka akan membuat STMIK LIKMI memperoleh nilai positif

dari masyarakat sehingga dapat menjadi pilihan prioritas bagi para pendaftar calon mahasiswa/i baru yang akan melanjutkan studi jenjang sarjana.



Gambar 1. 1
Data Perguruan Tinggi Swasta Wilayah Kopertis-IV
(Sumber : www.ristekdikti.go.id)



Gambar 1. 2
Data Penerimaan Mahasiswa STMIK LIKMI Periode 2012 – 2017
(Sumber : BAAK STMIK LIKMI)

Berdasarkan paparan di atas, penulis mencoba melakukan penelitian berupa pengolahan data (nilai) mahasiswa lulusan STMIK LIKMI untuk menghasilkan suatu informasi yang diharapkan dapat membantu dosen wali dalam menyusun strategi yang

tepat untuk menghasilkan lulusan mahasiswa dengan predikat *Cumlaude* (Dengan Pujian), maka penulis mengangkat topik *data mining* dalam penelitian ini dengan judul: "**KAJIAN *DATA MINING* UNTUK MEMPREDIKSI KELULUSAN MAHASISWA DENGAN METODE KLASIFIKASI**".

1.2 Rumusan Masalah

Permasalahan yang menjadi fokus penelitian yaitu:

1. Bagaimana menentukan suatu pola pencapaian prestasi studi lulusan mahasiswa dari tahun ke tahun?
2. Rancangan model seperti apa yang digunakan untuk dapat menganalisa mahasiswa mana yang paling besar kemungkinannya lulus dengan predikat Memuaskan (rata-rata)?
3. Subjek (mata kuliah) mana yang paling mempengaruhi kelulusan mahasiswa untuk memperoleh predikat *Cumlaude* (Dengan Pujian) pada saat lulus?

1.3 Tujuan Penelitian

Adapun tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Menentukan suatu pola pencapaian studi lulusan mahasiswa dari tahun ke tahun sehingga dapat diperoleh suatu fungsi atau *rule* yang menentukan hasil akhir kelulusan mahasiswa.
2. Merancang suatu model untuk dapat menganalisa mahasiswa mana yang paling besar kemungkinannya lulus dengan predikat memuaskan (rata-rata) sedini mungkin, sehingga dapat menjadi acuan dosen wali dalam menyusun strategi yang tepat untuk menghasilkan lulusan mahasiswa dengan predikat *Cumlaude* (Dengan Pujian) pada saat lulus.
3. Dapat memanfaatkan pengelolaan data (nilai) historis mahasiswa secara efektif untuk mengidentifikasi subjek (mata kuliah) apa saja yang paling mempengaruhi kelulusan mahasiswa untuk memperoleh predikat *Cumlaude* (Dengan Pujian) pada saat lulus nantinya.

1.4 Ruang Lingkup Penelitian

Agar penelitian ini dapat memberikan hasil maksimal dan tepat sasaran, maka ruang lingkup bahasan dalam penelitian ini penulis batasi dalam:

1. Penulis melakukan pengelolaan data (nilai) mahasiswa STMIK LIKMI lulusan tahun 2010 – 2014 yang berfokus pada program studi sistem informasi.
2. Penelitian ini menggunakan fungsi Klasifikasi dalam *data mining* dengan membandingkan model *Discriminant Analysis*, model *K-Nearest Neighbor* dan model *Decision Tree*.
3. Penelitian ini menggunakan bantuan program aplikasi IBM SPSS Statistics 24 dan IBM SPSS Modeler 18.0 untuk mendukung analisis data yang tepat dan akurat.
4. Penelitian ini tidak memberikan implementasi berupa rancangan sistem informasi terhadap hasil usulan rekomendasi.

1.5 Sistematika Penulisan

Sistematika penulisan laporan penelitian ini tersusun dalam lima (5) bab, yaitu:

BAB I PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, tujuan penelitian, ruang lingkup penelitian, serta sistematika penulisan yang menguraikan pokok bahasan setiap bab.

BAB II TINJAUAN PUSTAKA

Bab ini berisi teori-teori dasar yang menjadi acuan penulis dalam melakukan penelitian, proses pengolahan data hingga membuat rancangan kerangka model *data mining* sesuai topik dalam penelitian yang dilakukan oleh penulis.

BAB III OBJEK DAN METODOLOGI PENELITIAN

Bab ini berisi mengenai objek dan metode penelitian yang digunakan penulis dalam menyelesaikan penyusunan laporan tesis ini.

BAB IV ANALISIS DAN PEMBAHASAN HASIL

Bab ini berisi tentang pengolahan *dataset* menjadi model prediksi, yaitu berupa sebuah aturan yang dapat dimanfaatkan untuk menghasilkan informasi dan pengetahuan bagi bidang akademik STMIK LIKMI itu sendiri.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari solusi atas permasalahan yang telah dirumuskan oleh penulis, serta saran untuk pengembangan lebih lanjut dari penelitian ini terhadap STMIK LIKMI.

BAB II

TINJAUAN PUSTAKA

2.1 *Data Mining*

Berikut ini merupakan uraian mengenai *data mining*, yang dikemukakan oleh Han dalam bukunya yang berjudul "*Data Mining: Concepts and Techniques 3rd edition*":

Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. (Han, 2012:8)

Menurut Sumathi dan Sivanadam dalam "*Introduction to Data Mining and its Applications*" data mining didefinisikan sebagai:

Data mining is streamlining the transformation of masses of information into meaningful knowledge. It is a process that helps identify new opportunities by finding fundamental truths in apparently random data. The patterns revealed can shed light on application problems and assist in more useful, proactive deciosion making. (Sumathi, 2006:9)

Sedangkan Tan dkk, dalam bukunya yang berjudul "*Introduction to Data Mining*" mendefinisikan data mining sebagai proses pencarian pola-pola yang menarik dan tersembunyi dari suatu kumpulan data yang berukuran besar yang tersimpan dalam suatu basis data, *data warehouse*, atau tempat penyimpanan data lainnya. (Tan, 2006:2)

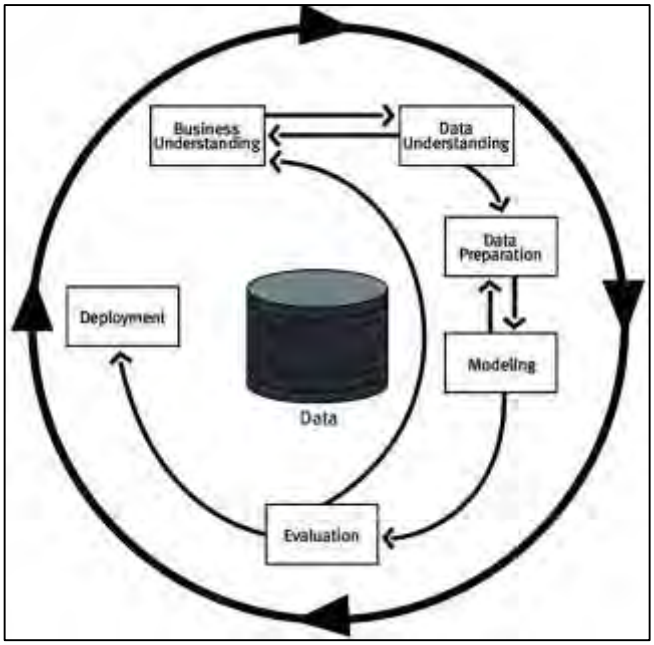
Berdasarkan definisi dari para ahli, dapat dipetik beberapa poin penting mengenai *Data Mining*, diantaranya:

1. *Data Mining* merupakan suatu rangkaian proses untuk mengolah data yang sangat besar yang diperoleh dari *database*.
2. Tujuan *Data Mining* adalah untuk mendapatkan pola-pola tersembunyi yang mungkin memberikan informasi dan pengetahuan bermanfaat.
3. *Data Mining* merupakan dasar dari kecerdasan buatan dalam hal membantu pembuatan keputusan.

Demikian pula halnya mengenai contoh penggunaan *data mining* salah satu contohnya adalah dapat memprediksi kelulusan mahasiswa di suatu institusi perguruan tinggi, apakah mahasiswa tersebut akan lulus dengan predikat *Cumlaude* (Dengan Pujian), Sangat Memuaskan, ataukah Memuaskan (Rata-rata).

2.2 Langkah-langkah *Data Mining*

Dalam mengkaji setiap proses *data mining,* dalam penulisan tesis ini penulis menggunakan siklus CRISP-DM (*Cross Industry Standard Process for Data Mining*). CRISP-DM merupakan suatu metodologi *data mining* yang sudah dikembangkan sejak tahun 1996 oleh tim dari NCR, SPSS, dan Daimler Chrysler. Metodologi ini dipilih karena fleksibilitasnya terhadap unit bidang bisnis apapun, baik untuk pemecahan permasalahan maupun untuk unit penelitian. Terdapat enam (6) fase dalam CRISP-DM yang saling berkaitan antara fase satu dengan fase lainnya, seperti pada gambar 2.1 antara lain sebagai berikut (Chapman, 2000:10):



Gambar 2. 1
Fase dari CRISP-DM
(Sumber: Chapman, 2000:10)

1. *Business Understanding*, merupakan fase pertama yang fokus pada pemahaman tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menerjemakan pengetahuan ini kedalam pendefinisian masalah dalam *data mining*. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.
2. *Data Understanding*, fase ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi

adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. *Data Preparation*, fase ini meliputi semua kegiatan untuk membangun dataset akhir (data yang akan diproses pada tahap *modeling*) dari data mentah. Fase ini dapat diulang hingga beberapa kali. Pada fase ini ini juga mencakup pemilihan tabel, *record*, dan atribut-atribut data, termasuk proses pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan (*modeling*).
4. *Modeling*. Dalam fase ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal. Secara khusus, ada beberapa teknik berbeda yang dapat diterapkan untuk masalah *data mining* yang sama. Pada fase ini masih memungkinan kembali ke tahap sebelumnya.
5. *Evaluation*. Pada fase ini, model sudah terbentuk dan diharapkan memiliki kualitas baik jika dilihat dari sudut pandang analisa data. Pada fase ini akan dilakukan evaluasi terhadap keefektifan dan kualitas model sebelum digunakan dan menentukan apakah model dapat mencapat tujuan yang ditetapkan pada fase awal (*Business Understanding*). Kunci dari fase ini adalah menentukan apakah ada masalah bisnis yang belum dipertimbangkan. Pada akhir dari tahap ini harus ditentukan penggunaan hasil proses *data mining*.
6. *Deployment*. Pada fase ini, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Fase *deployment* dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses *data mining* yang berulang dalam sebuah unit bisnis.

2.3 Fungsi-fungsi dalam *Data Mining*

Mengacu kepada Larose dalam bukunya yang berjudul "*Discovering Knowledge in Data : An Introduction to Data Mining*" terdapat enam fungsi *data mining*, yaitu (Larose, 2005:11):

1. Fungsi Deskripsi (*description)*

   Deskripsi bertujuan untuk mengidentifikasi pola yang muncul secara berulang pada suatu data dan mengubah pola tersebut menjadi aturan dan kriteria yang dapat mudah dimengerti oleh para ahli pada domain aplikasinya. Aturan yang dihasilkan harus mudah dimengerti agar dapat dengan efektif meningkatkan tingkat pengetahuan (*knowledge*) pada sistem. Fungsi deskripsi merupakan fungsi *data mining* yang sering dibutuhkan pada teknik *postprocessing* untuk melakukan validasi dan menjelaskan hasil dari proses *data mining*. *Postprocessing* merupakan proses yang digunakan untuk memastikan hanya hasil yang valid dan berguna yang dapat digunakan oleh pihak yang berkepentingan.

2. Fungsi Prediksi (*prediction*)

   Fungsi prediksi memiliki kemiripan dengan klasifikasi, akan tetapi data diklasifikasikan berdasarkan perilaku atau nilai yang diperkirakan pada masa yang akan datang. Contoh dari fungsi prediksi misalnya untuk memprediksikan adanya pengurangan jumlah pelanggan dalam waktu dekat dan prediksi harga saham dalam tiga bulan yang akan datang.

3. Fungsi Estimasi (*estimation*)

   Estimasi hampir sama dengan prediksi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan 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, berat badan, dan tingkat sodium darah. Hubungan antara tekanan darah

sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi.

4. Fungsi Klasifikasi (*classification*)

   Klasifikasi merupakan proses menemukan sebuah model atau fungsi yang mendeskripsikan dan membedakan data ke dalam kelas-kelas. Fungsi klasifikasi melibatkan proses pemeriksaan karakteristik dari objek dan memasukkan objek ke dalam salah satu kelas yang sudah didefinisikan sebelumnya.

5. Fungsi Pengelompokkan (*clustering*)

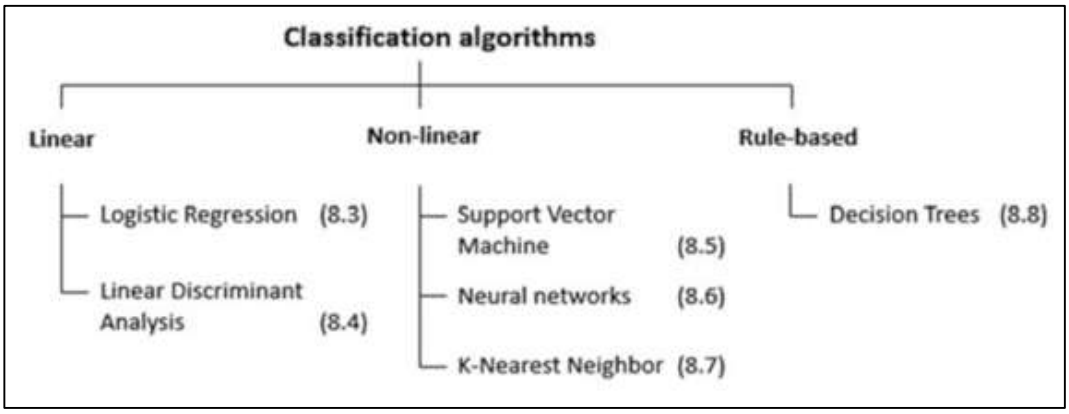
   Fungsi c*lustering* merupakan pengelompokan data tanpa berdasarkan kelas data tertentu ke dalam kelas objek yang sama. Sebuah *cluster* adalah kumpulan *record* yang memiliki kemiripan suatu dengan yang lainnya dan memiliki ketidakmiripan dengan record dalam *cluster* lain. Tujuannya adalah untuk menghasilkan pengelompokan objek yang mirip satu sama lain dalam kelompok-kelompok. Semakin besar kemiripan objek dalam suatu *cluster* dan semakin besar perbedaan tiap *cluster* maka kualitas analisis *cluster* semakin baik.

6. Fungsi Asosiasi (*association*)

   Fungsi asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (*market basket analisys*). Fungsi asosiasi berusaha untuk mengungkap aturan untuk mengukur hubungan antara dua atau lebih atribut.

2.4 Fungsi Klasifikasi

Dalam penelitian tesis ini penulis menggunakan salah satu fungsi dalam *data mining* yaitu fungsi klasifikasi. Terdapat berbagai macam algoritma dengan pendekatan dan teknik yang berbeda untuk menangani data dan menemukan aturan klasifikasi, yang dikelompokkan ke dalam tiga jenis kelas, yaitu: *linear*, *non-linear*, *rule-based*.



Gambar 2. 2
Algoritma Klasifikasi
(Sumber: Wendler, 2016:719)

Sampai saat ini belum ada salah satu model yang mampu mengungguli model lainnya untuk setiap penanganan masalah klasifikasi. Setiap model klasifikasi memiliki kelebihan dan kekurangan masing-masing yang dapat digunakan sesuai dengan kebutuhan dan fungsinya, dapat dilihat pada tabel 2.1

Tabel 2. 1
Kelebihan dan Kekurangan Model Klasifikasi

Model Klasifikasi	Kelebihan (+)	Kekurangan (-)
Logistic Regression	1. Kuat (hampir tidak terpengaruh oleh *noisy data*) 2. (Umumnya) presisi tinggi 3. Bekerja dengan ukuran sampel yang kecil 4. Interpretasi probabilistik dari hasil	1. Bermasalah dengan tingginya *collinearity* 2. *Gaussian* mendistribusikan residu
Linear discriminant analysis	1. Sangat cepat 2. Bekerja dengan sampel kecil 3. Optimal jika asumsi data terpenuhi	1. Kebutuhan persiapan data 2. Kelas *Gaussian* didistribusikan 3. Sensitif terhadap *outlier* 4. Hanya berlaku untuk masalah linier
Support vector machine	1. Kuat (hampir tidak terpengaruh oleh *noisy data*) 2. Presisi tinggi	1. Lambat dalam latihan 2. Hasil sulit dipahami 3. Sensiti, berkenaan dengan pilihan fungsi kernel 4. Memilih parameter yang salah dapat berisiko *overfitting*
Neural Network	1. Bagus untuk sampel besar 2. Dapat menangani interaksi yang sangat kompleks antar variabel 3. Nonparametrik. Tidak ada asumsi distribusi yang dibutuhkan	1. Tidak bisa menangani terlalu banyak variabel 2. Hasil sulit dipahami 3. Solusi tidak selalu optimal

Model Klasifikasi	Kelebihan (+)	Kekurangan (-)
	4. Tahan terhadap data yang rusak	
K-Nearest Neighbor	1. Kecepatan (tidak dibutuhkan pelatihan) 2. Konsep kompleks bisa dipelajari melalui prosedur sederhana	1. Kinerja tergantung dari jumlah dimensi (*Curse of dimensionality*) 2. Masalah dengan data yang sangat tidak seimbang
Decision Tree	1. Tidak masalah dengan outlier 2. Hasil dan aturan bisa dimengerti 3. Bekerja dengan banyak tipe data 4. Cepat dalam prediksi 5. Tidak memerlukan asumsi pada distribusi variabel yang dibutuhkan	1. Lambat dalam melatih model 2. Cenderung untuk overfit 3. Mendeteksi lokal, tidak global 4. Lebih menyukai variabel dengan banyak kategori atau data numerik

(Sumber: Wendler, 2016:720)

Melihat kelebihan dan kekurangan masing-masing model klasifikasi, maka dalam penelitian tesis ini penulis membandingkan model *Discriminant Analysis*, model *K-Nearest Neighbor* dan model *Decision Tree* untuk digunakan dalam membuat sebuah aturan baru dalam mengkaji *data mining* untuk memprediksi kelulusan mahasiswa di STMIK LIKMI.

2.4.1 Model *Discriminant Analysis* (Analisis Diskriminan)

Discriminant Analysis atau Analisis Diskriminan merupakan salah satu teknik analisis multivariat dalam metode perhitungan statistika. Menurut Singgih Santoso dalam bukunya yang berjudul "*Statistik Multivariat dengan SPSS*", analisis multivariat didefinisikan sebagai berikut:

"Analisis multivariat adalah analisis multivariabel dalam satu atau lebih hubungan. Analisis ini berhubungan dengan semua teknik statistik yang secara simultan menganalisis sejumlah pengukuran pada individu atau objek." (Santoso, 2017:1)

Singgih Santoso mendefinisikan analisis diskriminan sebagai teknik multivariat yang termasuk *dependece method*, yakni adanya variabel dependen dan independen. Dengan demikian ada variabel yang hasilnya tergantung dari data variabel independen. Ciri khusus data variabel dependen harus berupa data kategori, sedangkan data variabel independen berupa data non-kategori (Santoso, 2017).

Adapun tujuan dari analisis diskriminan yaitu:

1. Ingin mengetahui apakah ada perbedaan yang jelas antar grup pada variabel dependen? Dengan kata lain, apakah ada perbedaan antara anggota grup satu dengan anggota grup lainnya?
2. Jika ada perbedaan, variabel independen manakah pada fungsi diskriminan yang membuat perbedaan tersebut?
3. Membuat fungsi atau model diskriminan
4. Melakukan klasifikasi terhadap objek (dalam terminologi SPSS disebut baris), apakah suatu objek termasuk ke dalam grup satu, grup dua, atau lainnya.

Untuk memodelkan analisis diskriminan digunakan persamaan (2.1) berikut:

$$Y_1 = X_1 + X_2 + X_3 + \cdots X_n \quad (2.1)$$

(Sumber: Santoso, 2017:4)

Keterangan:

X_1 : Variabel independen (X_1, X_2, danseterusnya) yang merupakan data metrik, yakni data berjenis interval atau rasio. Misal: Usia seseorang, tinggi badan, dan sebagainya.

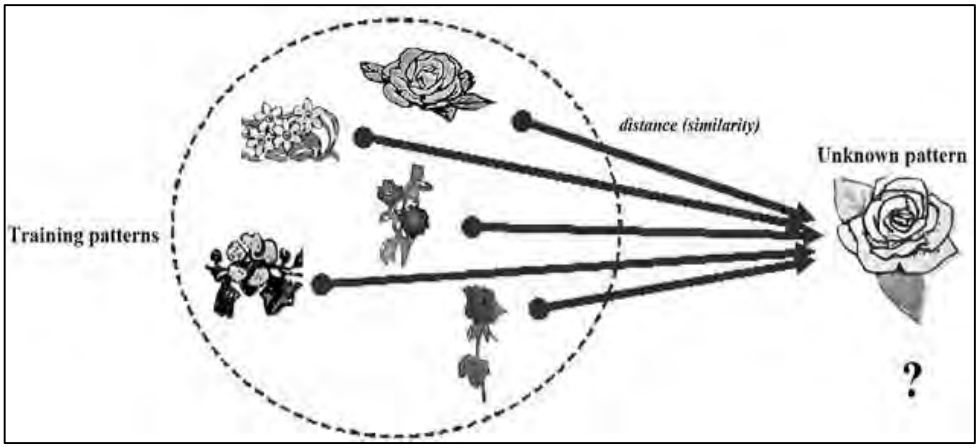
Y_1 : Variabel dependen yang merupakan data kategorikal atau nominal. Misal: kategori seseorang yang banyak minum (kode 1), sedangkan kategori seseorang yang sedikit minum (kode 0), dan sebagainya.

Model analisis diskriminan yang hanya terdiri dari dua kategori saja (misal: banyak minum dan sedikit minum) disebut dengan *Two-Group Discriminant Analysis*, sedangkan jika terdiri lebih dari dua kategori disebut dengan *Multiple Discriminant Analysis*. Berdasarkan keterangan di atas, terdapat perbedaan dalam penempatan data yang sekilas mirip, seperti halnya usia seseorang (dalam tahun). Jika usia yang dimaksud adalah berupa angka (14 tahun, 20 tahun, dan seterusnya), maka data tersebut adalah data metrik (rasio) yang otomatis diperlakukan sebagai Variabel Independen. Namun, jika usia yang dimaksud adalah penggolongan dan dimasukkan ke dalam kategori-kategori tertentu, misal: usia antara 15-20 tahun digolongkan Remaja, usia antara 21-40 tahun digolongkan

Dewasa, dan usia diatas 40 tahun digolongkan Lansia, maka apabila terdapat data seseorang dengan usia 25 tahun tidak akan langsung ditulis dengan angka 25, melainkan ditulis ke dalam kategori Dewasa. Data hasil kategorisasi ini merupakan data nominal dan termasuk Variabel Dependen.

2.4.2 Model *K-Nearest Neighbor*

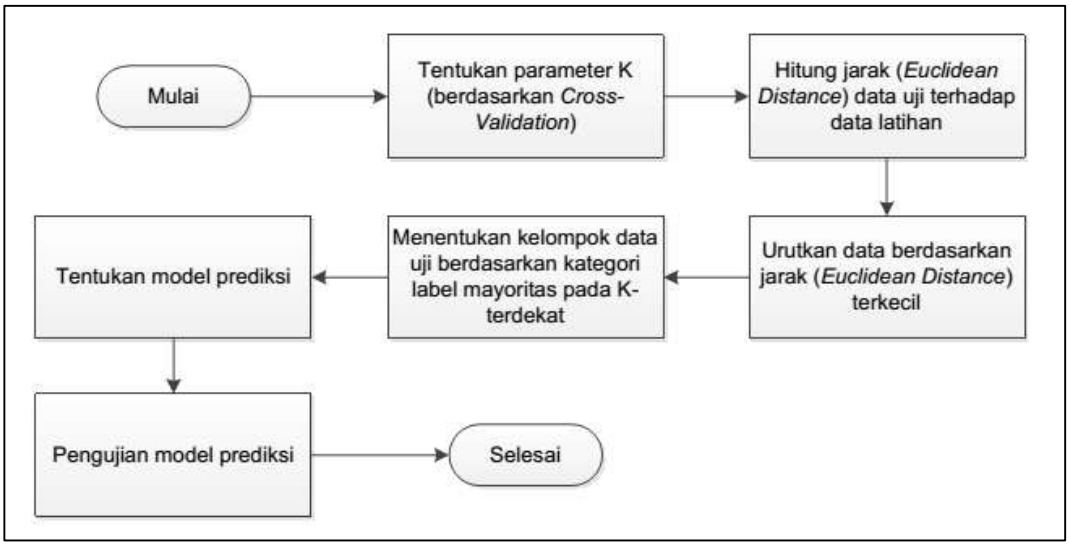
Model *K-Nearest Neighbor* (K-NN) merupakan metode klasifikasi yang mengelompokkan data baru berdasarkan jarak terdekat (paling mirip) dengan objek pada data baru (*instance-based learning*) (Larose, 2015). Dalam area pengenalan pola, K-NN mewakili metode klasifikasi, dimana objek baru diberi label berdasarkan objek terdekatnya-*k* (Gorunescu, 2011). Secara sederhana penjelasan model K-NN dapat dilihat pada gambar 2.3.



Gambar 2. 3
Ilustrasi Model *K-Nearest Neigbor*
(Sumber: Gorunescu, 2011)

Pada prinsipnya, pola dari data latihan (gambar yang berada dalam lingkaran garis putus-putus) akan dijadikan acuan dalam mengklasifikasikan objek/ data baru berdasarkan hasil perhitungan jarak (seberapa jauh kesamaan/ kemiripan pola).

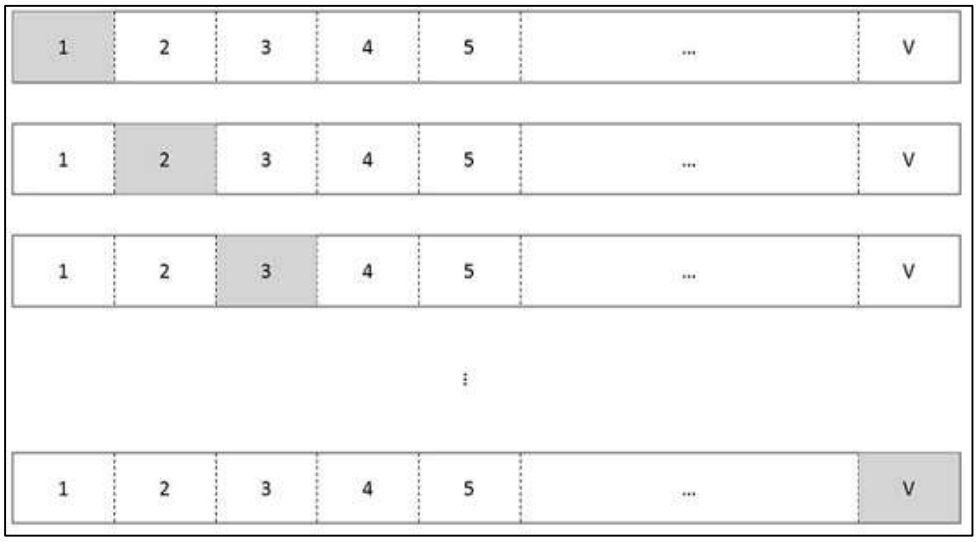
Langkah-langkah algoritma K-NN dapat diuraikan secara sederhana ke dalam gambar 2.4.



Gambar 2. 4
Langkah-langkah Algoritma *K-Nearest Neighbor*
(Sumber: Wendler, 2016)

1. Menentukan parameter (nilai) *K*

Dalam model K-NN diperlukan nilai *K* untuk membatasi kelompok hasil perhitungan *nearest neighbor* (tetangga terdekat). Parameter (nilai) *K* yang terbaik untuk model K-NN tergantung pada data. Secara umum bahwa nilai *K* yang tinggi akan mengurangi dampai *noise* pada klasifikasi, tetapi mempengaruhi batasan antar klasifikasi menjadi kabur (Larose, 2015). Mengacu pada (Wendler, 2016), dalam menentukan parameter (nilai) *K* terbaik dalam kisaran tertentu dapat menggunakan pengaturan proses *Cross-validation* dengan teknik *V-Fold Cross-validation*. Dalam teknik ini dataset dibagi menjadi sejumlah V-buah lipatan (partisi) secara acak. Kemudian dilakukan sejumlah V-kali eksperimen, dimana masing-masing eksperimen menggunakan data partisi ke-V sebagai data testing dan memanfaatkan sisa partisi lainnya sebagai data training, seperti terlihat pada gambar 2.5 kotak abu-abu merupakan kumpulan tes dan sisanya adalah himpunan bagian dari data latihan dalam setiap kasus.



Gambar 2. 5
Visualisasi proses *V-Fold Cross-validation*.
(Sumber: Wendler, 2016:889)

2. Menghitung jarak terdekat

Untuk menghitung seberapa jauh jarak tetangga terdekat, syarat utama semua variabel yang terkait harus bertipe numerik, umumnya dapat menggunakan rumus *Euclidean Distance* pada persamaan (2.2) (Wendler, 2016):

$$d\,(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2.2)$$

(Sumber: Wendler, 2016:880)

Keterangan :

d : jarak terdekat (*distance*)

x : data latihan

y : data uji

n : banyak variabel (variabel data latihan = variabel data uji)

x_i : variabel data latihan baris ke – i

y_i : variabel data uji baris ke – i

i : 1, 2, 3 ... dst

Apabila tipe data variabel bukan numerik, maka langkah yang harus dilakukan yaitu membuat transformasi ke dalam bentuk numerik terlebih dahulu, misal

variabel jenis kelamin (Pria dan Wanita) dapat terelbih dahulu ditransformasikan ke dalam bentuk numerik, yakni nilai 0 untuk Wanita dan nilai 1 untuk Pria atau dapat pula menggunakan angka lain yang sederhana, baru kemudian dapat dihitung jaraknya menggunakan rumus di atas. Selain menggunakan *Euclidean Distance*, terdapat beberapa rumus untuk mengukur jarak tetangga terdekat diantaranya: *Mikwoski Distance* dan *Manhattan Distance* (Witten, 2011:587).

3. Mengurutkan data berdasarkan jarak minimum terhadap *K*

Setelah jarak terdekat data uji terhadap data latihan telah diperoleh, maka selanjutnya adalah mengurutkan nilai jarak tersebut menggunakan ranking jarak dengan nilai paling kecil, dapat dilihat pada tabel 2.2.

Tabel 2. 2
Urutan Data Hasil Uji Berdasarkan Jarak Terpendek

No	Nama	Distance	Ranking
1	d1,d6	1,66	2
2	d2,d6	2,02	5
3	d3,d6	1,53	1
4	d4,d6	1,76	3
5	d5,d6	1,89	4

4. Menentukan kelompok data uji

Pada tahap ini dilakukan penyesuaian kelompok data uji berdasarkan kategori label mayoritas yang didapat dari *K*-terdekat, dapat dilihat pada tabel 2.3 dan tabel 2.4

Tabel 2. 3
Data Latihan Status Kelulusan Mahasiswa

NO	NAMA	IPK					KELULUSAN
		SMT 1	SMT 2	SMT 3	SMT 4	SMT 5	
1	Ahmad	3	4	3,5	4	3	Tepat Waktu
2	Antoni	4	4	4	3	3	Tepat Waktu
3	Mahmud	2	2,5	4	3	4	Terlambat
4	Syam Aji	3	4	4	3	3	Tepat Waktu
5	Sulaiman	2	3	2,5	3	2,5	Terlambat
6	Abdul Kodir	3	3	3	3,3	4	?

Tabel 2.3 merupakan tabel yang berisi lima baris dataset sebagai data latihannya. Pada tabel tersebut dapat dilihat pola IPK mahasiswa yang telah lulus dari semester I hingga semester V, sehingga diperoleh status lulusan Tepat Waktu/ Terlambat. Pada baris terakhir, terdapat satu baris data sebagai data uji yang akan

diprediksi status kelulusannya. Setelah ditentukan parameter (nilai) *K* (misal, nilai K=3) dan diperoleh hasil hitung jarak ketetanggaan terdekat dan diurutkan ranking berdasarkan jarak terpendek (nilai terkecil) seperti pada tabel 2.2. Oleh karena kategori label (dalam hal ini adalah status kelulusan) mayoritas pada *K*-terdekat yaitu Tepat Waktu, maka hasil prediksi untuk pola nilai IPK semester I hingga semester V untuk data atas nama Abdul Kodir dapat diklasifikasikan kedalam kelompok Tepat Waktu, dapat dilihat secara ringkas pada tabel 2.4.

Tabel 2. 4
Data Hasil Prediksi K-NN Status Kelulusan Abdul Kodir

No	Nama	Distance	Ranking	Kelulusan
1	d1,d6	1,66	2	Tepat Waktu
3	d3,d6	1,53	1	Terlambat
4	d4,d6	1,76	3	Tepat Waktu

5. Menentukan model prediksi

Tahap selanjutnya adalah menentukan model prediksi yang nantinya akan dijadikan acuan dalam memprediksi data-data baru sehingga dapat dilakukan suatu langkah-langkah strategis guna memanfaatkan informasi dan pengetahuan yang didapat dari pola data yang ada.

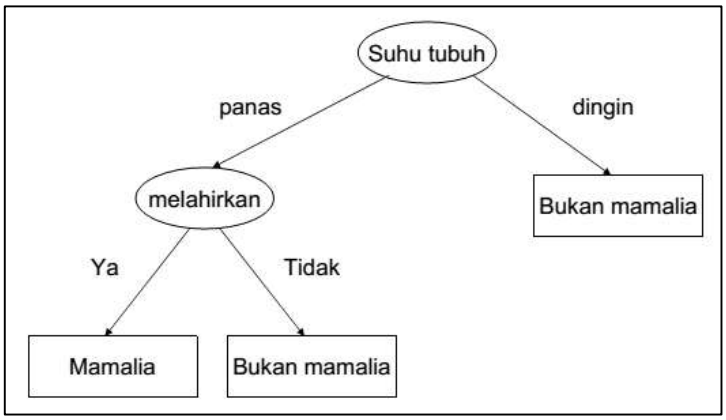
6. Menguji model prediksi

Setelah model prediksi dibuat, maka tahap terakhir yang harus dilakukan yaitu menguji model prediksi yang sudah dibuat tersebut. Fungsi dari pengujian model prediksi ini adalah untuk memastikan bahwa model prediksi yang dihasilkan dari pola data yang ada benar-benar valid dan dapat digunakan sebagai acuan dalam memprediksi data-data baru selanjutnya..

2.4.3 Model *Decision Tree*

Decision Tree (pohon keputusan) merupakan sebuah diagram alir (*flowchart*) yang serupa dengan struktur pohon, dimana setiap internal *node* menotasikan atribut yang diuji, setiap cabang merepresentasikan hasil dari atribut uji tersebut, dan *leaf node* (*terminal node*) merepresentasikan kelas-kelas tertentu atau distribusi dari kelas-kelas (Han, 2006). Model ini menggunakan pemisahan *univariate*, sehingga mudah untuk dipahami oleh

pengguna dikarenakan strukturnya yang sederhana, selain itu batasan-batasan yang diterapkan pada representasi aturan dan pohon tertentu dapat secara signifikan membatasi bentuk fungsional dari model (Han, 2012). Gambaran sederhana mengenai *Decision Tree* dapat dilihat pada gambar 2.6.



Gambar 2. 6
Contoh *Decision Tree* dalam Klasifikasi Mamalia

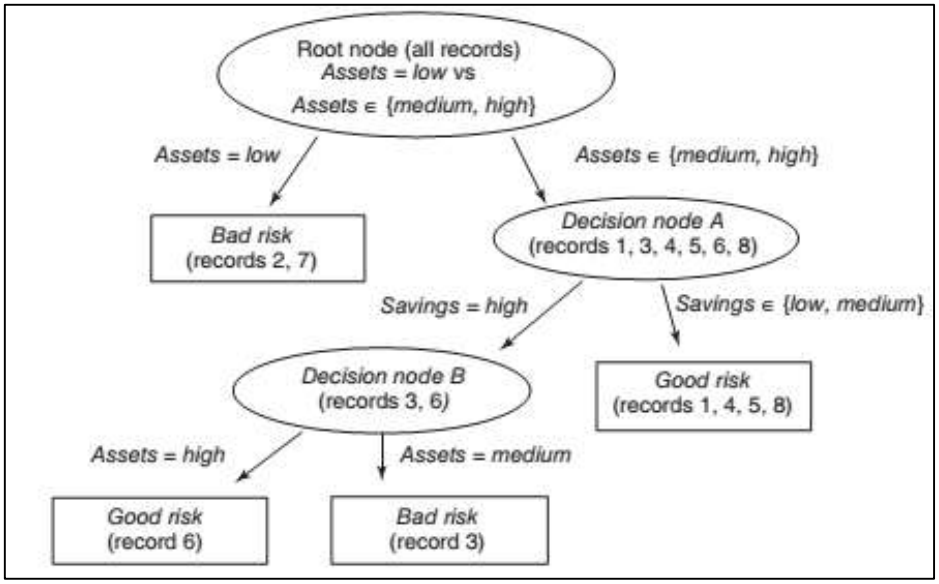
Mengacu pada (Han, 2006), dalam *decision tree* terdapat tiga jenis *node*, yaitu:

1. *Root Node*, merupakan *node* paling atas. Pada *node* ini tidak ada *input* dan bisa saja tidak mempunyai *output* atau mempunyai *output* lebih dari satu.
2. *Internal Node*, merupakan *node* percabangan. Pada *node* ini hanya terdapat satu *input* dan mempunyai *output* minimal dua.
3. *Leaf Node (Terminal Node)*, merupakan *node* akhir. Pada *node* ini hanya terdapat satu *input* dan tidak mempunyai *output*.

Terdapat berbagai macam jenis algoritma pada model *Decision Tree*, diantaranya: Algoritma ID3, Algoritma C4.5, Algoritma J48, Algoritma C5.0, dan Algoritma *Regression Tree* (biasa dikenal dengan nama CART – *Classification and Regression Tree*). Dalam penelitian ini, penulis menggunakan algoritma CART.

Algoritma CART adalah salah satu algoritma *data mining* yang diterapkan secara khusus pada model *decision tree* dalam metode klasifikasi dan masuk dalam *The Top Ten Algorithms in Data Mining* (Susanto,2010:57). Algoritma ini mempunyai dua noktah, noktah pertama disebut noktah keputusan, yakni bergambar elips. Noktah ini masih dapat dicabangkan lagi pada kondisi tertentu. Sedangkan noktah lainnya disebut noktah

terminasi, berbentuk persegi panjang. Noktah ini digunakan apabila suatu kondisi tidak dapat bercabang lagi, seperti terlihat pada gambar 2.7.



Gambar 2. 7
Contoh Model CART
(Sumber : Larose, 2015:324)

Algoritma CART tersusun berdasarkan empat (4) langkah (Susanto, 2010:58):

1. Tentukan dahulu data yang akan diproses, contoh seperti gambar 2.8

Customer	Savings	Assets	Income ($1000s)	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

Gambar 2. 8
Dataset yang siap diproses
(Sumber : Larose, 2015, 318)

2. Menilai kinerja keseluruhan calon cabang yang ada pada daftar calon cabang mutakhir, dengan mempergunakan persamaan (2.3)

$$\emptyset\,(s|t) = 2P_L P_R \sum_{j=1}^{jumlah\ kategori} |P(j|t_L) - P(j|t_R)| \quad (2.3)$$

(Sumber : Larose:2015:320)

Keterangan:

t_L = cabang calon kiri dari noktah keputusan *t*

t_R = cabang calon kanan dari noktah keputusan *t*

P_L = $\frac{\text{jumlah catatan pada calon cabang kiri } t_L}{\text{jumlah catatan pada latihan}}$

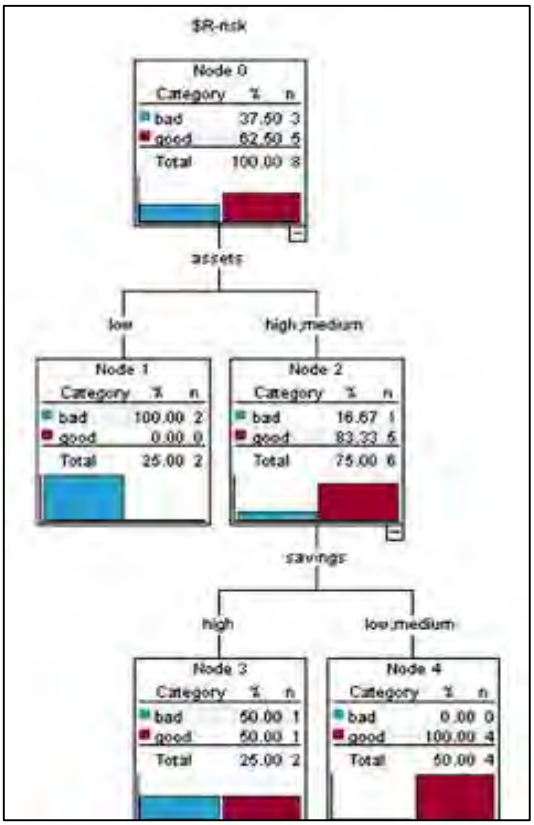
P_R = $\frac{\text{jumlah catatan pada calon cabang kanan } t_R}{\text{jumlah catatan pada latihan}}$

P(j|tL) = $\frac{\text{jumlah catatan berkategori } j \text{ pada calon cabang kiri } t_L}{\text{jumlah catatan pada noktah keputusan } t}$

P(j|tR) = $\frac{\text{jumlah catatan berkategori } j \text{ pada calon cabang kanan } t_R}{\text{jumlah catatan pada noktah keputusan } t}$

3. Menentukan calon cabang manakah yang akan dijadikan cabang dengan memilih calon cabang yang memiliki nilai kesesuaian terbesar.
4. Setelah itu gambarkan percabangan, jika tidak ada percabangan maka proses CART dihentikan. Namun jika ada percabangan maka akan dilanjutkan ke langkah nomor 2 setelah sebelumnya menghapus terlebih dahulu calon cabang yang berhasil menjadi cabang.

Pada saat menggunakan *software data mining* (IBM SPSS Modeler 18.0 misalnya), terdapat perbedaan simbol namun memiliki makna yang berbeda. Seperti contoh pada gambar 2.9, terlihat node 1 sudah tidak memiliki cabang sehingga dianggap sebagai noktah terminasi. Sedangkan node 2 masih mempunyai dua cabang keputusan lainnya, ditandai dengan tanda "-" pada kotak kecil di kanan bawah. Apabila tanda tersebut diklik maka cabang akan tersembunyi sehingga tanda "-" pada kotak kecil di kanan bawah tersebut akan berubah menjadi "+", yang artinya node tersebut masih mempunyai dua cabang keputusan lainnya.



Gambar 2. 9
Model yang dihasilkan oleh *software* IBM SPSS Modeler 18.0

2.5 Evaluasi Performa Model Klasifikasi

Setiap model yang dihasilkan melalui teknik *data mining* perlu dievaluasi. Tujuan dari evaluasi ini adalah untuk memastikan bahwa model *data mining* yang dihasilkan benar-benar dapat dipertanggungjawabkan tingkat akurasinya. Dalam penelitian ini, penulis menggunakan alat ukur berupa *confusion matrix* karena dapat memberikan informasi dalam bentuk angka, sehingga rasio keberhasilan model klasifikasi yang dibuat dapat dihitung. Pada pengukuran kinerja menggunakan *confusion matrix*, terdapat empat istilah sebagai representasi hasil proses klasifikasi, seperti yang tertuang pada tabel 2.5.

Tabel 2. 5
Confusion Matrix

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif
Positif	*True Positive* (TP)	*False Negative* (FN)
Negatif	*False Positive* (FP)	*True Negative* (TN)

(Sumber: Sokolova, 2009)

Selanjutnya, mengacu pada (Gorunescu, 2011) akurasi klasifikasi dapat dibandingkan sesuai dengan tabel 2.6.

Tabel 2. 6
Klasifikasi Akurasi

Presentase Akurasi	Klasifikasi
50 – 59	*Failure*
60 – 69	*Poor*
70 – 79	*Fair*
80 – 89	*Good*
90 – 100	*Excellent*

(Sumber: Gorunescu, 2011)

2.6 Penelitian-penelitian Terdahulu

Berikut merupakan uraian penelitian-penelitan terdahulu yang disajikan dalam tabel 2.7.

Tabel 2. 7
Penelitian Terdahulu

No	Tahun	Peneliti	Metode	Hasil
1	2009	Hanief Meinanda, dkk.	*Artificial Neural Network*	– IPK, jumlah mata kuliah yang diambil, jumlah mata kuliah mengulang, dan jumlah pengambilan mata kuliah tertentu mempengaruhi masa studi – Model regresi akan menghasilkan prediksi masa studi yang bias karena asumsi *homescedasticity error* tidak dapat dipenuhi – *Artificial Neural Network* arsitektur *multilayer Perceptron* dalam penelitian merupakan model terbaik untuk memprediksi lama masa studi
2	2013	Asep Saefulloh dan Moedjiono	*algoritma C4.5, Naïve Bayes,* dan *Neural Network*	– Dari ketiga algoritma, algoritma dengan tingkat akurasi terbaik adalah *algoritma C4.5,* dan *Neural Network* – Algoritma tersebut dapat menampilkan data mahasiswa beserta prediksi kelulusan dengan menggunakan *engine java*
3	2014	Sandy Kosasi	Jaringan Saraf Tiruan-*backpropagation*	– Jaringan saraf tiruan *backpropagation* cukup mampu melakukan prediksi nilai ujian sekolah siswa SD dengan tingkat akurasi yang cukup baik – Kombinasi parameter cukup bervariasi an terbaik yang

No	Tahun	Peneliti	Metode	Hasil
				dihasilkan cukup bervariasi jika dikaitkan dengan nilai *learning rate* dan jumlah maksimum *epoch* yang dibutuhkan untuk mencapai nilai MSE (*Mean Square Error*) terkecil – Tidak ada patokan/ pedoman pasti untuk mencapai nilai *error* terkecil dalam hal penggunaan parameter pelatihan berupa jumlah maksimum *epoch* dan besar *learning rate*
4	2015	Mambang dan Finki Dona Marleny	*Decision Tree*	– Pohon keputusan menggunakan *weight by Information Gain Ratio* dapat digunakan pada data sejenis, karena telah memiliki akurasi 88,24%, – Pohon keputusan menggunakan *weight by uncertainty* yang sedikit lebih rendah akurasinya yang hanya memiliki 80,39% akurasi. – Algoritma decision tree C4.5 dapat juga digunakan untuk klasifikasi model prediksi calon mahasiswa baru.
5	2015	I Made Budi Adnyana	*Random Forest*	– Terdapat 5 buah fitur data yang digunakan (IPK, total jumlah SKS, jumlah tidak aktif dan cuti, jumlah nilai mta kuliah A dan B , dan nilai mata kuliah D dan E) semua data yang digunakan per semester 4 – Klasifikasi yang terdiri dari 2 kelas, yaitu “lulus tepat waktu” dan “lulus lewat batas waktu studi”. – *Overall accuracy* yang didapatkan adalah 83.54% (Kappa Coefficient 0.564).
6	2016	Rudy Ansari	Jaringan Syaraf Tiruan	– Metode jaringan syaraf tiruan menghasilkan akurasi sebesar 92,49% – menghasilkan 12 *input layer* – menghasilkan 9 *hidden layer* – menghasilkan 3 *output layer*
7	2017	Dony Susandi, dkk	*Knowledge Based System – K-nearest neighbor*	– Prediksi dilakukan dengan data akademik dan latar belakang sekolah sebelumnya – Pengujian r*apidminer* dapat digunakan untuk membandingkan antara algoritma *k-nearest neighbor* dengan agoritma *backpropagation neural network* – Dari proses pengujian dengan *confusion matrix* diketahui bahwa metode algoritma *k-nearest neighbor* menghasilkan nilai akurasi tertinggi

No	Tahun	Peneliti	Metode	Hasil
				– Penerapan *rule* dari algoritma *k-nearest neighbor* yang digunakan dalam klasifikasi kelulusan mahasiswa kedalam aplikasi *knowledge based system* menggunakan CLIPS, dapat membantu dalam proses prediksi kelulusan mahasiswa
8	2017	Supardi Salmu dan Achmad Solichin	*Naive Bayes*	– Atribut yang digunakan untuk memprediksi kelulusan mahasiswa : Jenis kelamin, Jenis seleksi, Pendapatan ayah, Pendapatan ibu, IP Semester 1-4 dan SKS semester 1-4 – Pengujian data diperoleh penelitian sebesar 80,72% dari 1162 data yang digunakan *training* dan 587 data untuk *testing*

Dari beberapa penelitian sebelumnya tersebut dapat disimpulkan bahwa penggunaan teknik *data mining* mampu membantu bagian akademik untuk memberikan prediksi kelulusan mahasiswa ke depan dan rekomendasi model yang dihasilkan untuk kondisi tertentu. Perbedaan penelitian penulis dengan penelitian sebelumnya terletak pada model yang digunakan oleh penulis yaitu membandingkan model *Discriminant Analysis*, model *K-Nearest Neighbor* dan model *Decision Tree*, dimana peneliti sebelumnya belum ada yang melakukan penelitian dengan ketiga model tersebut secara bersamaan dalam satu penelitian memprediksi kelulusan mahasiswa. Perbedaan lainnya terletak pada objek penelitian yang dipilih oleh penulis dalam menyusun laporan tesis ini.

BAB III

OBJEK DAN METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian (STMIK LIKMI)

STMIK LIKMI merupakan salah satu perguruan tinggi swasta yang berlokasi di kota Bandung, tepatnya di Jl. Ir. H. Juanda No.96 Bandung. STMIK LIKMI didirikan tahun 1994 yang awalnya bernama Lembaga Ilmu Komputer Terapan Indonesia (LIKTI) pada tahun 1985 dan mempelopori sistem praktek langsung komputer. Pada tahun 2004 STMIK LIKMI mulai berkembang menjadi kampus digital, dimana menggunakan smart card sebagai kartu pengenal mahasiswa dan pada tahun 2006 STMIK LIKMI membuka pendidikan Magister Sistem Informasi (Magister Komputer).

STMIK LIKMI membuka pendidikan dengan program sekretaris, diploma III untuk jurusan sistem informasi, pendidikan sarjana dengan jurusan sistem informasi (yang memilki 3 bidang keahlian) dan jurusan teknik informasi (dengan 3 bidang keahlian) dan pendidikan pasca sarjana dengan jurusan sistem informasi bisnis dan rekayasa sistem informasi. Program magister STMIK LIKMI sendiri menjadi salah satu dari lima penyelenggara S2 bidang Sistem Informasi terbaik di Indonesia.

Sama seperti perguruan tinggi lainnya, proses bisnis STMIK LIKMI memulai bisnisnya dengan datang ke sekolah-sekolah menengah atas yang ada di kawasan propinsi Jawa Barat untuk menawarkan beasiswa bagi pelajar berprestasi dan mempromosikan STMIK LIKMI sendiri. Terdapat dua kegiatan pada proses bisnis perguruan tinggi khususnya STMIK LIKMI:

1. Kegiatan Utama, meliputi Pelaksanaan Proses Pembelajaran, Penelitian dan Pengabdian, Penerimaan Mahasiswa Baru, Kerjasama dengan Instansi serta Pemasaran.
2. Kegiatan Pendukung, merupakan kegiatan yang mendukung kegiatan utama meliputi Layanan Akademik, Layanan Keuangan, Layanan Kepegawaian, Layanan Sistem Informasi, Layanan Sarana dan Prasarana, dan Layanan Kemahasiswaan.

3.2 Metodologi Penelitian

Pada bagian ini akan diuraikan mengenai metode penelitian yang penulis gunakan untuk menyelesaikan laporan tesis ini. Metodologi penelitian ini diharapkan mampu membantu penulis dalam melakukan penelitian dari awal sampai dengan penelitian ini selesai dilakukan, sehingga laporan tesis dapat tersusun secara sistematis. Secara ringkas metodologi penelitian yang digunakan penulis diuraikan sebagai berikut (dapat dilihat pada gambar 3.1).

3.2.1 Menentukan Topik Penelitian

Tahap awal yang harus dilakukan penulis sebelum melakukan penelitian yaitu menentukan topik. Topik yang telah ditentukan (*data mining*) selanjutnya menjadi persiapan awal penelitian. Data mining menarik untuk dibahas karena dapat menghasilkan suatu *knowledge* dari kumpulan data yang sebelumnya tidak pernah terpikirkan untuk dapat diolah sedemikan rupa.

3.2.2 Menentukan Objek Penelitian

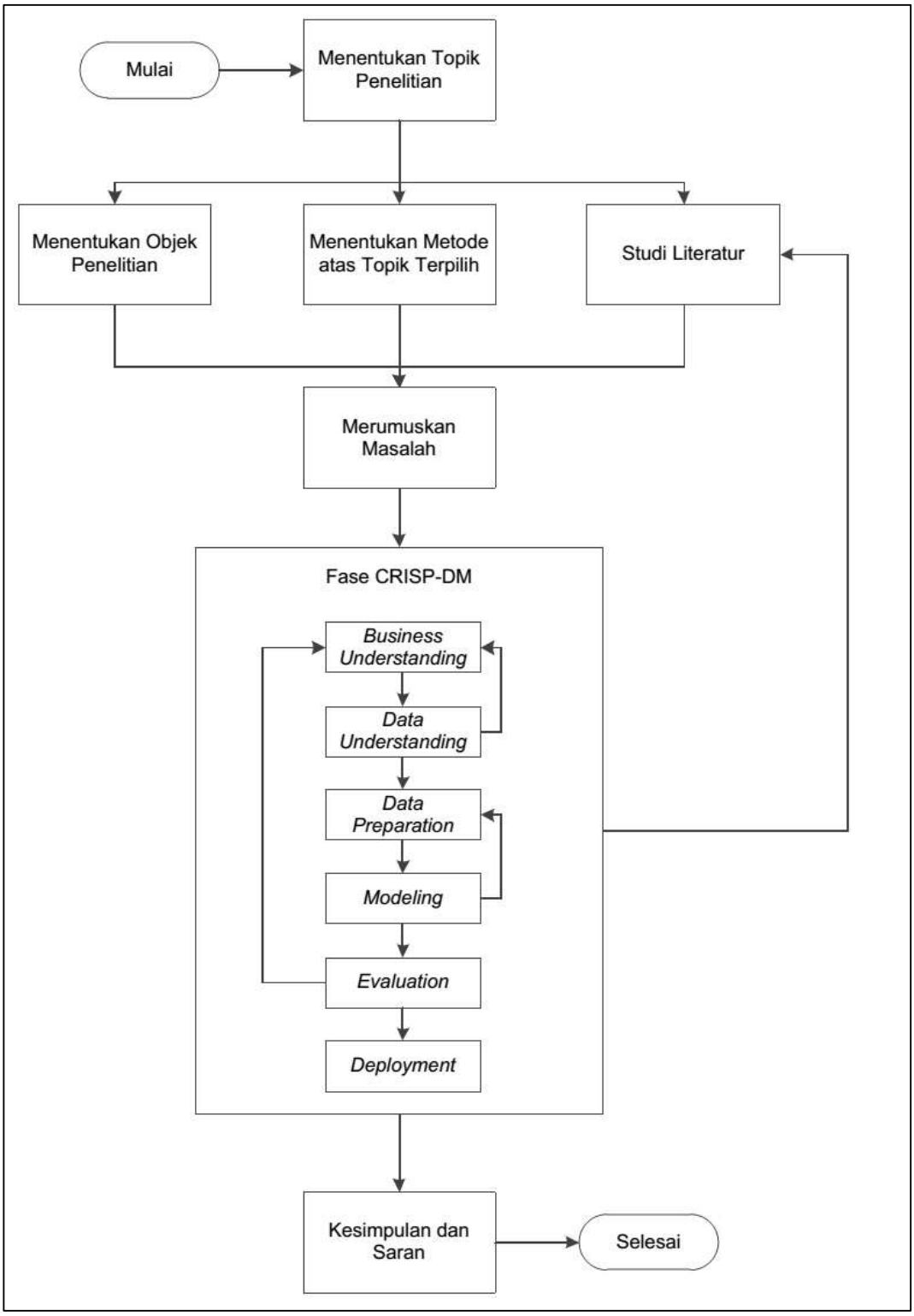
Tahap selanjutnya setelah topik ditentukan yaitu menentukan objek penelitian. Objek penelitian yang dipilih penulis dalam melakukan penelitian tesis ini adalah STMIK LIKMI, yakni dengan menggunakan data nilai mahasiswa alumni STMIK LIKMI.

3.2.3 Menentukan Metode atas Topik Terpilih

Setelah topik penelitian dipilih, selanjutnya penulis menentukan metode yang akan digunakan penulis untuk membantu penulis dalam melakukan penelitian, yakni model *Discriminant Analysis*, model *K-Nearest Neighbor* dan model *Decision Tree.* Ketiga metode ini penulis pilih dikarenakan metode-metode ini lebih mudah untuk dimengerti dan dapat menghasilkan prosedur yang lebih sederhana sehingga proses kerja dapat lebih efisien.

3.2.4 Studi Literatur

Pada tahap ini penulis memulai dengan mencari dan membaca teori-teori yang berhubungan dengan penelitian yang akan dilakukan penulis. Pada tahap ini diharapkan penulis memiliki pedoman teoritis dalam penelitian ini. Studi literatur ini bersumber dari buku-buku *data mining* dan jurnal-jurnal penelitian sebelumnya.



Gambar 3. 1
Metodologi Penelitian

3.2.5 Merumuskan Masalah

Setelah menentukan topik penelitian, objek penelitian dan melakukan studi literatur, maka penulis merangkum setiap pertanyaan besar yang menjadi kajian penelitian ke dalam rumusan masalah dan membuat peta konsep *data mining* sebagai solusi untuk penyelesaian masalah yang menjadi pertanyaan besar penulis tersebut.

3.2.6 Fase-fase CRISP-DM

Merujuk pada subbab 2.2 mengenai siklus *data mining*, penulis menyisipkan fase-fase CRISP-DM (*Cross Industry Standard Process for Data Mining*) ke dalam bagian metodologi penelitian sebagai kerangka kerja dalam pemrosesan *data mining*. Siklus CRISP-DM meliputi enam (6) fase yang masing-masing memiliki fungsi tersendiri untuk memberikan kontribusi terhadap proses pengerjaan *data mining*, dimulai dari fase *business understanding,* fase *data understanding,* fase *data preparation,* fase *modeling,* fase *evaluation, dan* fase *deployment.*

3.2.7 Kesimpulan dan Saran

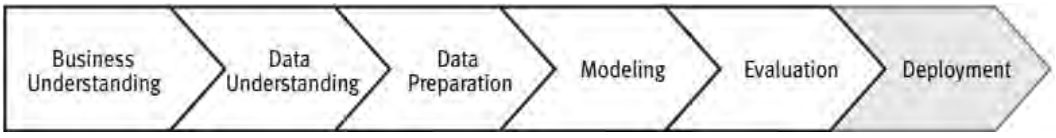
Tahap akhir dalam metodologi penelitian ini mencakup kesimpulan dan saran atas penelitian yang sudah dilakukan oleh penulis. Kesimpulan dari penelitian ini merupakan jawaban dari rumusan masalah yang menjadi pertanyaan penulis pada saat merumuskan masalah. Saran dari penelitian ini merupakan usulan membangun untuk pengembangan penelitian ini atas permasalahan yang mungkin belum terjawab secara lengkap dan terperinci.

BAB IV

ANALISIS DAN PEMBAHASAN HASIL

4.1 Analisis Proses *Data Mining*

Mengacu pada urutan metodologi penelitian yang telah diuraikan di Bab III sub-bab 3.2, pada bagian ini akan dijelaskan secara rinci dari poin 3.2.6 mengenai fase-fase CRISP-DM sebagai urut-urtan proses pengerjaan *data mining* seperti pada gambar 4.1.



Gambar 4. 1
Urutan Proses Pengerjaan *Data Mining*
(Sumber: Chapman, 2000:13)

4.1.1 Fase *Business Understanding*

Pada fase ini menggambarkan mengenai permasalahan yang ada di STMIK LIKMI sehingga perlunya dibuat permodelan *data mining*. Selanjutnya diuraikan pula mengenai tujuan bisnis dari organisasi, penilaian kondisi organisasi, dan tujuan dari *data mining* itu sendiri.

4.1.1.1 Tujuan bisnis

Saat ini (tahun 2017) STMIK LIKMI menginjak usia ke-23 tahun. Di tengah persaingan antar institusi perguruan tinggi, terdapat beberapa universitas besar di kota Bandung yang terlebih dahulu berdiri dan dikenal oleh masyarakat luas, ditambah pula dengan bermunculnya berbagai insitusi perguruan tinggi baru yang ada di kota Bandung memberikan alternatif pilihan bagi para calon mahasiswa/i untuk melanjutkan studi ke jenjang yang lebih tinggi. Salah satu dampak langsung terhadap STMIK LIKMI adalah penurunan jumlah pendaftar calon mahasiswa/i. Tujuan bisnis yang ingin dicapai STMIK LIKMI yakni dapat mampu mempertahankan eksistensinya di tengah persaingan antar institusi perguruan tinggi dengan menjadi perguruan tinggi terkemuka dalam bidang bisnis dan informatika yang mampu menghasilkan lulusan yang berkarakter unggul dan dapat memenuhi kebutuhan masyarakat. Agar hasil dari proses kajian *data mining* ini berhasil dan tepat sasaran maka diperlukan penentuan kriteria keberhasilan, yakni:

1. Dapat memprediksi subjek (mata kuliah) mana yang menentukan hasil akhir kelulusan mahasiswa.
2. Dapat memberikan *knowledge* sebagai informasi mahasiswa mana yang paling besar kemungkinannya lulus dengan predikat Memuaskan (rata-rata), sehingga dosen wali dengan cepat dapat menyusun strategi yang tepat dalam menghasilkan mahasiswa lulus dengan predikat *Cumlaude* (Dengan Pujian) atau paling tidak lulus dengan predikat Sangat Memuaskan.
3. Dapat memanfaatkan pengelolaan data (nilai) historis mahasiswa secara efektif.

4.1.1.2 Penilaian kondisi (*Assess Situation*)

Dalam menjalankan proses *data mining* perlu adanya sumber daya yang mendukung selama berlangsungnya proses *data mining*. Sumber daya yang nantinya dibutuhkan untuk menjalankan proses *data mining* ini terbagi ke dalam beberapa komponen, di antaranya:

1. Personel
   a. *Database administrator*, diperlukan untuk mengatur kebutuhan dan ketersediaan data yang selanjutnya diperlukan untuk proses *data mining*.
   b. *Analyst system*, diperlukan untuk merancang dan mengntegrasikan *data mining* guna mendukung sistem yang beroperasi di instansi.
   c. *Programmer*, diperlukan untuk membuat aplikasi prediksi lulusan mahasiswa dari proses *data mining* dengan mengambil data dari *database* operasional.
   d. Teknisi jaringan komputer, diperlukan untuk pengembangan dan penambahan jalur jaringan komputer yang diperlukan berkaitan dengan proses *data mining*.
2. Data

   Pengaksesan data terhadap *database* untuk mendapatkan data historis yang digunakan sebagai dasar pembuatan model prediksi. Data historis yang digunakan merupakan data yang valid.
3. *Platform,* terdiri dari:
   a. *Operating System* : Windows Server

b. *Database* : SQL Server 2008

4. *Software,* terdiri dari:

   a. IBM SPSS Statistics 24 dan IBM SPSS Modeler 18.0 untuk melakukan proses *data mining*

   b. Ms.Excel untuk pembersihan *dataset*

   c. Java dan Visual Foxpro 9 untuk pengaksesan data ke *database*

4.1.1.3 Tujuan *data mining*

Penggunaan teknik *data mining* ini tidak lain adalah untuk mendukung tujuan dari bisnis organisasi itu sendiri. Adapun tujuan dari *data mining* ini adalah:

1. Memprediksi seberapa besar mahasiswa yang berpotensi lulus dengan predikat *Cumlaude* (Dengan Pujian), sehingga dapat dibuat matrik klasifikasi mahasiswa dengan memetakan mahasiswa berdasarkan grup status kelulusannya, yakni *Cumlaude* (Dengan Pujian), Sangat Memuaskan, dan Memuaskan (Rata-rata)
2. Memberikan *knowledge* kepada dosen wali mengenai fungsi dan *rule* yang dibentuk dari hasil proses *data mining* setelah melakukan serangkaian ketentuan algoritma klasifikasi yang dipilih.
3. Memberikan informasi mengenai subjek (mata kuliah) mana yang paling berpengaruh terhadap prestasi belajar mahasiswa se-dini mungkin..

4.1.2 Fase *Data Understanding*

Fase ini menjelaskan mengenai aktvitas-aktivitas yang dimulai dari pengambilan data hingga data tersebut siap digunakan untuk membuat pemodelan *data mining*.

4.1.2.1 Pengumpulan data awal

Agar proses pemahaman data ini dapat berjalan dengan baik, maka perlu diuraikan mengenai data-data awal apa saja yang harus dipersiapkan untuk melakukan proses *data mining*. Adapun *database* yang digunakan yaitu *database* Mahasiswa dan *database* Nilai_Mahasiswa

4.1.2.2 Gambaran data yang dibutuhkan

Setelah *dataset* diperoleh, selanjutnya akan digambarkan secara rinci mengenai informasi tabel data yang akan digunakan dalam pemrosesan *data mining*. Dikarenakan data-data kemahasiswaan (profil, kuliah, keuangan, dll) bersifat sangat rahasia, maka penulis hanya dapat membuat daftar *field* apa saja yang dibutuhkan untuk pemrosesan *data mining* ini, yang selanjutnya dibantu oleh bagian IT STMIK LIKMI untuk dilakukan penggabungan relasi antar tabel master hingga penarikan (pengambilan) data tersebut. Daftar *field* yang dibutuhkan penulis dapat dilihat pada tabel 4.1.

Tabel 4. 1
Daftar *field* yang dibutuhkan

No	Nama *Field*	Keterangan
1	NPM	Nomor Pokok Mahasiswa STMIK LIKMI
2	NAMA	Nama mahasiswa STMIK LIKMI (disamarkan)
3	MK	Mata Kuliah yang diambil selama mahasiswa studi di STMIK LIKMI (Semester I s/d Semester VIII)
4	NILAI AKHIR	Nilai mata kuliah yang diperoleh
5	IPK	Indeks Prestasi Kumulatif akhir pada saat lulus
6	STATUS LULUS	Menyatakan status kelulusan seorang mahasiswa (Dengan Pujian, Sangat Memuaskan, Memuaskan)

4.1.2.3 *Explore data*

Pada bagian ini akan dijelaskan mengenai langkah-langkah untuk mendapatkan data yang kemudian akan diproses *data mining* adalah sebagai berikut:

1. Pengambilan data dilakukan pada 19 Juli 2017 dibantu oleh bagian IT kampus STMIK LIKMI. Data yang dimaksud yakni data nilai mahasiswa angkatan tahun 2010 sampai dengan 2014. Data diambil dengan merelasikan beberapa tabel sehingga menjadi sebuah tabel utuh yang terdiri atas NPM, NAMA, THN_SEMESTER, KODE, MK, NILAI AKHIR, IPK, SKS_LULUS. Untuk kepentingan kerahasiaan maka nama mahasiswa disamarkan.
2. Selanjutnya data nilai mahasiswa yang telah diperoleh dilakukan penyesuaian nama mata kuliah (dalam hal ini sebagai variabel prediktor). Karena terdapat beberapa perubahan nama mata kuliah, misal: Teknik Riset Operasional menjadi Manajemen Operasional.

3. Kemudian dilakukan pembersihan data, yakni data-data yang tidak ada kaitannya dengan mata kuliah (variabel) yang akan digunakan untuk *data mining* dihapus, agar tidak menimbulkan *error* pada saat *generate* di program aplikasi.
4. Proses berikutnya adalah melakukan *cross-tab* sehingga data nilai matakuliah setiap mahasiswa yang awalnya berurutan ke bawah diubah menjadi berurutan ke samping kanan. Proses *cross-tab* penulis lakukan agar mempermudah dalam memberikan gambaran peta dari tiap-tiap mahasiswa yang telah menempuh mata kuliah wajib program studi sistem informasi. Sehingga diperoleh field yang sudah siap yakni NPM, MATA_KULIAH, IPK, STATUS_LULUS.

Adapun langkah-langkah detail mengenai tahapan ini penulis tuangkan ke dalam Lampiran 1 (Pengambilan dan pemrosesan data awal).

4.1.2.4 Verifikasi kualitas data

Agar data yang diolah dalam proses *data mining* ini benar-benar valid dan terbebas dari kesalahan, maka perlu dilakukan verifikasi data, yakni:

1. Memastikan bahwa setiap baris tidak ada duplikasi data selama berlangsungnya proses pengambilan data.
2. Pengambilan data berasal dari sumber data yang dapat dipertanggungjawabkan, artinya sudah dilakukan validasi oleh bagian IT STMIK LIKMI.

4.1.3 Fase *Data Preparation*

Fase ini merupakan aktivitas yang dilakukan dalam menyiapkan data menjadi *dataset*. *Dataset* inilah yang akan digunakan sebagai acuan untuk pembentukan model *data mining*.

4.1.3.1 Pemilihan data dan penentuan variabel

Dalam penelitian ini penulis telah memilih data yang akan menjadi parameter status kelulusan mahasiswa seperti pada tabel 4.2

Tabel 4. 2
Variabel Prediktor dan Variabel Target

Variabel Prediktor		
No	Nama *Field*	Keterangan
1	MK	Mata kuliah yang digunakan yakni mata kuliah wajib program studi Sistem Informasi (Semester I s/d Semester VIII) mulai dari mahasiswa angkatan 2010 hingga 2014 yang telah lulus. Berdasarkan pemilihan ini diperoleh sebanyak tiga puluh satu (31) Mata Kuliah, yakni: Matematika Bisnis, Pengembangan Proyek Sistem Informasi. Sistem Basis Data, Jaringan Komputer dll.
Variabel Target		
No	Nama *Field*	Keterangan
1	STATUS LULUS	Menyatakan status kelulusan seorang mahasiswa (Dengan Pujian, Sangat Memuaskan, Memuaskan).

4.1.3.2 *Clean data*

Sebelum membuat pemodelan dataset terlebih dahulu dataset dibersihkan (*cleaning data*), tujuannya yaitu untuk menghilangkan data-data yang dianggap *error*. *Cleaning data* ini dikategorikan menjadi tiga, yakni:

1. Standarisasi penamaan

   Untuk memastikan keseragaman data yang ada pada tabel daftar nilai mahasiswa. Contoh: penyeragaman penulisan mata kuliah Manajemen Operasional yang sebelumnya ditulis Teknik Riset & Operasi.

2. Perubahan bentuk

   Untuk memastikan kesesuaian data yang ada pada tabel daftar nilai mahasiswa. Contoh: untuk mahasiswa yang beragama muslim akan mengambil mata kuliah Agama, sedangkan mahasiswa yang beragama non-muslim mengambil mata kuliah Fenomenologi Agama. Maka, penamaan dalam tabel disesuaikan menjadi Agama saja.

3. Penghapusan nilai

   Langkah ini perlu dilakukan apabila data tersebut sudah tidak dapat ditelusuri keberadaannya. Contoh: ada daftar mahasiswa yang sudah terlalu banyak mata kuliah bernilai kosong/ sudah tidak melanjutkan studi, maka data ini tidak dapat diikutsertakan dalam pemrosesan *data mining*.

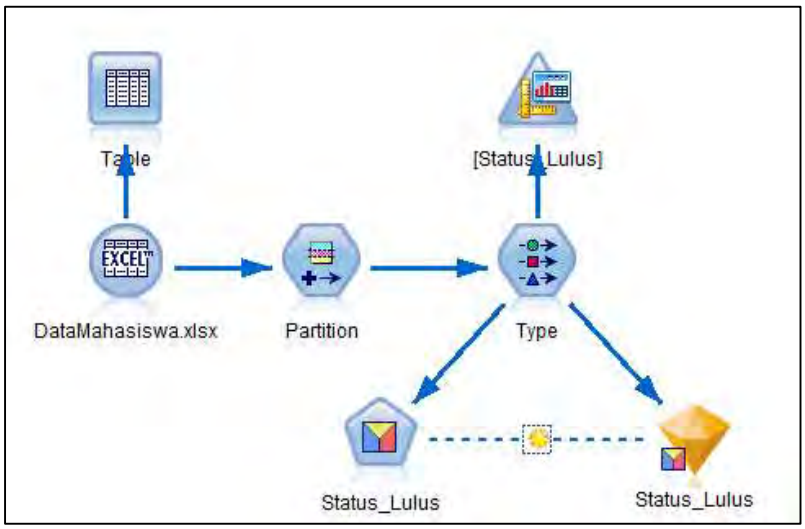
4.1.4 Fase *Modeling*

Fase selanjutnya adalah pemilihan model klasifikasi beserta *sodtware* yang digunakan untuk pembuatan model *data mining* ini. Penulis memilih model dalam metode klasifikasi yakni model *Discriminant Analysis*, model *Decision Tree*, dan model *K-Nearest Neighbor*. Sedangkan untuk *software* yang digunakan untuk membantu pembuatan model ini adalah IBM SPSS Statistics 24 dan IBM SPSS Modeler 18.0.

4.1.4.1 Model *Discriminant Analysis*

Discriminant Analysis atau Analisis Diskriminan bekerja sebagai berikut:

1. Sebelum melakukan pemodelan *Discriminant Analysis* terlebih dahulu data nilai mahasiswa yang telah disusun, dilakukan transformasi terlebih dahulu yakni: Nilai A = 4, Nilai B = 3, Nilai C = 2, Nilai D = 1. Transformasi ini perlu dilakukan karena untuk membuat model prediksi *Discriminant Analysis* diperlukan tipe data numerik. Adapun skema dari proses pemodelan *Discriminant Analysis* dapat dilihat pada gambar 4.2.



Gambar 4. 2
Proses akhir dalam pemrosesan model *Discriminant Analysis*

2. Berdasarkan atribut mahasiswa yang ada, teknik ini mengidentifikasi mahasiswa mana yang melakukan diskriminasi di antara kelompok. Atribut yang sesuai didapat dari mata kuliah yang terkait dengan program studi Sistem Informasi. Dari dua puluh lima (25) variabel prediktor yang sudah ditentukan di awal (mata kuliah semester 1 hingga semester 4) tidak semua variabel tersebut dapat digunakan

untuk pemodelan prediksi ini, sebanyak delapan (8) variabel prediktor telah terpilih berdasarkan hasil generate variabel *Mahalanobis Distance* di program aplikasi *IBM SPSS Statistics 24* menghasilkan *Standardized Canonical Discriminant Function Coefficients,* dapat dilihat pada gambar 4.3.

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Bhs_Inggris_I	.253	.476
Peng_Tek_Info	.372	-.535
Mat_Bisnis	.282	.378
Stat_I	.390	-.506
Sist_Info_Manaj_II	.214	.479
Apl_Peng_Proyek	.271	-.370
Peng_Sist_Info_I	.227	.041
Kec_Antar_Personal	.404	.347

Gambar 4. 3
Variabel prediktor terpilih untuk model analisis diskriminan

3. Menggunakan atribut yang diidentifikasi guna mengembangkan fungsi diskriminan untuk menghitung dua atau lebih nilai diskriminan dan untuk menghasilkan peta teritorial. Berdasarkan variabel prediktor terpilih, maka akan didapat fungsi diskriminan seperti gambar 4.4.

Canonical Discriminant Function Coefficients

	Function	
	1	2
Bhs_Inggris_I	.318	.599
Peng_Tek_Info	.533	-.766
Mat_Bisnis	.389	.520
Stat_I	.535	-.695
Sist_Info_Manaj_II	.263	.588
Apl_Peng_Proyek	.371	-.506
Peng_Sist_Info_I	.300	.055
Kec_Antar_Personal	.507	.435
(Constant)	-9.903	-.225

Unstandardized coefficients

Gambar 4. 4
Kriteria fungsi diskriminan yang terbentuk

Dari tiga grup dalam kasus prediksi kelulusan mahasiswa pada penelitian ini (Dengan Pujian, Sangat Memuaskan, dan Memuaskan) terbentuk dua fungsi diskriminan dengan kriteria:

a. Fungsi Diskriminan 1, untuk memilah mana yang masuk ke grup Memuaskan atau ke grup Sangat Memuaskan, dengan fungsi:

   Z Score_1=(-9,903)+(0,318 Bhs_Inggris_I)+(0,533 Peng_Tek_Info) +(0,389 Mat_Bisnis)+(0,535 Stat_I) + (0,263 Sist_Info_Manaj_II) + (0,371 Apl_Peng_Proyek) + (0,300 Peng_Sist_Info_I) + (0,507 Kec_Antar_Personal)

b. Fungsi Diskriminan 2, untuk memilah mana yang masuk ke grup Sangat Memuaskan atau ke grup Dengan Pujian, dengan fungsi:

   Z Score_2=(-0,225)+(0,599 Bhs_Inggris_I)+(-0,766 Peng_Tek_Info) +(0,520 Mat_Bisnis)+(-0,695 Stat_I) + (0,588 Sist_Info_Manaj_II) + (-0,506 Apl_Peng_Proyek) + (0,055 Peng_Sist_Info_I) + (0,435 Kec_Antar_Personal)

Maka, kedua fungsi diskriminan di atas harus digunakan secara bersama-sama, tidak bisa hanya menggunakan salah satu fungsi diskriminan saja untuk kemudian memodelkan prediksi analsis diskriminan.

4. Menentukan nilai *centroid* (rata-rata tiap grup) yang dihitung pada peta teritorial agar mengklasifikasikan siswa menjadi satu dari beberapa kelompok yang ada. *Centroid* adalah nilai rata-rata fungsi Z Score dari setiap objek yang ada pada grup (dalam hal ini objek adalah mahasiswa). Dikarenakan terdapat tiga grup maka akan ada tiga nilai *centroid* yang terbentuk, sehingga terdapat satu *centroid* yang menyatakan nilai rata-rata dari semua data mahasiswa yang ada di grup lulus Dengan Pujian, satu *centroid* yang menyatakan nilai rata-rata dari semua data mahasiswa yang ada di grup lulus Sangat Memuaskan, dan satu *centroid* yang menyatakan nilai rata-rata dari semua data mahasiswa yang ada di grup lulus Memuaskan (dilihat pada gambar 4.5). Kegunaan *centroid* yakni untuk mengetahui bagaimana penyebaran data dari tiap grup dan bagaimana kedekatan antar *centroid* dari masing-masing grup.

Functions at Group Centroids

Status_Lulus	Function 1	Function 2
Memuaskan	-2.752	.392
Sangat Memuaskan	.142	-.440
Dengan Pujian	2.073	.500

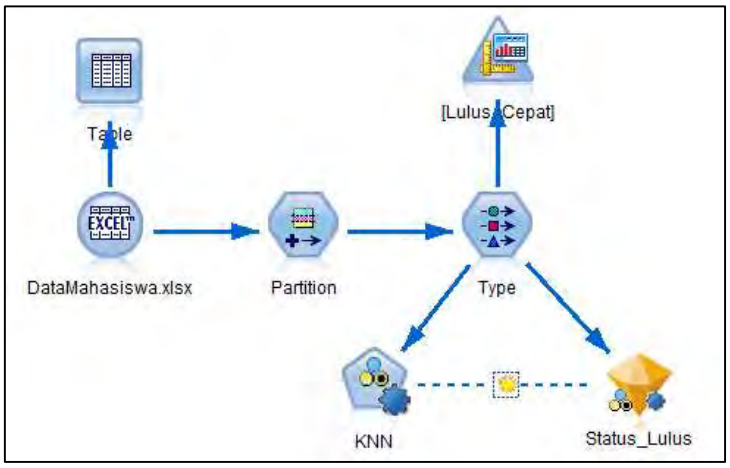
Unstandardized canonical discriminant functions evaluated at group means

Gambar 4. 5
Functions at Group Centroids

Dari data tersebut, titik *centroid* untuk grup lulus Memuaskan adalah -2,752 pada fungsi 1, dan 0,392 pada fungsi 2. Demikian seterusnya untuk titk *centroid* dua grup lainnya. Untuk lebih jelas, angka-angka di atas dibuat grafik yang disebut Peta Teritorial (*Territorial Map*) yang dapat dilihat pada bagian lampiran beserta hasil prediksi menggunakan model ini.

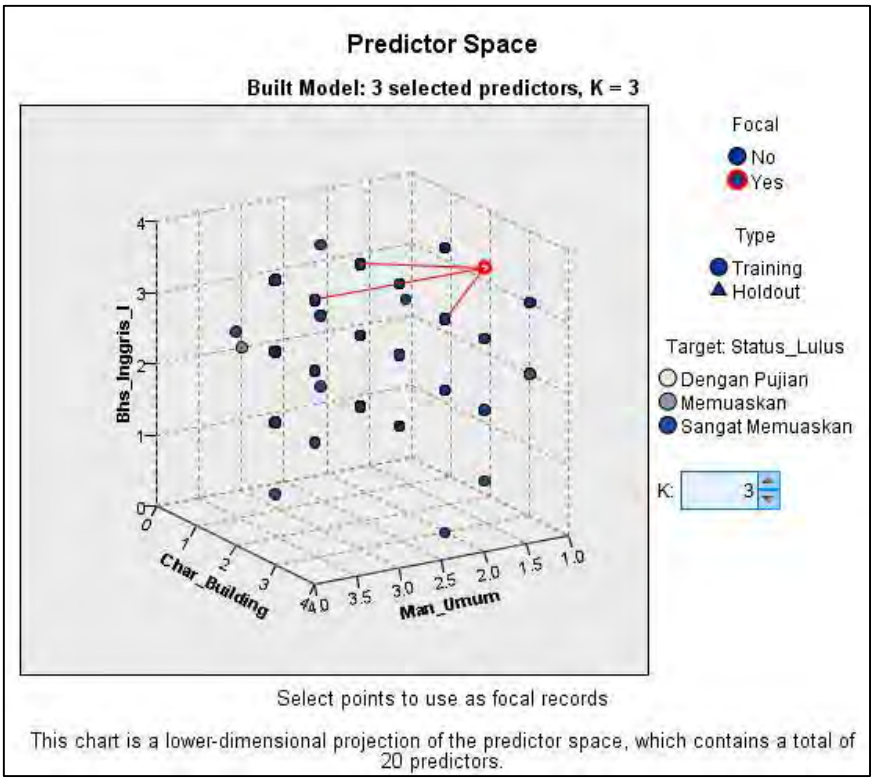
4.1.4.2 Model *K-Nearest Neighbor*

Berdasarkan daftar nilai mahasiswa yang telah disusun, dilakukan transformasi nilai terlebih dahulu yakni: Nilai A = 4, Nilai B = 3, Nilai C = 2, Nilai D = 1, dan Nilai E = 0, karena untuk membuat model prediksi *K-Nearest Neighbor* diperlukan tipe data numerik.



Gambar 4. 6
Proses akhir dalam pemrosesan model *K-Nearest Neighbor*

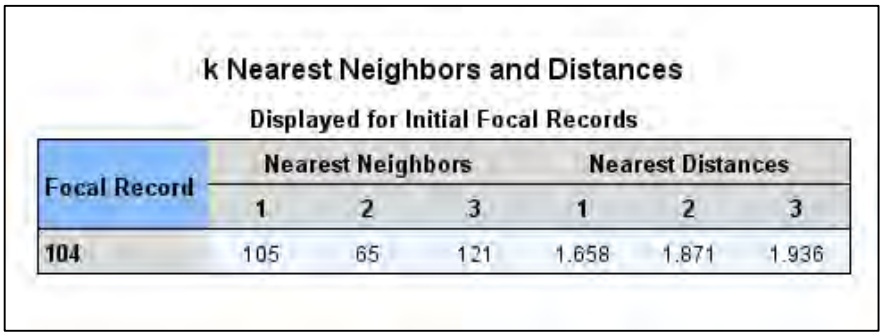
Gambar 4.6 merupakan skema dari proses pemodelan *K-Nearest Neighbor*, dimana nilai *K*-terdekat yang akan dibuat acuan sebagai ranking menggunakan teknik *V-Fold Cross-validation* mengacu pada subbab 2.4.2 poin 1, sehingga diperoleh nilai *K*=3 yang berarti akan ada peringkat 3 teratas dalam mengidentifikasi kesamaan atau ketetanggaan terdekat, seperti pada gambar 4.7.



Gambar 4. 7
Hasil *V-Fold Cross-validation*

Pada gambar 4.7 dapat dilihat bahwa terdapat tiga (3) variabel prediktor yang ditampilkan dalam model ini, yaitu Man_Umum, Char_Building, dan Bhs_Inggris_I merupakan variabel prediktor yang paling menonjol dalam model ini. Bidang dimensi tersebut (gambar 4.7) dinamakan *scatterplot*. *Scatterplot* tersebut interaktif dan dapat diputar (dipilih satu per satu) oleh kursor *mouse*, maka variabel sumbu dapat berubah.

Selanjutnya dilakukan identifikasi mengenai jarak terdekat dari nilai *K* yang sudah ditentukan di awal. Seperti contoh pada gambar 4.8, terdapat objek pada *focal record* ke-104 memiliki kemiripan atau kedekatan pada tiga tetangga dengan nomor *record* 105, 65, dan 121, pada jarak berturut-turut 1.658, 1.871, dan 1.936. Jarak ketetanggaan terdekat ini diperoleh dari rumus *euclidean distance* (lihat persamaan 2.2 pada subbab 2.4.2) yang sudah dihitung secara otomatis oleh program aplikasi IBM SPSS Modeler 18.0.



k Nearest Neighbors and Distances

Displayed for Initial Focal Records

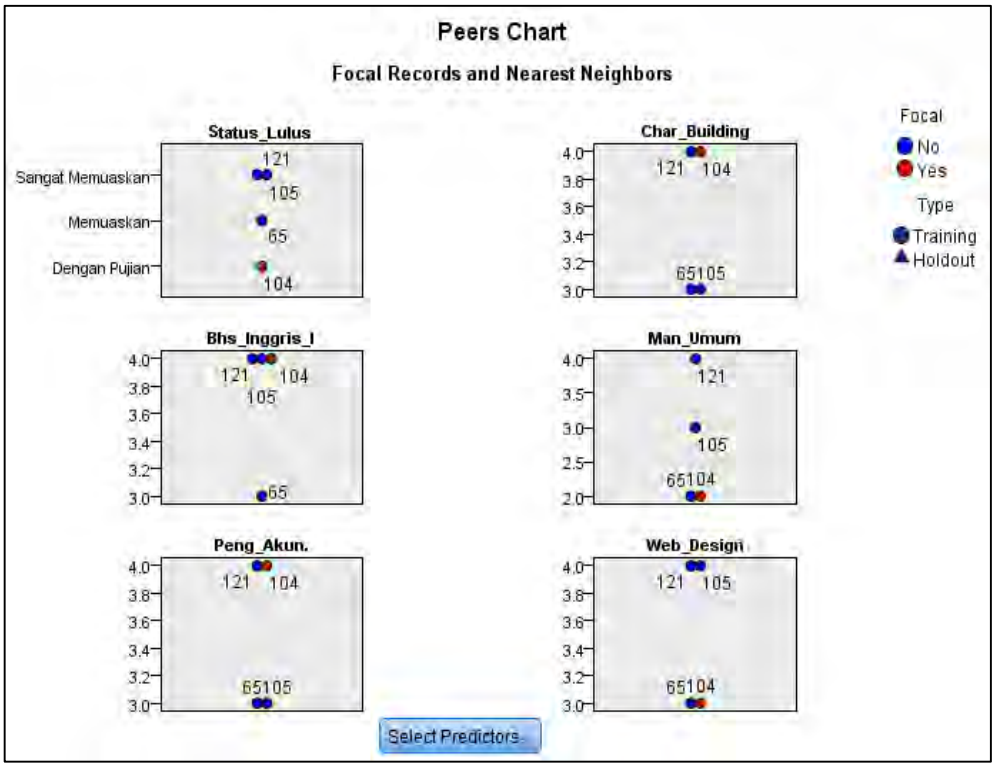
Focal Record	Nearest Neighbors			Nearest Distances		
	1	2	3	1	2	3
104	105	65	121	1.658	1.871	1.936

Gambar 4. 8
Hasil *euclidean distance*

Diagram *peers* (gambar 4.9) menampilkan *focal cases* dan tetangga terdekatnya masing-masing prediktor dan terhadap target. Diagram ini dapat tersedia apabila sebuah *focal case* dipilih pada *predictor space* seperti gambar 4.7. Apabila tidak ada satupun *focal case* yang dipilih maka diagram *peers* ini tidak akan muncul pada tampilan layar.

Diagram *peers* membaca hasil nilai *K* yang telah didapat, dalam kasus ini nilai *K* yang diperoleh adalah 3, sehingga *focal record* yang diprediksi (*focal record* 104 pada gambar 4.8 misalnya) akan selalu ditampilkan berdampingan dengan 3 buah *focal cases* lainnya sebagai tetangga terdekatnya pada masing-masing variabel prediktor. Demikian pula halnya apabila nilai *K* yang diperoleh adalah 5 ataupun 7 dan seterusnya, maka *focal record* yang diprediksi akan selalu ditampilkan berdampingan dengan banyaknya *focal cases* sesuai dengan nilai *K* yang diperoleh.

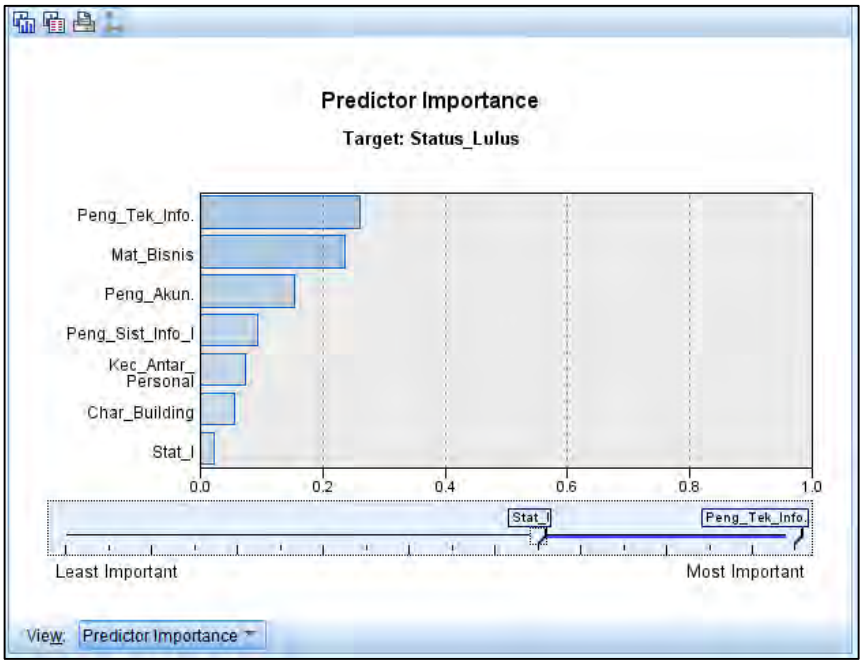
Diagram *peers* ini sebenarnya ingin menggambarkan tingkat ranking objek prediksi yang dibandingkan pada masing-masing variabel prediktor, seperti yang telah dipaparkan pada subbab 2.4.2 poin 3 yaitu mengurutkan data uji berdasarkan jarak minimum terhadap *K*. Sehingga pada akhirnya dapat ditentukan objek yang diprediksi tersebut akan termasuk pada kelas (grup) yang dimaksud (Dengan Pujian, Sangat Memuaskan ataukah Memuaskan).



Gambar 4. 9
Diagram *peers*

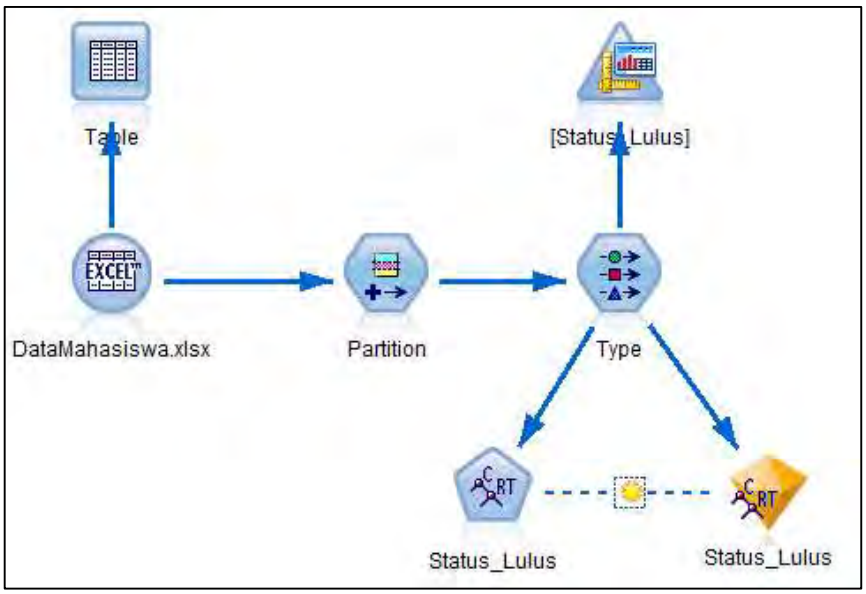
4.1.4.3 Model *Decision Tree*

Berdasarkan variabel prediktor yang telah ditentukan sebelumnya, yakni dua puluh (20) mata kuliah dan variabel target yaitu status kelulusan (Dengan Pujian, Sangat Memuaskan, dan Memuaskan) selanjutnya dibuat model *Decision Tree* (algoritma CART) yang nantinya akan menghasilkan sebuah *rule* atau aturan *If . . .Then . . .* (Jika ...Maka...). Model ini mampu membaca tipe data numerik maupun karakter, sehingga tidak perlu dilakukan transformasi data. Tidak semua variabel prediktor yang ada dapat digunakan untuk pemodelan prediksi ini. Setelah melalui analisis *feature selection* maka didapat beberapa variabel prediktor yang dinilai dapat memberikan hasil yang baik pada pemodelan ini, seperti pada gambar 4.10.



Gambar 4. 10
Variabel prediktor terpilih untuk memodelkan algoritma CART

Setelah proses pemilihan variabel prediktor, selanjutnya dilakukan proses pemilihan model algoritma CART. Adapun skema dari proses pemodelan algoritma CART dapat dilihat pada gambar 4.11.



Gambar 4. 11
Proses akhir dalam pemrosesan algoritma CART

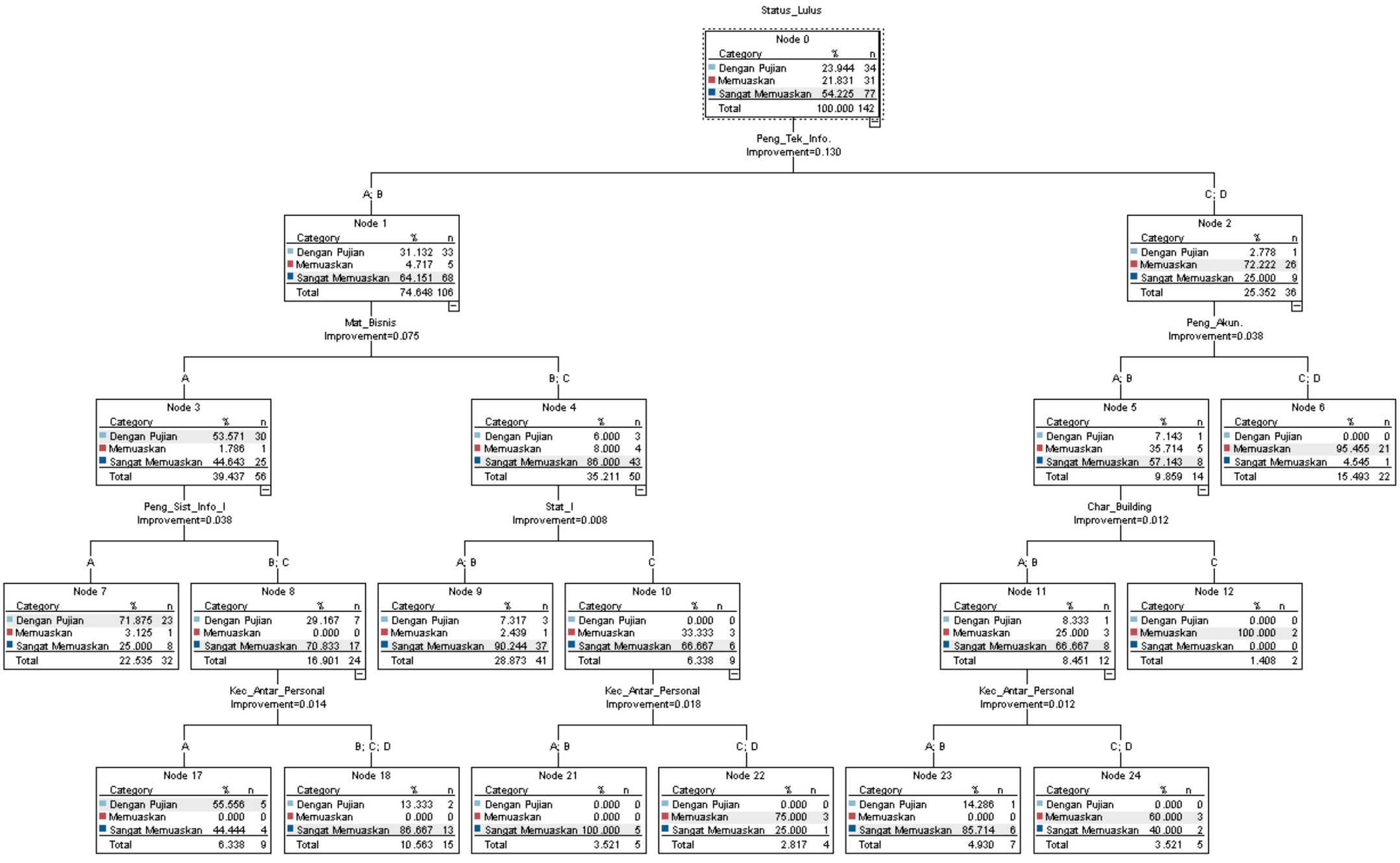
Hasil akhir dari pemodelan ini adalah terbentuknya model CART (*Classification and Regression Trees*), yang dapat dilihat pada gambar 4.12. Dari model seperti pada

gambar 4.12 dapat disimpulkan sementara bahwa dari 142 data latih yang menyandang status lulus dengan predikat *Cumlaude* (Dengan Pujian) sebesar 23,944% (34 mahasiswa), predikat Memuaskan sebesar 21,831% (31 Mahasiswa), sedangkan yang menyandang predikat Sangat Memuaskan sebesar 54,225% (77 mahasiswa).

Sesuai dengan analisis *feature selection* pada gambar 4.10, dari pemodelan CART ini terdapat tujuh (7) variabel prediktor (mata kuliah) yang sangat berpengaruh terhadap hasil prediksi kelulusan mahasiswa, diantaranya:

1. Pengantar Teknologi Informasi, didasarkan atas node 1 dan 2.
2. Matematika Bisnis, didasarkan atas node 3 dan 4.
3. Pengantar Akuntansi, didasarkan atas node 5 dan 6.
4. Pengembangan Sistem Informasi I, didasarkan atas node 7 dan 8.
5. Kecakapan Antar Personal, didasarkan atas node 17, 18, 21, 22, 23, dan 24.
6. Character Building, didasarkan atas node 11 dan 12.
7. Statistik I, didasarkan atas node 9 dan 10.

Contoh cara membaca gambar 4.12 ini dimulai dari node 0 yang merupakan node dasar. Node dasar akan membuat keputusan menjadi node 1 dan node 2 karena variabel Pengantar Teknologi Informasi memiliki nilai kesesuaian yang lebih besar (lihat gambar 4.10), selanjutnya node 1 melakukan keputusan menjadi node 3 dan node 4, dimana pada node ini variabel Matematika Bisnis memiliki nilai kesesuaian yang lebih besar. Kemudian node 3 melakukan keputusan menjadi node 7 dan node 8 dimana variabel Pengembangan Sistem Informasi I memiliki nilai kesesuaian yang lebih besar. Selanjutnya node 7 tidak melakukan keputusan sehingga disebut dengan node terminasi, yang berarti bahwa node tersebut sudah dapat ditentukan klasifikasinya. Dari gambar 4.12 ini dihasilkan *rule* yang diuraikan pada tabel 4.3



Gambar 4. 12
Model prediksi kelulusan mahasiswa algoritma CART

Tabel 4. 3
Rule algoritma CART

No	*Rule*	Status Lulus
1	**JIKA** nilai Peng_Tek_Info = A atau nilai Peng_Tek_Info = B dan nilai Mat_Bisnis = A dan nilai Peng_Sist_Info_I = A **MAKA**	Dengan Pujian
2	**JIKA** nilai Peng_Tek_Info = A atau nilai Peng_Tek_Info = B dan nilai Mat_Bisnis = A dan nilai Peng_Sist_Info_I = B atau nilai Peng_Sist_Info_I = C dan nilai Kec_Antar_Personal = A **MAKA**	Dengan Pujian
3	**JIKA** nilai Peng_Tek_Info = A atau nilai Peng_Tek_Info = B dan nilai Mat_Bisnis = A dan nilai Peng_Sist_Info_I = B atau nilai Peng_Sist_Info_I = C dan nilai Kec_Antar_Personal = B atau nilai Kec_Antar_Personal = C atau nilai Kec_Antar_Personal = D **MAKA**	Sangat Memuaskan
4	**JIKA** nilai Peng_Tek_Info = A atau nilai Peng_Tek_Info = B dan nilai Mat_Bisnis = B atau nilai Mat_Bisnis =C dan nilai Stat_I = A atau nilai Stat_I = B **MAKA**	Sangat Memuaskan
5	**JIKA** nilai Peng_Tek_Info = A atau nilai Peng_Tek_Info = B dan nilai Mat_Bisnis = B atau nilai Mat_Bisnis =C dan nilai Stat_I = C dan nilai Kec_Antar_Personal = A atau nilai Kec_Antar_Personal = B **MAKA**	Sangat Memuaskan
6	**JIKA** nilai Peng_Tek_Info = A atau nilai Peng_Tek_Info = B dan nilai Mat_Bisnis = B atau nilai Mat_Bisnis =C dan nilai Stat_I = C dan nilai Kec_Antar_Personal = C atau nilai Kec_Antar_Personal = D **MAKA**	Memuaskan
7	**JIKA** nilai Peng_Tek_Info = C atau nilai Peng_Tek_Info = D dan nilai Peng_Akun = A atau nilai Peng_Akun = B dan nilai Char_Building = A atau nilai Char_Building = B dan nilai Kec_Antar_Personal = A atau nilai Kec_Antar_Personal = B **MAKA**	Sangat Memuaskan
8	**JIKA** nilai Peng_Tek_Info = C atau nilai Peng_Tek_Info = D dan nilai Peng_Akun = A atau nilai Peng_Akun = B dan nilai Char_Building = A atau nilai Char_Building = B dan nilai Kec_Antar_Personal = C atau nilai Kec_Antar_Personal = D **MAKA**	Memuaskan
9	**JIKA** nilai Peng_Tek_Info = C atau nilai Peng_Tek_Info = D dan nilai Peng_Akun = A atau nilai Peng_Akun = B dan nilai Char_Building = C **MAKA**	Memuaskan
10	**JIKA** nilai Peng_Tek_Info = C atau nilai Peng_Tek_Info = D dan nilai Peng_Akun = C atau nilai Peng_Akun = D **MAKA**	Memuaskan

4.1.5 Fase *Evaluation*

Untuk menilai tingkat akurasi atas model yang sudah dibuat, maka perlu dilakukan pengujian model yang akan menjamin kebenaran dalam melakukan prediksi.

4.1.5.1 Model *Discriminant Analysis*

Dari hasil pengujian *Cross-validation* model *Discriminant Analysis* (Analisis Diskriminan) pada IBM SPSS Statistics 24 diperoleh tingkat akurasi prediksi model ini dalam melakukan fungsi klasifikasi sebesar 82,4%, seperti terlihat pada poin a gambar

4.13. Mengacu pada subbab 2.5 tabel 2.6 mengenai tingkat akurasi klasifikasi, maka model *discriminant analysis* (analisis diskriminan) dalam kasus ini menempati tingkat akurasi *Good*, yang artinya model ini telah layak dan dapat digunakan untuk membedakan ketiga grup status kelulusan mahasiswa tersebut.

Classification Results[a,c]

			Predicted Group Membership			
		Status_Lulus	Memuaskan	Sangat Memuaskan	Dengan Pujian	Total
Original	Count	Memuaskan	35	5	0	40
		Sangat Memuaskan	7	67	15	89
		Dengan Pujian	0	4	43	47
	%	Memuaskan	87.5	12.5	.0	100.0
		Sangat Memuaskan	7.9	75.3	16.9	100.0
		Dengan Pujian	.0	8.5	91.5	100.0
Cross-validated[b]	Count	Memuaskan	33	7	0	40
		Sangat Memuaskan	9	64	16	89
		Dengan Pujian	0	6	41	47
	%	Memuaskan	82.5	17.5	.0	100.0
		Sangat Memuaskan	10.1	71.9	18.0	100.0
		Dengan Pujian	.0	12.8	87.2	100.0

a. 82.4% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 78.4% of cross-validated grouped cases correctly classified.

Gambar 4. 13
Hasil klasifikasi model *Discriminant Analysis*

4.1.5.2 Model *K-Nearest Neighbor*

Berdasarkan pengujian terhadap data latih untuk model *K-Nearest Neighbor* yang dibantu dengan program aplikasi IBM SPSS Modeler 18.0, diperoleh tingkat akurasi prediksi model ini dalam melakukan fungsi klasifikasi sebesar 71,9%, seperti terlihat pada 4.14. Mengacu pada subbab 2.5 tabel 2.6 mengenai tingkat akurasi klasifikasi, maka model *K-Nearest Neighbor* dalam kasus ini menempati tingkat akurasi *Fair*, yang artinya model ini cukup layak digunakan untuk membedakan ketiga grup status kelulusan mahasiswa tersebut.

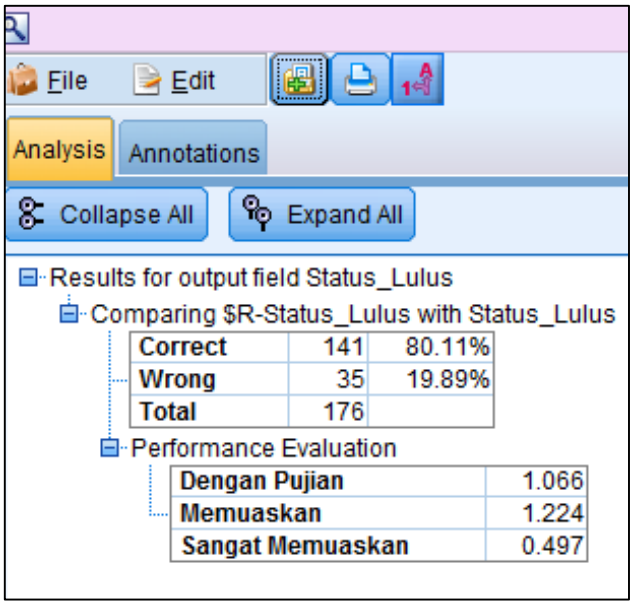
Classification Table

Partition	Observed	Predicted			
		Dengan Pujian	Memuaskan	Sangat Memuaskan	Percent Correct
Training	Dengan Pujian	33	0	4	89.2%
	Memuaskan	0	24	1	96.0%
	Sangat Memuaskan	18	11	30	50.8%
	Overall Percent	42.1%	28.9%	28.9%	71.9%

Gambar 4. 14
Hasil klasifikasi model *K-Nearest Neighbor*

4.1.5.3 Model *Decision Tree*

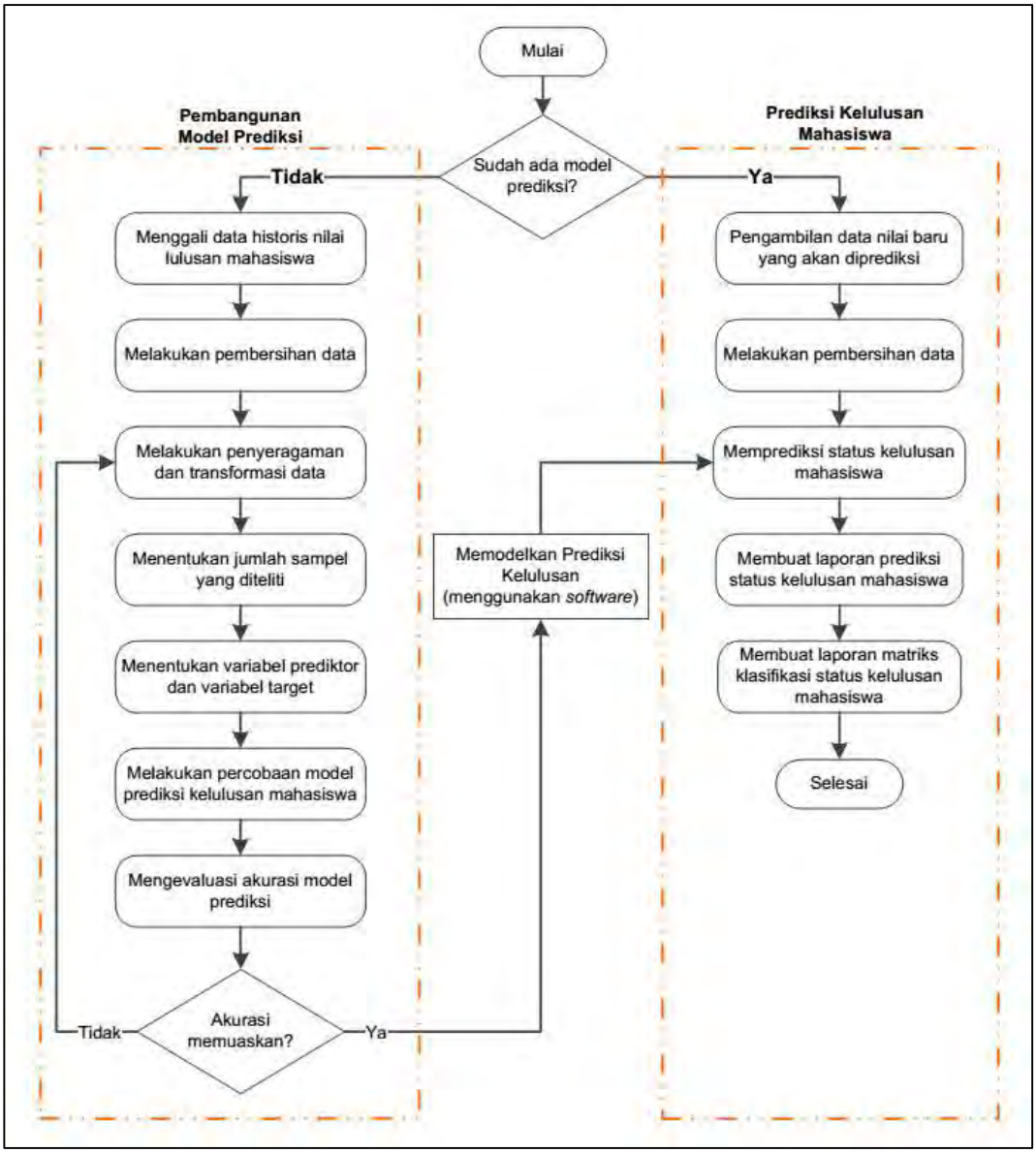
Berdasarkan pengujian terhadap data latih untuk model *Decision Tree* menggunalkan algoritma CART (*Classification and Regression Trees*) yang dibantu dengan program aplikasi IBM SPSS Modeler 18.0, diperoleh tingkat akurasi prediksi model ini dalam melakukan fungsi klasifikasi sebesar 80,11%, seperti terlihat pada 4.15. Mengacu pada subbab 2.5 tabel 2.6 mengenai tingkat akurasi klasifikasi, maka model K-Nearest Neighbor dalam kasus ini menempati tingkat akurasi *Good*, yang artinya model ini telah layak dan dapat digunakan untuk membedakan ketiga grup status kelulusan mahasiswa tersebut.



Gambar 4. 15
Hasil klasifikasi model *Decision Tree* (CART)

4.1.6 Fase *Deployment*

Fase ini merupakan fase terakhir dari siklus yang terdapat dalam CRSIP-DM, dimana nantinya user akan mempergunakan hasil dari pemodelan yang sudah dibuat. Agar dapat dengan sederhana dalam menggunakan model prediksi ini oleh pengguna, maka diperlukannya bagian IT untuk membantu menjalankan model prediksi ini, sesuai dengan yang telah penulis uraikan pada fase *Business Understanding*, adapun langkah prosesnya penulis uraikan pada gambar 4.16.



Gambar 4. 16
Deployment model prediksi status kelulusan mahasiswa

4.2 Pembahasan Hasil Pemodelan

Setelah ketiga model tersebut (*Discriminant Analysis, Decision Tree, dan K-Nearest Neighbor*) dibuat, selanjutnya akan dibandingkan model mana yang akan menghasilkan fungsi paling akurat dan efisien, baik dari proses pembuatan fungsi hingga menghasilkan suatu kalimat matematis yang mudah dimengerti oleh pengguna.

4.2.1 Model *Discriminant Analysis*

Berdasarkan pemodelan pada model *Discriminant Analysis* (Analisis Diskriminan) menghasilkan suatu fungsi yang mampu memetakan objek yang diprediksi sehingga dapat diklasifikasikan menurut grup status kelulusan (Dengan Pujian, Sangat Memuasan, atau Memuaskan). Dari kedua fungsi yang dibentuk oleh model *Discriminant Analysis* (Analisis Diskriminan) pada subbab 4.1.4.1 poin 3, mata kuliah yang berpengaruh dapat dituangkan ke dalam tabel 4.4 mengacu pada dataset awal.

Tabel 4. 4
Daftar mata kuliah berpengaruh pada model *Discriminant Analysis*

No	Mata Kuliah	Semester
1	Bahasa Inggris I	1
2	Pengantar Teknologi Informasi	1
3	Matematika Bisnis	2
4	Statistik I	3
5	Sistem Informasi Manajemen II	3
6	Aplikasi Pengelolaan Proyek	3
7	Pengembangan Sistem Informasi I	3
8	Kecakapan Antar Personal	3

Maka penulis mencoba menjelaskan pemanfaatan fungsi ini yang bertujuan untuk mengetahui se-dini mungkin mahasiswa tingkat awal (semester 1 hingga semester 4) apakah nantinya pada saat lulus, mahasiswa tersebut mampu menyandang predikat *Cumlaude* atau tidak. Adapun pemanfaatan tersebut yaitu:

1. Pada semester 1 dosen wali dapat menginformasikan kepada mahasiswa yang dibimbing bahwa mata kuliah Bahasa Inggris I dan Pengantar Teknologi Informasi diharapkan mendapat nilai baik, agar mahasiswa tersebut dapat memperoleh predikat *Cumlaude* pada saat lulus nanti.
2. Pada semester 2 dosen wali dapat menginformasikan kepada mahasiswa yang dibimbing bahwa mata kuliah Matematika Bisnis diharapkan mendapat nilai baik,

agar mahasiswa tersebut dapat memperoleh predikat *Cumlaude* pada saat lulus nanti.

3. Pada semester 3 dosen wali dapat menginformasikan kepada mahasiswa yang dibimbing bahwa mata kuliah Statistik I, Sistem Informasi Manajemen II , Aplikasi Pengelolaan Proyek, Pengembangan Sistem Informasi I, Kecakapan Antar Personal diharapkan mendapat nilai baik, agar mahasiswa tersebut dapat memperoleh predikat *Cumlaude* pada saat lulus nanti.
4. Agar kedelapan mata kuliah tersebut diperhatikan lebih serius oleh STMIK LIKMI selaku lembaga pendidikan sekolah tinggi, sehingga setiap mahasiswa yang sedang menempuh studi benar-benar mampu fokus memahami sekaligus berkeinginan memperoleh nilai mata kuliah yang baik, khususnya dari kedelapan mata kuliah yang dibentuk dari fungsi diskriminan tersebut.

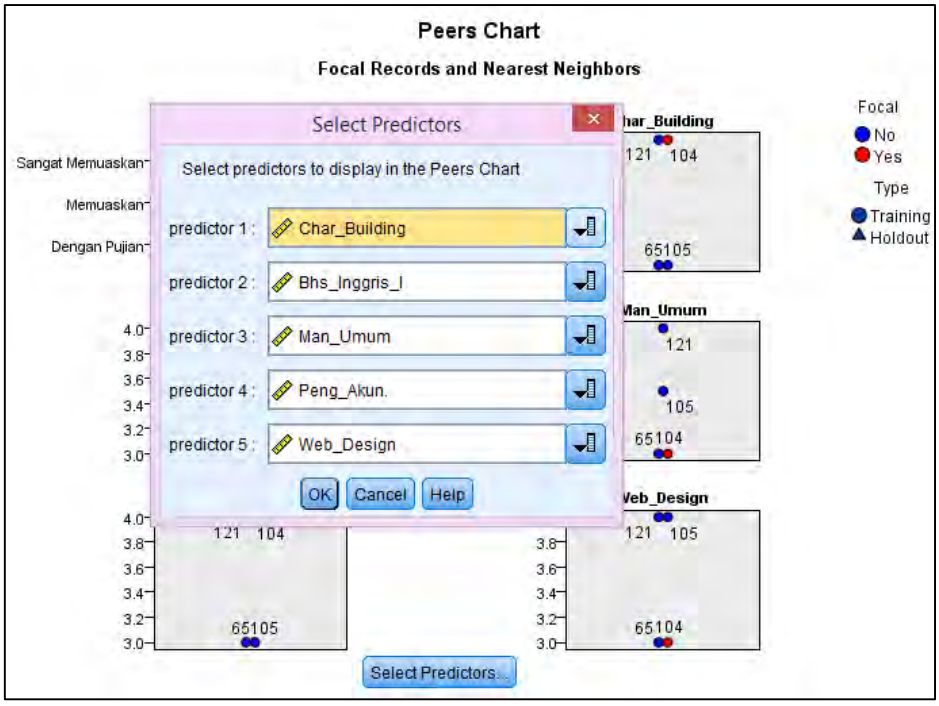
Dengan demikian, model ini mampu menjawab pertanyaan besar penulis pada penelitian ini dalam menentukan suatu pola pencapaian prestasi studi lulusan mahasiswa dari tahun ke tahun, menentukan model untuk menganalisa mahasiswa mana yang berpotensi lulus dengan predikat Memuaskan (rata-rata) se-dini mungkin, dan mengidentifikasi subjek (mata kuliah) apa saja yang paling mempengaruhi kelulusan mahasiswa untuk memperoleh predikat *Cumlaude* (Dengan Pujian) pada saat lulus nantinya.

4.2.2 Model *K-Nearest Neighbor*

Berdasarkan pemodelan yang dibentuk oleh model *K-Nearest Neighbor*, dihasilkan bidang *scatterplot* dan diagram *peers* (lihat gambar 4.7 dan gambar 4.9). Kedua gambar tersebut menampilkan hasil prediksi sebuah objek berdasarkan nilai *K*-terdekat yang dibentuk dari program aplikasi IBM SPSS Modeler 18.0. Selanjutnya dari gambar 4.7 dapat dianalisa bahwa terdapat tiga mata kuliah (Manajemen Umum, Character Building, dan Bahasa Inggris I) yang dianggap mempengaruhi status kelulusan mahasiswa. Bidang *scatterplot* tersebut hanyalah bentuk minimalis dari proyeksi dimensional atas bidang prediktor (*predictor space*). Dari hasil secara utuh (lihat gambar 4.7) ternyata terdapat dua

puluh variabel prediktor sebagai penentunya, yang berarti bahwa seluruh variabel prediktor mempengaruhi status kelulusan mahasiswa, sehingga untuk mengidentifikasi jarak terdekat dengan metode *Euclidean Distance* perlu dilakukan perhitungan pada keseluruhan variabel prediktor yang ada. Hal tersebut dipertegas pada gambar 4.17, dimana pada bagian bawah diagram *peers* terdapat tombol "Select Predictors" yang apabila ditekan maka akan muncul kotak dialog, sehingga dapat dipilih secara manual variabel-variabel prediktor yang dapat ditampilkan pada diagram *peers* tersebut.

Dengan demikian model ini hanya mampu menjawab pertanyaan penulis nomor 1 dan 2 dalam rumusan masalah penelitian ini, tidak mampu menjawab pertanyaan penulis nomor 3; yakni dalam hal mengidentifikasi subjek (mata kuliah) apa saja yang paling mempengaruhi kelulusan mahasiswa untuk memperoleh predikat *Cumlaude* (Dengan Pujian) pada saat lulus nantinya, sehingga hasil dari model ini sulit untuk dibuatkan pemanfaatannya dalam membimbing mahasiswa semester 1 hingga semester 4, sebab seluruh mata kuliah semester 1 hingga semester 4 diusahakan mendapat nilai yang baik agar dapat lulus dengan predikat *Cumlaude* (Dengan Pujian).



Gambar 4. 17
Variabel prediktor K-NN

4.2.3 Model *Decision Tree*

Berdasarkan sepuluh (10) *rule* pada tabel 4.3 yang dibentuk dengan menggunakan algoritma CART (*Classification and Regression Trees*), maka penulis akan mencoba menjelaskan pemanfaatan *rule* tersebut yang bertujuan untuk mengetahui se-dini mungkin mahasiswa tingkat awal (semester 1 hingga semester 4) apakah nantinya pada saat lulus, mahasiswa tersebut mampu menyandang predikat *Cumlaude* atau tidak.

Tabel 4. 5
Daftar mata kuliah berpengaruh pada algoritma CART

No	Mata Kuliah	Semester
1	Pengantar Teknologi Informasi	1
2	Matematika Bisnis	2
3	Pengantar Akuntansi	1
4	Pengembangan Sistem Informasi I	3
5	Kecakapan Antar Personal	3
6	Character Building	1
7	Statistik I	3

Adapun pemanfaatan *rule* tersebut adalah sebagai berikut:

1. Pada semester 1 dosen wali dapat menginformasikan kepada mahasiswa yang dibimbing bahwa mata kuliah Pengantar Teknologi Informasi, Pengantar Akuntansi dan Character Building paling tidak mendapatkan nilai B, agar mahasiswa tersebut dapat memperoleh predikat *Cumlaude* pada saat lulus nanti.
2. Pada semester 2 dosen wali dapat menginformasikan kepada mahasiswa yang dibimbing bahwa mata kuliah Matematika Bisnis diharapkan mendapatkan nilai A, agar mahasiswa tersebut dapat memperoleh predikat *Cumlaude* pada saat lulus nanti.
3. Pada semester 3 dosen wali dapat menginformasikan kepada mahasiswa yang dibimbing bahwa mata kuliah Pengembangan Sistem Informasi I paling tidak mendapatkan nilai B atau C, mata kuliah Kecakapan Antar Personal diharapkan mendapat nilai A, dan mata kuliah Statistik I paling tidak mendapatkan nilai B, agar mahasiswa tersebut dapat memperoleh predikat *Cumlaude* pada saat lulus nanti.

Dengan demikian model ini telah mampu menjawab pertanyaan besar penulis pada penelitian ini dalam menentukan suatu pola pencapaian prestasi studi lulusan mahasiswa dari tahun ke tahun, menentukan model untuk menganalisa mahasiswa mana

yang berpotensi lulus dengan predikat Memuaskan (rata-rata) se-dini mungkin, dan mengidentifikasi subjek (mata kuliah) apa saja yang paling mempengaruhi kelulusan mahasiswa untuk memperoleh predikat *Cumlaude* (Dengan Pujian) pada saat lulus nantinya.

4.2.4 Perbandingan Hasil Pemodelan

Pada bagian ini akan diuraikan mengenai perbandingan hasil pemodelan prediksi kelulusan mahasiswa dari ketiga model di atas: *Discriminant Analysis, K-Nearest Neighbor,* dan *Decision Tree* (algoritma CART) berdasarkan hasil pengamatan penulis selama pemrosesan berlangsung, dapat dilihat pada tabel 4.6.

Tabel 4. 6
Perbandingan hasil model prediksi

Kriteria	Model Prediksi		
	Discriminant Analysis	***K-Nearest Neighbor***	***Decision Tree (CART)***
Waktu pemrosesan	12 detik	17 detik	15 detik
Kemudahan pemrosesan	Mudah	Sedang	Sedang
Akurasi prediksi	82,4%	71,9%	80,11%
Kemudahan pemahaman model/ fungsi / *rule*	Sedang	Cukup sulit	Mudah
Tingkat konstan	Agak konstan	Kurang konstan	Paling konstan

Dalam kasus ini, Model *Decision Tree* adalah yang paling konstan dalam menghasilkan model prediksi. Karena setelah penulis melakukan beberapa kali uji dengan jumlah data yang berbeda, bagan (*tree*) dan *rule*, serta variabel prediktor yang dihasilkan tetap sama dan sangat presisi. Sehingga model ini sangat cocok dan direkomendasikan untuk dituangkan kedalam rancangan sistem informasi prediksi status kelulusan mahasiswa di STMIK LIKMI.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Setelah melakukan proses analisis dan pembahasan hasil model prediksi kelulusan mahasiswa dalam penelitian yang telah dilakukan penulis di STMIK LIKMI hingga penyelesaiannya, maka dapat disimpulkan bahwa:

1. Dalam menentukan pola pencapaian prestasi studi lulusan mahasiswa dari tahun ke tahun yang dihasilkan dari masing-masing model, yakni:
   a. Pada model *Discriminant Analysis*, dihasilkan dua buah fungsi (lihat sub bab 4.1.4.1 hal 38) sebagai penentu status predikat kelulusan seorang mahasiswa (Dengan Pujian, Sangat Memuaskan, atau Memuaskan) berdasarkan perhitungan nilai mata kuliah yang paling berpengaruh dan kemudian disesuaikan nilainya dengan *territorial map* yang terbentuk dari hasil *generate* program aplikasi IBM SPSS Statistics 24 (tersaji dalam lampiran 2.
   b. Pada model *K-Nearest Neighbor*, pola yang terbentuk didasarkan pada nilai K-terdekat yang diperoleh dari proses *Cross-validation*. Akan tetapi nilai K-terdekat ini belum mampu menunjukkan pola pencapaian prestasi studi lulusan mahasiswa dari tahun ke tahun secara jelas dan mendetail.
   c. Pada model *Decision Tree* dengan algoritma CART (*Classification and Regression Trees*) diperoleh sepuluh (10) *rule* yang kemudian dapat digunakan sebagai acuan untuk memprediksi status predikat kelulusan mahasiswa (Dengan Pujian, Sangat Memuaskan, atau Memuaskan), tersaji pada sub bab 4.1.4.3 dalam tabel 4.3 (hal 47).
2. Berkenaan dengan rancangan model seperti apa yang digunakan untuk menganalisa mahasiswa mana yang paling besar kemungkinannya lulus dengan predikat Memuaskan (rata-rata) dan subjek (mata kuliah) mana yang paling mempengaruhi kelulusan mahasiswa untuk memperoleh predikat *Cumlaude* (Dengan Pujian) pada saat lulus, yakni:

a. Pada model *Discriminant Analysis* Pada *output* program aplikasi IBM SPSS Statistics 24, terbentuk rancangan model berupa *territorial map* atas fungsi yang dihasilkan dari data histori, juga terdapat *Standardized Canonical Discriminant Function Coefficients* yang memberikan informasi mengenai mata kuliah apa saja yang menentukan status predikat kelulusan seorang mahasiswa, yaitu: Bahasa Inggris I, Pengantar Teknologi Informasi, Matematika Bisnis, Statistik I, Sistem Informasi Manajemen II, Aplikasi Pengelolaan Proyek, Pengembangan Sistem Informasi I, dan Kecakapan Antar Personal.

b. Pada model *K-Nearest Neighbor* dihasilkan diagram *peers* yang menunjukkan hasil prediksi dengan mengacu pada nilai *K*-terdekat yang terbentuk pada saat proses *Cross-validation*. Akan tetapi pada model ini sulit untuk membaca *rule* prediksi yang dihasilkan dalam menentukan status predikat kelulusan mahasiswa (Dengan Pujian, Sangat Memuaskan, atau Memuaskan). Di samping itu model ini pun belum cukup jelas memberikan gambaran mengenai mata kuliah mana saja yang dapat menentukan status predikat kelulusan seorang mahasiswa.

c. Pada model *Decision Tree* dihasilkan diagram pohon keputusan yang selalu bercabang dua pada setiap noktah keputusan, sesuai dengan ciri dari algoritma CART (*Classification and Regression Trees*). Berdasarkan menu *feature selection* pada IBM SPSS Modeler 18.0, dihasilkan beberapa mata kuliah yang berpengaruh dalam pembuatan model prediksi ini berurutan dimulai dari mata kuliah yang paling menentukan, yaitu: Pengantar Teknologi Informasi, Matematika Bisnis, Pengantar Akuntansi, Pengembangan Sistem Informasi I, Kecakapan Antar Personal, Character Building, dan Statistik I.

3. Berdasarkan perbandingan ketiga model tersebut, model *Decision Tree* dengan algoritma CART (*Classification and Regression Trees*) merupakan model terbaik dan sangat cocok untuk memprediksi status kelulusan mahasiswa STMIK LIKMI, karena kemudahannya dalam pemahaman diagram dan *rule* hasil prediksi, dan

presentase tingkat akurasinya sebesar 80,11%, tidak jauh berbeda dengan akurasi model *Discriminant Analysis* yaitu 82,4%.

5.2 Saran

Berikut ini merupakan beberapa saran yang dapat penulis usulkan dengan tujuan untuk mengembangkan model prediksi kelulusan mahasiswa STMIK LIKMI:

1. Model prediksi kelulusan mahasiswa yang baru dibuat ini diharapkan dapat diterapkan pada STMIK LIKMI, agar mahasiswa lulusannya benar-benar memiliki kualitas yang mampu bersaing secara unggul untuk memenuhi kebutuhan masyarakat dalam lingkungan pekerjaan.
2. Model *K-Nearest Neighbor* dalam penelitian ini agar dapat dikembangkan lebih lanjut, misal menggunakan bantuan program aplikasi lainnya (Clementine, RapidMiner, dan sebagainya) agar menghasilkan model prediksi yang benar-benar valid dan memeiliki akurasi yang baik, serta dapat mudah dipahami oleh pembaca (pengguna).
3. Model prediksi yang telah penulis rancang ini agar dapat dikembangkan lebih lanjut dengan menggunakan variabel-variabel lain sebagai prediktornya, misal: umur mahasiswa saat masuk kuliah semester I, asal daerah (jauh atau dekat), status mahasiswa (menikah atau belum menikah), kelas yang ditempuh (reguler atau karyawan), dan sebagainya.

DAFTAR PUSTAKA

Adnyana, I Made Budi. 2015. *Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus : Stikom Bali)*. CSRID Journal, Vol.8 No.3, Hal. 201-208. ISSN: 2085-1367, Oktober 2015.

Ansari, Rudy. 2016. *Prediksi Kelulusan Mahasiswa Dengan Jaringan Syaraf Tiruan*. JTIULM – Vol.1 , No.1 , Hal. 18-23, Januari-Juni 2016.

BAN-PT, 2011. *Pedoman Penilaian Borang Akreditasi Institusi Perguruan Tinggi*. Jakarta.

Berry, M.W dan M. Browne. 2006. *Lecture Notes in Data Mining*. Singapore: World Scientific.

Chapman, Peter. dkk. 2000. *CRISP-DM 1.0 Step-by-Step Data Mining Guide.*

Gorunescu, F. 2011. *Data Mining Concept Model and Techniques*. ISBN 978-3-642-19720-8. Berlin: Springer.

Han, J , dan Kamber. 2006. *Data Mining Concept and Techniques*. San Fransisco: Morgan Kauffman.

Han, J. dkk. 2012. *Data Mining: Concepts and Techniques 3rd edition*. Massachusetts (US): Morgan Kaufmann.

Kosasi, Sandy. 2014. *Penerapan Metode Jaringan Saraf Tiruan Backpropagation Untuk Memprediksi Nilai Ujian Sekolah*. Jurnal Teknologi, Vol. 7 No.1, Hal. 20-28, Juni 2014.

Larose, Daniel T. 2005. *Discovering Knowledge in Data : An Introduction to Data Mining*. New Jersey: John Willey & Sons, Inc

Larose, Daniel T dan Chantal D. Larose. 2015. *Data Mining and Predictive Analytics*. New Jersey: John Willey & Sons, Inc.

Mambang dan Finki Dona Marleny. 2015. *Prediksi Calon Mahasiswa Baru Mengunakan Metode Klasifikasi Decision Tree*. CSRID Journal, Vol.7 No.1 Februari 2015, Hal. 46-54

Meinanda, Hanief dkk. 2009. *Prediksi Masa Studi Sarjana Dengan Artificial Neural Network*. Internetworking Indonesian Journal. Vol.1 No.2 (2009). ISSN: 1942-9703

Moedjiono dan Asep Saefulloh. 2013. *Penerapan Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu*. Infosys Journal, Vol.2 No.1 Februari 2013.

Pandya, Jayati dan Rutvija Pandya. 2015. *C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning*. International Journal of Computer Applications (0975-8887), Vol.117 No.16 May 2015.

Ristekdikti. 2016. *Data Perguruan Tinggi Wilayah Kpertis-IV*. Dipetik 2 Desember 2017, dari www.ristekdikti.go.id

Salmu, Supardi dan Achmad Solichin. 2017. *Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naive Bayes: Studi Kasus UIN Syarif Hidayahtullah Jakarta*. Prosiding Seminar Nasional Multidisiplin Ilmu Universitas Budi Luhur. ISSN : 2087-0930.

Santoso, Singgih. 2017. *Statistik Multivariat dengan SPSS*. Jakarta: PT Elex Media Komputindo.

Sokolova, M dan Guy Lapalme. 2009. *A Systematic Analysis of Performance Measures for Classification Tasks*. Information Processing and Management, Vol. 45 No.4, Hal. 427-437.

Sumathi, S. dan Sivanadam. 2006. *Introduction to Data Mining and its Applications*. Verlag Berlin Heidelberg: Springer.

Susandi, Dony, dkk. 2016. *Prediksi Kelulusan Mahasiswa Pada Perguruan Tinggi Kabupaten Majalengka Berbasis Knowledge*. Seminar Nasional Telekomunikasi Dan Informatika (Selisik 2016) ISSN : 2503-2844

Susanto, Sani dan Suryadi D. 2010. *Pengantar Data Mining Menggali Pengetahuan dari Bongkahan Data*. CV. ANDI OFFSET.

Tan, dkk. 2006. *Introduction to Data Mining*. Pearson Education Inc.

Wahana. 2010. *Mengolah Data Statistik Hasil Penelitian dengan IBM SPSS Statistics 24*. Yogyakarta: Wahana Komputer dan Andi.

Wendler, Tilo dan Soren Grottrup. 2016. *Data Mining with SPSS Modeler*. Switzerland: Springer.

Witten, Ian H. dkk. 2011. *Data Mining: Pratical Machine Learning Tools and Techniques, 3rd Edition*. Burlington: Morgan Kaufmann.