**TESIS – KS142501**
**PEMILIHAN FITUR UNTUK KLASIFIKASI LOYALITAS PELANGGAN TERHADAP MEREK PRODUK*FAST MOVING CONSUMER GOODS***
**(Studi Kasus: Mie Instan)**

HENI SULISTIANI
5214201009

DOSEN PEMBIMBING
Dr. Ir. Aris Tjahyanto, M.Kom.
NIP. 196503101991021001

PROGRAM MAGISTER
JURUSAN SISTEM INFORMASI
FAKULTAS TEKNOLOGI INFORMASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2016

**THESES – KS142501**
**FEATURE SELECTION FOR CLASSIFICATION OF CUSTOMER LOYALTY TO PRODUCTS BRAND ON FAST MOVING CONSUMER GOODS (Case Study: Instant Noodless)**

HENI SULISTIANI
5214201009

SUPERVISOR
Dr. Ir. Aris Tjahyanto, M.Kom.
NIP. 196503101991021001

MAGISTER PROGRAM
MAJOR IN INFORMATION SYSTEM
FACULTY OF INFORMATION TECHNOLOGY
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2016

# LEMBAR PENGESAHAN TESIS

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Komputer (M.Kom)
di
Institut Teknologi Sepuluh Nopember

oleh:

Heni Sulistiani
NRP. 5214201009

Tanggal Ujian : 13 Juli 2016
Periode Wisuda : September 2016

Disetujui Oleh:

Dr. Ir. Aris Tjahyanto, M.Kom.
NIP. 19650310 199102 1 001 (Pembimbing)

Dr. Eng. Febriliyan Samopa, S.Kom., M.Kom.
NIP. 19730219 199802 1 001 (Penguji I)

Mahendrawathi E. R., ST, M.Sc., Ph.D.
NIP. 19761011 200604 2 001 (Penguji II)

Direktur Program Pascasarjana

Prof. Ir. Djauhar Manfaat, M.Sc., Ph.D.
NIP. 19601202 198701 1 001

# PEMILIHAN FITUR UNTUK KLASIFIKASI LOYALITAS PELANGGAN TERHADAP MEREK PRODUK PADA *FAST MOVING CONSUMER GOODS* (Studi Kasus: Mie Instan)

Nama Mahasiswa : Heni Sulistiani
NRP : 5214201009
Pembimbing : Dr. Ir. Aris Tjahyanto, M.Kom.

## ABSTRAK

Pemilihan fitur merupakan salah satu bagian penting dan teknik yang sering digunakan dalam praproses penggalian data yang membawa efek langsung untuk mempercepat algoritma penggalian data dan meningkatkan kinerja pertambangan seperti akurasi prediksi dan hasil yang komprehensif. Penelitian ini membahas mengenai pemilihan subset fitur dalam klasifikasi loyalitas pelanggan terhadap merek bagi pengguna *fast moving consumer goods*(dalam penelitian ini mengambil studi kasus pada salah satu produknya yaitu mie instan) dan melakukan analisis terhadap fitur-fitur yang mempengaruhi performa klasifikasi pohon keputusan.

Data yang digunakan pada penelitian ini merupakan hasil penyebaran kuisoner kepada para pelanggan mie instan di Propinsi Lampung. Data yang diperoleh memiliki fitur yang bersifat heterogen, untuk itu dilakukan pengubahan fitur menjadi fitur homogen. Dalam penelitian ini, mengkombinasikan metode UFT (*unsupervised feature transformation*) dan metode DMI (*dynamic mutual information*) untuk seleksi fitur. Metode UFT digunakan untuk transformasi fitur non-numerik menjadi fitur numerik, sehingga fitur yang bersifat heterogen menjadi fitur homogen. Metode DMI digunakan untuk pemilihan fitur. Hasil transformasi fitur diklasifikasikan menggunakan algoritmapohon keputusan. Hasil klasifikasi digunakan untukmelakukan perbandingan performa antara dataset sebelum pemilihan fitur, setelah dilakukanpemilihan fitur menggunakan metode DMI, *p-Value* dan perkiraan peneliti.

Dari hasil pengujian terhadap model prediksi klasifikasi diperoleh fitur-fitur yang mempengaruhi performa klasifikasi pohon keputusanloyalitas pelanggan. Peningkatan performa tersebut dapat dilihat pada pengimplementasian metode pemilihan fitur DMIdengan jumlah fitur sebanyak lima. Nilai akurasi, presisi, *recall* dan *f-measure* mengalami peningkatan bila dibandingkan dengan penggunaan seluruh fitur (sebelum dilakukan pemilihan fitur), metode pemilihan fitur *p-value* dan hasil perkiraan, masing-masing nilai tersebut secara berturut-turut adalah sebesar 76.68%, 74.4%, 76.7% dan 73.5%.Fitur-fitur yang berpengaruh tersebut antara lain jumlah pengeluaran, rata-rata konsumsi, usia, alamat dan alasan berpindah merek.

Kata kunci: klasifikasi, loyalitas pelanggan, mutual informasi, transformasi fitur

*(lembar ini sengaja dikosongkan)*

# FEATURE SELECTION FOR CLASSIFICATION OF CUSTOMER LOYALTY TO PRODUCTS BRAND ON FAST MOVING CONSUMER GOODS (Case Study: Instant Noodless)


By : Heni Sulistiani
Student Identify Number : 5214201009
Supervisor : Dr. Ir. Aris Tjahyanto, M.Kom.


## ABSTRACT


Feature selection is one of the important parts and techniques used in data mining preprocess to bring immediate effect in accelerate the data mining algorithms and improve the performance of mining such as the prediction accuracy and comprehensive results. This study discusses the subset features selection in the classification of customer loyalty to the brand for the fast moving consumer goods (this study took a case study on one of its products, i.e instant noodles) and an analysis of the features that affect the performance classification of decision tree.

The used data in this study is the result of spread questionnaires to customers instant noodles in Lampung Province. The obtained data has a heterogeneous features, it is neededto carried out the transformation of features into a homogeneous features. In this study, we combine UFT (unsupervised feature transformation) and DMI (dynamic mutual information)methods for features selection. UFT methods used for transformation of non-numerical features into a numerical features, so heterogeneous features became homogeneous features. DMI methods used for feature selection.Feature transformation result is classified using decision trees algorithm. The results of classification is used to performance comparisons between the datasets before the feature selection, after the feature selection using DMI, p-Value and researchers estimate.

The test results of the predictive models of classification obtained the features that affect the decision tree algorithm performance of customer loyalty. The performance enhancement can be seen in the implementation of the DMI feature selection method with a number of features as many as five features. Value of accuracy, precision, recall and F-measure increased when compared to the use of all features (prior to the selection of features), methods of feature selection p-value and methods of researcher's estimate, respectively of values is 76.68%, 74.4 %, 76.7% and 73.5%. The features that affect the performance of classification, ie expenditures, average of consumption, age of costumer, address and the reason for switching brands.



Keywords : classification, customer loyalty, mutual information, feature transformation

*(lembar ini sengaja dikosongkan)*

# KATA PENGANTAR

Segala puji bagi Allah atas rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan tesis yang berjudul **"Pemilihan Fitur untuk Klasifikasi Loyalitas Pelanggan Terhadap Merek Produk pada *Fast Moving Consumer Goods* (Studi Kasus: Mie Instan)"** sebagai salah satu syarat kelulusan dari Program Pascasarjana Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember Surabaya. Proses pengerjaan tesis ini telah banyak mendapatkan bantuan, bimbingan, masukan serta dukungan dari berbagai pihak. Sehingga dalam kesempatan ini, penulis mengucapkan terima kasih kepada:

1. Kedua orang tuaku, adikku dan seluruh keluarga besar yang selalu memberikan do'a, motivasi, semangat serta dukungan setiap saat dan tanpa batas.
2. Bapak Dr. H. M. Nasrullah Yusuf, S.E., M.B.A., selaku Ketua Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Teknokrat Lampung.
3. Bapak Dr. H. Mahathir Muhammad, S.E., M.M., selaku Ketua Yayasan Pendidikan Tinggi Teknokrat Lampung.
4. Bapak Dr. Ir. Aris Tjahyanto, M.Kom., selaku pembimbing yang telah memberikan waktu, motivasi, saran dan imunya selama proses membimbing sehingga tesis ini dapat terselesaikan dengan baik.
5. Bapak Dr. Eng. Febriliyan Samopa, S.Kom., M.Kom., selaku dosen penguji I dan Ibu Mahendrawathi E. R., ST, M.Sc., Ph.D., selaku penguji II yang telah banyak memberikan masukan dan saran dalam perbaikan tesis ini.
6. Keluarga Teknokrat-ITS Lampung, Ibu Damay, Ajeng, Ryan, Donaya, Mbak Ayu yang selalu mendukung dan memberikan motivasi serta semangat dalam penyelesaian tesis ini dan menjadi keluarga dalam suka dan duka.
7. Keluarga besar S2 SI ITS 2014 yang selalu kompak dan berbagi ilmu serta pemberian motivasi dalam proses belajar mulai dari kuliah penyegaran sampai dengan terselesaikannya tesis ini.
8. Teman-teman S2 SI ITS angkatan 2012, 2013 dan 2015 yang memberikan motivasi dan masukan dalam penulisan tesis ini.

9. Teman-teman jurusan Teknik Informatika ITS 2014 yang membantu dalam proses penyelesaian Tesis ini.
10. Sahabat-sahabat dan adik-adik tersayang di Lampung yang selalu menghibur dikala sedih dan memberikan motivasi.
11. Dosen dan karyawan/wati Perguruan Tinggi Teknokrat Lampung.
12. Dosen dan Karyawan/wati Institut Teknologi Sepuluh Nopember Surabaya.
13. Seluruh responden dan informan penelitian.

Penulis menyadari bahwa tesis ini masih memiliki kekurangan dan ketidaksempurnaan dalam penulisan. Oleh karena itu, penulis mengharapkan kritik dan saran dari pembaca yang bersifat membangun sebagai bahan acuan penelitian selanjutnya. Akhir kata, semoga penelitian ini dapat memberikan manfaat bagi kita semua.

Surabaya, Juli 2016

Penulis

# DAFTAR ISI

# DAFTAR GAMBAR

Gambar Halaman

*(lembar ini sengaja dikosongkan)*

# DAFTAR TABEL

Tabel Halaman

*(lembar ini sengaja dikosongkan)*

# DAFTAR LAMPIRAN

# BAB 1
# PENDAHULUAN

Pada bab awal ini dijelaskan mengenai gambaran penelitian dari latar belakang penelitian, rumusan masalah, tujuan penelitian, batasan penelitian, kontribusi penelitian, hingga sistematika penulisan dokumen.

## 1.1. Latar Belakang

Data Kantar Worldpanel Indonesia menunjukkan bahwa dari tahun 2012 ke 2013 telah terjadi peningkatan penjualan produk-produk*Fast Moving Consumer Goods* (FMCG) sebesar 14% di seluruh Indonesia, baik di kawasan perkotaan (urban) maupun pedesaan (rural)(Sundari, 2014). Sepanjang tahun 2014, pasar FMCG di Indonesia menunjukkan tingkat pertumbuhan yang tinggi sebesar 15%. Indonesia menjadi pasar potensial untuk produk-produk FMCG dengan pertumbuhan industri FMGC yang masih menunjukkan angka dua digit dengan peningkatan jumlah kelas menengah dan rata-rata usia penduduk Indonesia relatif muda(Musriadi, 2014).Dalam menghadapi pasar terbuka ASEAN (Masyarakat Ekonomi ASEAN/MEA), para pelaku (industri FMCG)harus semakin banyak mengeksplore pola-pola pengembangan bisnisnya karena pasar akan semakin ketat dan banyak tantangan(Kurniawan, 2014). Dengan semakin tinggi dan ketatnya tingkat persaingan antar perusahaan tersebut, akan menyebabkan perusahaan pada umumnya berusaha untuk mempertahankan kelangsungan hidup(Rohman, 2015). Oleh sebab itu diperlukan strategi pemasaran yang tepat untuk dapat mempertahankan pelanggan agar mampu bertahan di lingkungan pasar. Salah satu strategi pemasaran yang tepat untuk bertahan dari persaingan yang tinggi adalah dengan mempertahankan loyalitas pelanggannya(Santoso, 2012) dan salah satu bentuk informasi penting yang dibutuhkan oleh produsen suatu produk konsumtif adalah informasi tentang tingkat kesetiaan konsumen terhadap merek dagang atau merek produk yang dijualnya serta kekuatan suatu merek terhadap konsumennya (Ariwibowo, 2013).

Mempertahankan pelanggan akan menghasilkan pendapatan dan margin yang lebih tinggi dari pelanggan baru (Buckinx & Poel, 2005). Sehingga, tindakan mengurangi pembelotan pelanggan atau mempertahankan loyalitas pelanggan dapat memberikan dampak yang sangat besar bagi perusahaan. Untuk memprediksi dan menganalisa loyalitas pelanggan dapat menggunakan metode klasifikasi. Prediksi loyalitas pelanggan menggunakan metode klasifikasi terhadap perusahaan retail diusulkan oleh Buckinx & Poel, (2005) menggunakan metode *logistic regression, ARD Neural Network* dan *random forest*. Penelitian tersebut menggunakan metode analisis regresi dalam pemilihan fitur untuk mengklasifikasikan loyalitas pelanggan. Pemilihan fitur yang didasarkan pada kriteria validasi statistik, belum tentu menyebabkan model dapat mengoptimalkan target yang ditetapkan oleh organisasi masing-masing (Maldonado, Flores, Verbraken, Baesens, & Weber, 2015). Model prediksi tersebut juga dibangun tanpa menggunakan variabel atau fitur kepuasan pelanggan, dikarenakan variabel kepuasan pelanggan tidak tersedia dalam *database* dan peneliti hanya memfokuskan pada perbandingan model prediksi. Sedangkan meningkatkan kepuasan pelanggan merupakan salah satu strategi untuk mencapai keunggulan bersaing di pasar (Aktepe, Ersoz, & Toklu, 2014). Salah satu rujukan yang paling sering disebutkan dalam literatur kepuasan pelanggan adalah loyalitas pelanggan. Loyalitas pelanggan dinyatakan sebagai kemungkinan untuk merekomendasikan perusahaan kepada pelanggan lain, kemungkinan untuk membeli kembali atau kembali dari pelanggan(Anderson & Mittal, 2014).

Memprediksi loyalitas pelanggan dapat dilakukan dengan menggunakan teknik penggalian data. Kesuksesan proses penemuan informasi dalam penggalian datadipengaruhi oleh beberapa faktor. Salah satu faktor kuncinya adalah kualitas data(Purbasari & Nugroho, 2013). Jika data memiliki terlalu banyak *noise*, atau banyak data yang redundan dan tidak relevan, proses pelatihan penemuan informasi akan mengalami kesulitan. Pada praproses penggalian data, pemilihan fitur adalah salah satu bagian yang penting (Blum & Langley, 1997)untuk mengurangi jumlah fitur, menghilangkan fitur yang tidak relevan, redundansi, atau *noise*, dan membawa efek langsung untuk aplikasi yaitu mempercepat algoritma penggalian data, meningkatkan kinerja pertambangan seperti akurasi

prediksi dan hasil yang komprehensif (Liu & Yu, 2005). Memilih subset fitur yang baik tidak hanya mengurangi beban komputasi, tetapi juga dapat meningkatkan akurasi(Forman, 2003). Yu dan Liu (2006) mendefinisikan fitur optimum terdiri dari semua fitur kuat yang relevan dan fitur lemah relevan tetapi tidak berlebihan.Dengan menggunakan fitur yang relevan, algoritma klasifikasi dapat secara umum meningkatkan akurasi prediksi mereka, mempersingkat periode pembelajaran, dan bentuk konsep sederhana(Liu & Setiono, 1997).Fitur dianggap relevan bila nilainya bervariasi secara sistematis dengan keanggotaan kategori(Hall, 2000). Tabel 1.1 menunjukkan perbandingan hasil akurasi prediksi antara menggunakan fitur asli dan fitur yang telah dipilih.

Tabel 1.1. Perbandingan Hasil Akurasi Prediksi

| **Penulis** | **Tahun** | **Domain** | **Tanpa Pemilihan Fitur** | | **Menggunakan Metode Pemilihan Fitur** | |
|---|---|---|---|---|---|---|
| | | | **Jumlah Fitur** | **Akurasi** | **Jumlah Fitur** | **Akurasi** |
| Chih-Fong Tsai, Mao-Yuan Chen | 2010 | Customer Churn Prediction | 22 | 89,30% | 12 | 93,49% |
| Lei Yu, Huan Liu | 2006 | Klasifikasi *high dimensional data* | 57 | 80,83% | 5 | 87,50% |
| | | | 59 | 86,91% | 4 | 87,73% |
| | | | 62 | 94,14% | 6 | 93,48% |
| | | | 68 | 98,27% | 2 | 98,08% |
| | | | 86 | 93,97% | 3 | 94,02% |
| | | | 151 | 94,65% | 4 | 95,51% |
| | | | 169 | 96,79% | 2 | 91,33% |
| | | | 280 | 67,25% | 6 | 72,79% |
| | | | 618 | 79,10% | 23 | 75,77% |
| | | | 650 | 94,30% | 14 | 95,06% |
| Setyoningsih Wibowo | 2014 | Klasifikasi Loyalitas Pelanggan | 20 | 86,04% | 9 | 91,52% |

Langkah pertama dan sering menantang dalam data proses penambangan melibatkan pemilihan fitur dan transformasi(Tremblay, Berndt, & Studnicki, 2006). Seringkali, ketika proses penambangan disajikan dengan jumlah atribut yang tidak sedikit, banyak atribut yang tidak berguna untuk prediksi, sementara yang lain mungkin hanya berlebihan. Pemilihan fitur dapat ditemukan di banyak bidang penggalian data seperti klasifikasi, *clustering*, aturan asosiasi, dan regresi(Liu & Yu, 2005).Data biasanya diperoleh dari berbagai sumber dan berisi

fitur yang beragam, seperti fitur numerik dan fitur non-numerik. Hal tersebut mengakibatkan sulitnya untuk mengevaluasi fitur beragam secara bersamaan(Wei, Chow, & Chan, 2015b). Beberapa metode diusulkan untuk pemecahan masalah pemilihan fitur yang beragam. Sebagai contoh, penelitian mengenai analisis regresi yang digunakan untuk mengidentifikasi variabel penting yang mempengaruhi perilaku pembelian pelanggan di perusahaan jasa dan menggunakan rantai Markov untuk model probabilitas transisi dari perubahan perilaku (Cheng, Chiu, Cheng, & Wu, 2012). Namun analisis regresi memiliki kelemahan yaitu sulit untuk menginterpretasikan koefisien *intercept* dan bila tidak berhati-hati akan mengakibatkan interpretasi yang tidak sesuai dengan kondisi yang sebenarnya (Novita, 2008).

Suatu ukuran probabilitas diperkenalkan dan diusulkan *mixed forward selection* untuk pemilihan fitur yang beragam (Tang & Mao, 2007). Metode ini beroperasi dengan cara membagi fitur ke dalam kelompok numerik dan kelompok non-numerik untuk evaluasi dan kemudian menghasilkan bermacam-macam pilihan subset fitur. Namun, metode ini hanya memungkinkan fitur numerik dan fitur non-numerik dievaluasi secara terpisah (Wei, Chow, & Chan, 2015b). Sebuah usulan pemilihan fitur juga diusulkan oleh Liu & Setiono(1997), metode diskritisasi untuk memilih fitur langsung dari atribut numerik. Metode ini membagi nilai-nilai dari fitur numerik ke beberapa interval dan mewakili mereka dengan berbagai nilai-nilai non numerik. Namun, metode diskritisasi sering menyebabkan hilangnya informasi karena mengurangi jarak dan urutan dalam fitur numerik yang asli(Hu, Yu, Liu, & Wu, 2008; Chong & Wong, 1995). Metode *rough set theory* juga dapat digunakan untuk melakukan evaluasi fitur yang beragam(Slowinski & Vannderpooten, 2000; Pawlak & Skowron, 2007). Namun, metode tersebut memiliki kompleksitas komputasi yang tinggi atau bahkan NP-Hard (Li, Chow, & Tang, 2014).

Penelitian ini mengusulkan metode *unsupervised feature transformation* (UFT) yang mampu mengubah fitur non-numerik menjadi fitur numerik untuk pemilihan subset fitur dari dataset yang sifatnya heterogen. UFT dapat digunakan pada tahap praproses data dan menyatukan fitur heterogen.Transformasi ini hanya bergantung pada fitur non-numerik aslinya dan dapat menghindari bias informasi

pada label kelas(Wei, Chow, & Chan, 2015a). Dibandingkan dengan metode transformasi fitur lainnya,UFT memperkenalkan distorsi minimal informasi danlebih dapat diandalkan karena data yang diolah bebas bias. Untuk meminimalkan distorsi informasi dilakukan penghitungan keterkaitan antara fitur dengan label kelas menggunakan mutual informasi(MI). MI telah banyak digunakan karena memiliki dua keuntungan, yaitu MI dapat mengukur berbagai jenis relasi termasuk nonlinier dan MI kuat terhadap fitur yang mengandung *noise*(Li W. , 1990). Dan perlu dicatat bahwa di antara berbagai kriteria evaluasi, metode mutual informasimerupakan metrik yang efektif untuk skala relevansi antara fitur, mencapai kinerja yang sangat baik dan telah lebih banyak menarik perhatian(Qian & Shu, 2015).

Transformasi fitur merupakan cara lain untuk mengatasi pemilihan subset fitur yang beragam. Metode pemilihan subset fitur optimum dapat dilakukan dengan menggunakan metode filter. Metode *filter* dapat menghemat waktu dan mampu menangani masalah *over-fitting* yang disebabkan oleh ketergantungan klasifikasi pada *clasifier*(Zhang, Chen, Liang, & Li, 2008). Metode *filter* juga mampu untuk memperkirakan mutual informasiantara subset fitur dengan label kelas, beberapa contoh penelitian yang menggunakan metode mutual informasi dalam pemilihan fitur adalah *minumum redudancy maximum relevance* (mRMR)(Peng, Long, & Ding, 2005) dan *normalized mutual information feature selection* (NMIFS) (Estevez, Tesmer, Perez, & Zurada, 2009)untuk pemilihan subset fitur yang heterogen. Namun, kedua metode tersebut tidak dapat memberikan MI antara subset fitur dan label kelas secara langsung karena bergantung pada nilai redundansi yang diperkirakan oleh MI secara indivual (Chow & Huang, 2005).Dalam penelitian ini mengusulkan pemilihan fitur untuk data heterogen yang telah ditransformasi menjadi data homogen berdasarkan nilai mutual informasi menggunakan *dynamic mutual information (DMI)*, metode DMI mampu mengurangi redundansi dan data yang tidak relevan serta mampu memberikan mutual informasi antara subset fitur dengan label kelas secara langsung (Liu, Sun, Liu, & Zhang, 2009).

Hasil dari pencarian fitur optimum akan diuji coba dalam metode klasifikasi pohon keputusan, guna melihat hasil kinerja dari pencarian tersebut.

Bila dibandingkan dengan metode klasifikasi lainnya (SVM, *Naive Bayes, K-NN)* metode pohon keputusan memiliki beberapa kelebihan, yaitu pohon keputusan lebih mudah dipahami dan diinterpretasikan dalam bentuk pohon, membutuhkan sedikit ruang penyimpanan untuk data latih dibandingkan metode klasifikasi yang lain, dapat digunakan untuk tipe data kategorikal dan numerik, merepresentasikan model seperti *white box* (proses logika keputusannya dapat diikuti dengan mudah mengikuti arah dalam pohon keputusan), algoritmanya handal, cepat dan memproses dengan baik pada data latih yang banyak serta akurasinya dapat dibandingkan dengan metode klasifikasi yang lain bagi banyak data set sederhana(Prasetyo, 2014). Penelitian ini berfokus pada pemilihan fitur yang mempengaruhi dalam pembentukan pohon keputusan untuk mengklasifikasikan loyalitas pelanggan terhadap merek produk *fast moving consumer goods* dan melakukan analisis perbandingan dari beberapa metode pemilihan fitur.

## 1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan di atas, maka permasalahan utama pada penelitian dapat dirumuskan sebagai berikut:

a. Bagaimana menentukan fitur optimum yang paling terkait dengan label kelas dalam analisis loyalitas pelanggan terhadap merek produk bagi pelanggan *fast moving consumer goods?*
b. Bagaimana mengklasifikasikan loyalitas pelangganterhadap merek produk dengan menggunakan fitur yang telah diperoleh dari permasalahan pertama?

## 1.3. Tujuan Penelitian

Sesuai dengan rumusan masalah, tujuan dari penelitian ini adalah:

a. Menentukan fitur optimum yang paling terkait dengan label kelas untuk pengklasifikasian loyalitas pelanggan terhadap merek produk.
b. Membandingkan kinerja pemilihan fitur pada klasifikasi loyalitas pelanggan terhadap merek produk.
c. Merepresentasikan hasil klasifikasi loyalitas pelanggan dalam aturan IF – THEN.

## 1.4. Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah membantu perusahaan dalam proses pengambilan keputusan untuk menemukan faktor-faktor relevan yang dapat mempengaruhi loyalitas pelanggan terhadap merek dagang atau merek produk. Penelitian ini juga diharapkan dapat dijadikan dan kajian lebih lanjut pada penelitian-penelitian metode klasifikasi dan loyalitas pelanggan.

## 1.5. Batasan Penelitian

Batasan penelitian meliputi hal-hal di bawah ini:

a. Penelitian menggunakan data survey dengan penyebaran kuisoner terhadap pelanggan yang membeli atau pengguna barang konsumen yang bergerak cepat atau barang yang tidak tahan lama (dalam penelitian ini produk yang dijadikan studi kasus adalah mie instan).
b. Data yang akan digunakan meliputi profil pembeli, data psikografis, data transaksi, data produk dan promosi.

## 1.6. Kontribusi Penelitian

Kontribusi dari penelitian ini adalah mengubah fitur yang sifatnya heterogen menjadi fitur homogen untuk memudahkan dalam mengevaluasi dan menemukan fitur optimum atau fitur relevan yang paling terkait dengan label kelasdari fitur beragam dalammembangun model prediksi klasifikasi loyalitas pelanggan *fast moving consumer goods*. Selain itu, model prediksi klasifikasi pada penelitian ini melibatkan beberapa variabel baru yang belum digunakan pada usulan model prediksi klasifikasi loyalitas pelanggan sebelumnya, yaitu data produk dan promosi.

## 1.7. Sistematika Penulisan Dokumen

Sistematika penulisan dokumen laporan penelitian tesis ini dibagi menjadi lima bab, yaitu sebagai berikut:

**BAB 1 PENDAHULUAN**

Pada bab ini dijelaskan mengenai latar belakang, rumusan masalah, tujuan penelitian, batasan penelitian, kontribusi penelitian dan sistematika penulisan.

**BAB 2 LANDASAN TEORI DAN KAJIAN PUSTAKA**

Pada bab ini dijelaskan mengenai kajian pustaka dari berbagai penelitian yang berkaitan dengan penelitian ini. Kajian pustaka ini bertujuan untuk memperkuat dasar dan alasan dilakukannya penelitian ini. Selain kajian pustaka, pada bab ini juga dijelaskan mengenai teori-teori terkait yang bersumber dari buku, jurnal, maupun artikel yang berfungsi sebagai dasar dalam melakukan penelitian agar dapat memahami konsep atau teori penyelesaian permasalahan yang ada. Pada bab ini terdapat uraian mengenai penggalian data, pohon keputusan, pemilihan fitur dan manajemen hubungan pelanggan.

**BAB 3 METODOLOGI PENELITIAN**

Pada bab ini dijelaskan mengenai langkah-langkah penelitian beserta metode yang digunakan. Langkah-langkah penelitian akan dijelaskan dalam sebuah diagram alur yang sistematis dan akan dijelaskan tahap demi tahap.

**BAB 4 HASIL DAN PEMBAHASAN**

Pada bab ini akan dilakukan uji coba terhadap tahap pemilihan fitur berdasarkan skenario uji coba yang telah dirancang sebelumnya. Selain itu, pada bab ini juga dijelaskan mengenai hasil uji coba.

**BAB 5 KESIMPULAN DAN SARAN**

Pada bab ini berisi kesimpulan dari penelitian dan saran bagi penelitian berikutnya yang berasal dari kekurangan ataupun temuan dari penelitian ini.

# BAB 2
# LANDASAN TEORI DAN KAJIAN PUSTAKA

Pada bab ini dijelaskan mengenai teori-teori yang mendasari penelitian dan kajian pustaka mengenai penelitian-penelitian yang terkait. Teori yang dijelaskan antara lain mengenai penggalian data, pohon keputusan, pemilihan fitur dan manajemen hubungan pelanggan. Sedangkan penelitian-penelitian terkait yang dikaji antara lain penelitian mengenai pemilihan fitur untuk klasifikasi dan loyalitas pelanggan menggunakan teknik penggalian data.

## 2.1. Penelitian Terkait

Pemilihan fitur merupakan tahap praproses yang penting untuk menjamin tingkat akurasi yang tinggi, efisiensi, dan skalabilitas untuk proses klasifikasi, terutama ketika berurusan dengan satu set data yang besar atau bahkan luar biasa besar (Huang & Chow, 2005). Pemilihan fitur merupakan metode yang telah banyak digunakan untuk mendapatkan informasi penting dalam dataset untuk target tertentu (Blum & Langley, 1997). Pada klasifikasi, tujuan pemilihan fitur adalah untuk menemukan fitur optimum yang paling terkait dengan label kelas dan dengan demikian mengurangi redundansi atau informasi yang tidak berguna (Vergara & Este'vez, 2014). Usulan pemilihan fitur pada tahap praproses juga diusulkan oleh Tsai & Chen (2010) dalam pengembangan model prediksi *customer churn* pada *Multimedia on Demand* (MOD). Aturan asosiasi digunakan dalam pemilihan fitur dan model dikembangkan menggunakan metode *neural network* (NN) dan *decision tree* (DT). Hasil menunjukkan bahwa model NN dan DT yang diikuti dengan aturan asosiasi selama tahap praproses dapat memberikan hasil prediksi yang lebih baik. Dibandingkan dengan NN, DT memiliki kinerja yang lebih baik. Chow & Huang (2005) mengusulkan estimasi fitur yang optimal berbasis mutual informasi yang menggabungkan *pruned parzen windows estimator* dan *quadratic mutual information*. Fitur yang optimal atau yang mendekati optimal dapat diidentifikasi secara efektif dan diperkirakan dengan cara yang sistematis.

Wei, Chow, & Chan (2015a) melakukan penelitian untuk menyeleksi fitur yang beragam menggunakan transformasi fitur berdasarkan mutual informasi. Metode pemilihan fitur berdasarkan mutual informasi konvensial tidak dapat menangani fitur yang beragam karena perbedaan format ataupun karena perkiraan metode MI antara fitur dengan label kelas. Untuk memecahkan masalah tersebut, diusulkan metode transformasi fitur yang dapat mengubah fitur non-numerik menjadi fitur numerik. Usulan *unsupervised feature transformation* (UFT) melakukan transformasi fitur dengan menggunakan distribusi *Gaussian*, setiap subkelompok fitur non-numerik asli menggunakan kelompok nilai numerik yang mematuhi distribusi *Gaussian* untuk menggantikannya. Pada usulan tersebut menjelaskan bahwa fitur non-numerik ditransformasi dengan fitur numerik. Sementara itu, fitur numerik asli dari dataset dinormalisasikan untuk mengurangi skala yang berbeda pada data.

Karena pentingnya peningkatan kepuasan pelanggan dalam lingkungan bisnis saat ini, banyak perusahaan yang berfokus pada gagasan loyalitas pelanggan dan profitabilitas untuk meningkatkan pangsa pasar dan kepuasan pelanggan(Kim, Jung, Suh, & Hwang, 2006).Aktepe, Ersoz, & Toklu (2015) mengusulkan metode baru untuk mengidentifikasi kepuasan dan loyalitas pelanggan menggunakan algoritma klasifikasi dan model persamaan terstruktur. Dalam penelitian ini pelanggan dikelompokkan menjadi empat berdasarkan tingkat kepuasan dan loyalitas pelanggan dengan menganalisis kriteria dan kelompok menggunakan metode yang diusulkan. Peneliti menggunakan lima belas kriteria dalam dalam mengevaluasi kelompok pelanggan dengan membuat kuisoner yang disebarkan ke dua ratus responden. Kriteria yang berhubungan dengan loyalitas pelanggan adalah kualifikasi umum merek, penampilan fisik, iklan dari merek produk menggunakan TV dan internet, teknologi baru dari merek, nama merek/brand, kepercayaan pada informasi yang diberikan oleh perusahaan manufaktur, tingkat kepercayaan umum terhadap kualitas.

Metode penggalian data juga dilakukan untuk memprediksi nilai *lifetime* pelanggan sebuah perusahaan perbaikan dan pemeliharaan mobil di Taiwan, prediksi tersebut diestimasi menggunakan data demografis pelanggan dan histori transaksi pembelian (Cheng, Chiu, Cheng, & Wu, 2012). Penelitian tersebut

mengusulkan kerangka yang berisi tiga kelompok, teknik untuk mendapatkan perkiraan *lifetime* pelanggan dari histori transaksi nasabah. Kelompok 1 : menggunakan model *logistic regression* dan model pohon keputusan untuk memperkirakan probabilitas *churn* pelanggan dan untuk memprediksi panjang *lifetime* pelanggan. Kelompok 2 : analisis regresi untuk mengidentifikasi variabel penting yang mempengaruhi perilaku pembelian pelanggan, dan rantai Markov untuk model probabilitas transisi dari perubahan perilaku. Kelompok 3 menggunakan dua algoritma *neural network*yaitu *Backpropagation Neural Network* (BPN) dan*Radial Basis Function Network* (RBFN)untuk memprediksi keuntungan yang diberikan oleh seorang pelanggan berdasarkan berbagai perilaku pembelian. Kim, Jung, Suh, & Hwang (2006) menganalisis nilai *lifetime* pelanggan dan melakukan segmentasi pelanggan berdasarkan nilai pelanggan. Penelitian dilakukan untuk perusahaan komunikasi nirkabel, metode yang digunakan untuk analisis tersebut adalah pohon keputusan, *neural network* dan *logistic regression*. Variabel yang digunakan adalah data demografis pelanggan dan informasi penggunaan layanan nirkabel.

Loyalitas pelanggan dapat dinyatakan sebagai kemungkinan untuk merekomendasikan perusahaan kepada pelanggan lain, kemungkinan untuk membeli kembali atau kembali untuk menjadi pelanggan(Anderson & Mittal, 2014). Penelitian yang membahas tentang retensi pelanggan menggunakan teknik klasifikasi dan klustering telah dilakukan oleh Chu, Tsai, & Ho (2007) untuk memprediksi *customer churn* pada perusahaan telekomunikasi. Prediksi *customer churn* direpresentasikan menggunakan model pohon keputusan, kemudian membangun sebuah strategi retensi pelanggan menggunakan teknik klustering yaitu melakukan segmentasi pelanggan melalui modifikasi *GrowingHierarchical Self-Organizing Map* (GHSOM). Dalam membentuk model *customer churn* digunakan data histori dari *database* pelanggan.

Selain pada perusahaan telekomunikasi, penelitian mengenai pembelotan pelanggan secara parsial juga dilakukan di perusahaan retail, khususnya untuk barang konsumen yang bergerak cepat (Buckinx & Poel, 2005). *Logisticregression, automatic relevance determination* (ARD*) Neural Networks* dan*Random Forests* digunakan untuk mengklasifikasi pelanggan yang setia dan

tidak setia. Dalam memprediksi pembelotan pelanggan, variabel perilaku pelanggan (*Recency, Frequency* dan *Monetary*) lebih berpengaruh daripada variabel demografis. Mengetahui pembelotan parsial sedini mungkinakan menghasilkan informasi yang lebih penting daripada memprediksi jumlah pembelotan yang akan terjadi. Dengan demikian manajer pemasaran akan dapat mengetahui pelanggan yang tidak atau kurang setia terhadap perusahaan, sehingga dapat mengeksekusi tindakan-tindakan untuk mencegah terjadinya pembelotan.

## 2.2. Penggalian Data

Penggunaan komputer pada masyarakat telah meningkatkan kemampuan untuk menghasilkan dan mengumpulkan data dari berbagai sumber. Sejumlah besar data membanjiri hampir setiap aspek kehidupan kita. Data telah menghasilkan kebutuhan mendesak untuk teknik-teknik baru dan alat-alat otomatis yang cerdas membantu dalam mengubah data dalam jumlah besar menjadi informasi yang berguna dan pengetahuan.Penggalian data adalah proses menemukan pengetahuan menarik dari sejumlah besar data yang tersimpan baik di *database*, gudang data, atau repositori informasi lainnya (Han, Kamber, & Pei, 2012). Penggalian data juga didefinisikan sebagai proses menemukan pola dalam data(Witten, Frank, & Hall, 2011). Penggalian data (langkah analisis penemuan pengetahuan dalam basis data) merupakan teknologi baru yang kuat ditingkatkan dan begitu cepat berkembang. Ini adalah teknologi yang dengan potensi besar untuk membantu bisnis dan perusahaan untuk berfokus pada informasi yang paling penting dari data yang mereka punya dan harus mengumpulkan untuk mengetahui perilaku pelanggan mereka(AL-Nabi & Ahmed, 2013).

Berikut beberapa pengertian mengenai penggalian data yang diuraikan oleh Prasetyo (2014) mempunyai beberapa maksud yang mirip :

a. Pencarian otomatis pola dalam basis data yang besar, menggunakan teknik komputasional campuran dari statistik, pembelajaran mesin dan pengenalan pola;
b. Pengekstrakan implisit non-trivial, yang sebelumnya belum diketahui secara potensial adalah informasi berguna dari data;
c. Ilmu pengekstrakan informasi yang berguna dari set data atau basis data besar

d. Eksplorasi otomatis atau semiotomatis dan analisis dalam jumlah besar, dengan tujuan untuk menemukan pola yang bermakna;
e. Proses penemuan informasi otomatis dengan mengidentifikasi pola dan hubungan ‘tersembunyi’ dalam data.

Penggalian data mencakup analisis kecerdasan buatan dan atau analisis statistik, yang biasanya diterapkan pada data skala besar. Analisis statistik tradisional meliputi pendekatan yang biasanya diarahkan, dalam arti hasil-hasilnya telah diduga sebelumnya. Pendekatan ini disebut pendekatan *supervised*. Namun, penggalian data lebih dalam daripada sekedar perangkat-perangkat teknik yang digunakannya. Penggalian data mengandung semangat penemuan pengetahuan (*knowledge discovery*), yaitu belajar hal baru dan yang berguna, yang disebut sebagai pendekatan *unsupervised*. Banyak hasil yang dapat dicapai dengan menggunakan cara-cara otomatis, sebagaimana kita akan lihat, misalnya dalam analisis pohon keputusan (*decision tree analysis*). Tetapi penggalian data tidak terbatas hanya pada analisis terotomasi. Penemuan pengetahuan oleh manusia dapat ditingkatkan dengan perangkat grafis dan identifikasi pola-pola yang tidak diduga sebelumnya melalui kombinasi antara interaksi manusia dan komputer. Penggalian data berkembang dengan cepat, mendatangkan banyak manfaat bagi bisnis (Olson & Shi, 2008).

Penggalian data dapat dilakukan dengan menggunakan asosiasi, klasifikasi, prediksi, pola-pola sekuensial dan urutan waktu serupa (s*imiliar time sequences*) (Olson & Shi, 2008). Dalam asosiasi, hubungan hal tertentu dalam suatu transaksi data dengan hal lain dalam transaksi yang sama digunakan untuk memprediksi pola. Dalam klasifikasi, metode-metodenya ditujukan untuk pembelajaran fungsi-fungsi berbeda yang memetakan masing-masing data terpilih ke dalam salah satu dari kelompok kelas yang telah ditetapkan sebelumnya. Analisis pengelompokkan mengambil data yang belum dikelompokkan dan menggunakan teknik-teknik otomatis untuk mendapatkan data tersebut ke dalam berbagai kelompok. Analisis prediksi berhubungan dengan teknik-teknik regresi. Analisis pola sekuensial mencoba untuk menemukan pola-pola serupa dalam transaksi data selama suatu periode bisnis. Sedangkan urutan waktu serupa

diterapkan untuk menemukan urutan yang mirip dengan urutan yang telah diketahui baik dalam periode bisnis yang lalu maupun yang sekarang.

Metodologi penggalian data terdiri dari visualisasi data, pembelajaran mesin, teknik statistik, dan basis data deduktif(Lee, Chiu, Chou, & Lu, 2006). Aplikasi yang terkait dengan menggunakan metodologi ini dapat diringkas sebagai klasifikasi, prediksi, *clustering*, *summarization*, pemodelan ketergantungan, analisis keterkaitan, dan analisis sekuensial (Fayyad, 1996). Teknologi bagian dari penggalian data terdiri dari teknik seperti metode statistik, jaringan saraf, pohon keputusan, algoritma genetika, dan metode non-parametrik(Lee, Chiu, Chou, & Lu, 2006). Di antara aplikasi yang disebutkan di atas, pengamatan klasifikasi merupakan salah satu metode yang memainkan peranan penting dalam pengambilan keputusan bisnis karena aplikasi dalam pendukung keputusan yang luas, peramalan keuangan, deteksi penipuan, strategi pemasaran, pengendalian proses, dan bidang terkait lainnya (Chen, Han, & Yu, 1996;Fayyad, 1996).

Ngai, Xiu, & Chau (2009) mengidentifikasi delapan puluh tujuh artikel yang berhubungan dengan penerapan teknik penggalian data di CRM yang diterbitkan antara tahun 2000 dan 2006. Dari jumlah tersebut, 51,9% (28 artikel) dan 44,4% (24 artikel) yang terkait dengan program pemasaran dan loyalitas. Tabel 2.1 menunjukkan pembagian penelitian atau artikel berdasarkan CRM dan penggalian data. Sedangkan tabel 2.2 menunjukkan pembagian artikel berdasarkan teknik penggalian data. Berdasarkan tabel tersebut dapat diketahui bahwa model klasifikasi adalah model yang paling umum diterapkan di CRM untuk memprediksi perilaku pelanggan di masa depan. Hal ini tidak mengherankan, karena pemodelan klasifikasi dapat digunakan untuk memprediksi efektivitas atau profitabilitas dari strategi CRM melalui prediksi perilaku pelanggan.

Teknik penggalian data, seperti jaringan saraf dan pohon keputusan, bisa digunakan untuk mencari segmen pelanggan yang menguntungkan melalui analisis karakteristik dasar pelanggan. Teknik pohon keputusan dan aturan asosiasi memiliki peringkat popularitas setelah aplikasi jaringan saraf dalam CRM. Logika dari kedua teknik tersebut dapat diikuti dengan lebih mudah oleh

orang-orang bisnis daripada jaringan saraf.Pohon keputusan telah menjadi sangat populer untuk memecahkan tugas klasifikasi karena mereka dapat menangani dengan prediktor yang diukur pada tingkat pengukuran yang berbeda (termasuk variabel nominal) dan karena kemudahan penggunaan dan dapat diinterpretasikan(Duda, Hart, & Stork, 2001). Inti dari penggalian data adalah memiliki berbagai perangkat yang tersedia untuk membantu analis dan pengguna untuk mengetahui komponen-komponen data dengan lebih baik. Setiap metode memiliki cara yang berbeda, dan biasanya secara tidak langsung menyatakan bahwa masalah tertentu paling baik ditangani dengan jenis algoritma tertentu. Namun, terkadang jenis algoritma yang berbeda dapat digunakan untuk masalah yang sama. Kebanyakan di antaranya menyertakan pengaturan parameter, yang dapat menjadi penting dalam efektivitas metodenya. Lebih jauh lagi, keluaran hasil perlu diinterpretasikan maknanya.

### 2.2.1. Proses Penggalian Data

Secara sistematis, terdapat tiga langkah utama dalam penggalian data (Prasetyo, 2014) :

a. Eksplorasi/pemrosesan awal data: terdiri dari ‘pembersihan’ data, normalisasi data, penanganan data yang salah, reduksi dimensi, pemilihan fitur dan sebagainya.
b. Membangun model dan melakukan validasi terhadapnya: melakukan analisis berbagai model dengan kinerja prediksi yang terbaik. Dalam langkah ini digunakan metode-metode seperti klasifikasi, regresi, analisis pengelompokkan, deteksi anomali, analisis asosiasi, analisis pola sekuensial dan sebagainya.
c. Penerapan: menerapkan model pada data yang baru untuk menghasilkan perkiraan/prediksi masalah yang diinvestigasi.

Tabel 2.1. Pembagian Penelitian atau Artikel CRM dan Penggalian Data

| Dimensi CRM | Elemen CRM | Model Penggalian Data | Jumlah | | |
|---|---|---|---|---|---|
| Identifikasi Pelanggan | Segmentasi Pelanggan | | 8 | | |
| | | Klasifikasi | | 2 | |
| | | Klustering | | 5 | |
| | | Regresi | | 1 | |
| | Analisis Target Pelanggan | | 5 | | |
| | | Klasifikasi | | 3 | |
| | | Klustering | | 1 | |
| | | Visualisasi | | 1 | |
| | | | | | **13** |
| Daya Tarik Pelanggan | Pemasaran Langsung (*Direct Marketing)* | | 7 | | |
| | | Regresi | | 1 | |
| | | Klasifikasi | | 5 | |
| | | Klustering | | 1 | |
| | | | | | **7** |
| Retensi Pelanggan | Manajemen Komplain | | 2 | | |
| | | Klustering | | 1 | |
| | | *Sequence Discovery* | | 1 | |
| | Program Loyalitas | | 24 | | |
| | | Klasifikasi | | 20 | |
| | | Klustering | | 1 | |
| | | Regresi | | 2 | |
| | | *Sequence Discovery* | | 1 | |
| | Pemasaran *one-to-one* | | 28 | | |
| | | Asosiasi | | 13 | |
| | | Klasifikasi | | 7 | |
| | | Klustering | | 5 | |
| | | *Sequence Discovery* | | 3 | |
| | | | | | **54** |
| Membangun Pelanggan | Nilai *Lifetime* Pelanggan | | 5 | | |
| | | Klasifikasi | | 1 | |
| | | Klustering | | 2 | |
| | | Peramalan | | 1 | |
| | | Regresi | | 1 | |
| | Analisis Keranjang Belanja | | 6 | | |
| | | Asosiasi | | 4 | |
| | | *Sequence Discovery* | | 2 | |
| | *Up/Cross Selling* | | 2 | | |
| | | Asosiasi | | 1 | |
| | | *Sequence Discovery* | | 1 | |
| | | | | | 13 |
| **Total** | | | **87** | **87** | **87** |

Sumber : Ngai, Xiu, & Chau, 2009

Tabel 2.2 Pembagian artikel berdasarkan teknik penggalian data

| Teknik Penggalian Data | Jumlah |
|---|---|
| *Neural Network* | 30 |
| *Decision Tree* | 23 |
| *Association Rules* | 18 |
| *Regression* | 10 |
| *Genetic Algorithm* | 4 |
| *Markov Chain* | 4 |
| *Survival Analysis* | 4 |
| *K Means* | 3 |
| *K Means Neighbour* | 3 |
| *Bayesian Network Classifier* | 2 |
| *If-Then-Else* | 1 |
| *Set Theory* | 1 |
| *Support Vector Machine* | 1 |
| *Attribute Oriented Induction* | 1 |
| *Constructive Assignment* | 1 |
| *Customer Map* | 1 |
| *Data Envelopment Analysis* | 1 |
| *Data Mining by Evolutionary Learning* | 1 |
| *Expectation Max* | 1 |
| *Expectation Max Mod* | 1 |
| *Farthest First* | 1 |
| *Goal Oriented Sequential Pattern* | 1 |
| *Latent Class Model* | 1 |
| *Logical Analysis of Data* | 1 |
| *MARFS1/S2* | 1 |
| *Mixture Transition Distribution* | 1 |
| *Multi-Classifier Class Combiner* | 1 |
| *Multivariate Adaptive Regression Splines* | 1 |
| *Online Analytical Mining* | 1 |
| *Outlier Detection* | 1 |
| *Pattern Based Cluster* | 1 |
| *Rule-Based RIPPER* | 1 |
| *S-Means* | 1 |
| *S-Means Mod* | 1 |
| **Total*** | **125** |

* : tiap artikel mungkin menggunakan lebih dari satu teknik penggalian data

Sumber : Ngai, Xiu, & Chau, 2009

### 2.2.2. Data Set

Kata data dalam terminologi statistik adalah kumpulan objek dengan atribut-atribut tertentu, di mana objek tersebut adalah individu berupa data di

mana setiap data memilih sejumlah atribut (Prasetyo, 2014). Atribut tersebut berpengaruh pada dimensi data, semakin banyak atribut/fitur maka semakin besar dimensi data. Kumpulan data-data membentuk data set. Berikut adalah tiga jenis data set yang dikenal dan masing-masing penggolongannya:

a. *Record*
    1) Matriks data
    2) Data transaksi
    3) Data dokumen
b. *Graph*
    1) *World wide web*
    2) Struktur molekul
c. *Ordered data set*
    1) Data spasial
    2) Data temporal
    3) Data sekuensial
    4) Data urutan genetik (*genetic sequence*)

Dalam data set berbentuk *record* data, tidak ada hubungannya dengan baris data dengan baris data yang lain dan juga tidak punya hubungan dengan data set yang lain. Setiap baris data berdiri sendiri sebagai sebuah individu. Dalam sistem basis data, umumnya ada sejumlah tabel yang saling berhubungan menggunakan suatu kunci, tetapi dalam data set *record* data, diasumsikan bahwa hanya ada satu tabel yang berisi sejumlah baris data. Data grafik direpresentasikan dalam bentuk grafik atau diagram. Informasi yang diberikan dalam bentuk gambar dengan jenis tertentu, seperti rumus kimia, link HTML, struktur molekul dan sebagainya. Sedangkan *ordered data* adalah data-data yang tersusun dengan cara dalam urutan atau aturan tertentu, misalnya data struktur DNA mempunyai urutan genetik tertentu, data rekam medis seorang pasien di puskesmas/rumah sakit dengan pola terurut penyakit yang diderita dan sebagainya.

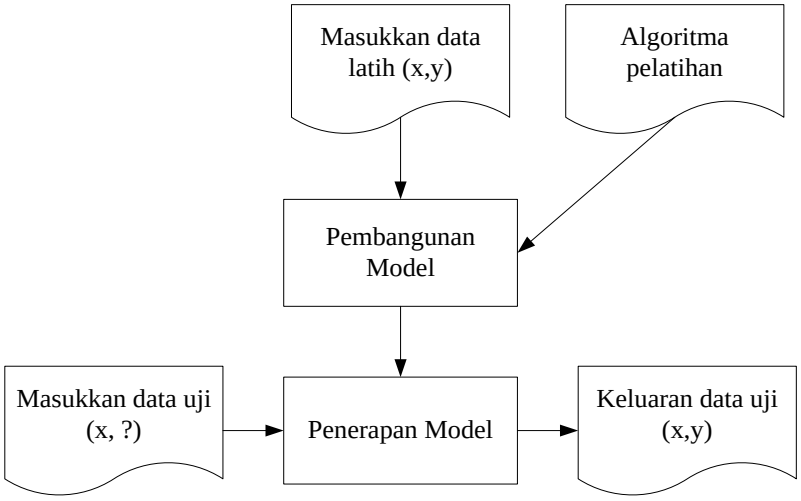
### 2.2.3. Konsep Klasifikasi

Olson & Shi (2008) menjelaskan bahwa metode-metode dalam klasifikasi dapat secara otomatis memprediksi kelas dari data lain yang belum diklasifikasikan. Dua masalah penelitian utama yang berkaitan dengan hasil klasifikasi adalah evaluasi kesalahan klasifikasi dan kekuatan prediksi. Teknik-teknik matematika yang sering kali digunakan untuk membangun metode-metode klasifikasi adalah pohon keputusan biner, jaringan saraf tiruan, pemrograman linier dan statistik. Klasifikasi dan prediksi adalah dua bentuk analisis data yang dapat digunakan untuk menggambarkan ekstrak model yang penting pada kelas data atau untuk memprediksi tren data masa depan (Patel & Rana, 2014).

Teknik klasifikasi dalam penggalian data yang mampu memproses data dalam jumlah yang besar. Hal ini dapat memprediksi kategoris label kelas dan data mengklasifikasikan berdasarkan set pelatihan dan kelas label dan karenanya dapat digunakan untuk mengelompokkan data yang baru tersedia. Dengan demikian dapat dijelaskan sebagai bagian tak terelakkan dari penggalian data dan mendapatkan popularitas yang lebih (AL-Nabi & Ahmed, 2013). Klasifikasi adalah teknik yang digunakan secara luas di berbagai bidang, termasuk data pertambangan, yang tujuannya adalah untuk mengklasifikasikan set besar objek ke dalam kelas yang telah ditetapkan, dijelaskan oleh satu set atribut, menggunakan metode pembelajaran yang terawasi. Karena ledakan pertumbuhan antara basis data bisnis dan ilmiah, penggalian aturan efisiensi klasifikasi dari basis data tersebut sangat penting (Mastrogiannisa, Boutsinas, & Giannikos, 2009). Zhang, Chen, Liang, & Li (2008) mengungkapkan bahwa analisis klasifikasi adalah menganalisis data dalam basis data, untuk membuat deskripsi yang akurat atau membangun model yang akurat atau menambang aturan pengklasifikasian untuk setiap kategori, dan kemudian menggunakan aturan untuk mengklasifikasikan catatan dalam basis data lainnya.

Klasifikasi dapat didefinisikan secara detail sebagai suatu pekerjaan yang melakukan pelatihan/pembelajaran terhadap fungsi target $f$ yang menetapkan setiap vektor (set fitur) $x$ ke dalam satu dari sejumlah label kelas $y$ yang tersedia. Pekerjaan pelatihan tersebut akan menghasilkan suatu model yang kemudian disimpan sebagai memori (Prasetyo, 2014). Model dalam klasifikasi mempunyai

arti yang sama dengan *blackbox*, di mana ada suatu model yang menerima masukan kemudian mampu melakukan pemikiran terhadap masukan tersebut dan memberikan jawaban sebagai keluaran dari hasil pemikirannya. Kerangka kerja klasifikasi ditunjukkan pada gambar 2.1. Pada gambar tersebut, disediakan sejumlah data latih (*x, y*) untuk digunakan sebagai data membangun model, kemudian menggunakan model tersebut untuk memprediksi kelas dari data uji (*x, ?*) sehingga data uji (*x, ?*) diketahui kelas *y* yang seharusnya.




Gambar2.1. Kerangka kerja klasifikasi (Prasetyo, 2014)

Model yang sudah dibangun pada saat pelatihan kemudian dapat digunakan untuk memprediksi label kelas dari data baru yang belum diketahui label kelasnya. Dalam pembangunan model selama proses pelatihan tersebut diperlukan adanya suatu algoritma untuk membangunnya yang disebut sebagai algoritma pelatihan (*learning algorithm*). Kerangka kerja seperti yang ditunjukkan pada gambar 2.1 meliputi dua langkah proses yaitu induksi dan deduksi. Induksi merupakan suatu langkah untuk membangn klasifikasi dari data latih yang diberikan atau disebut juga dengan proses pelatihan, sedangkan deduksi merupakan suatu langkah untuk menerapkan model tersebut pada data uji sehingga data uji dapat diketahui kelas yang sesungguhnya atau disebut juga dengan proses prediksi.

Berdasarkan cara pelatihan, algoritma-algoritma klasifikasi dapat dibagi menjadi dua macam, yaitu *lazy learner* dan *eager learner*. Algoritma-algoritma yang masuk kategori *lazy learner* hanya sedikit melakukan pelatihan (atau bahkan tidak sama sekali). Algoritma-algoritma ini hanya menyimpan sebagian atau seluruh data latih, kemudian menggunakan data latih tersebut ketika proses prediksi. Hal ini mengakibatkan proses prediksi menjadi lama karena model harus membaca kembali semua data latihnya untuk dapat memberikan keluaran label kelas dengan benar pada data uji yang diberikan. Kelebihan dari algoritma seperti ini adalah proses pelatihan berjalan dengan cepat. Algoritma-algoritma klasifikasi yang masuk kategori ini di antaranya adalah *rote classifier*, K-*Nearest Neighbor* (K-NN), *Fuzzy K-Nearest Neighbor* (FK-NN), regresi linear dan sebagainya.

Sedangkan algoritma-algoritma yang masuk kategori *eager learner* didesain untuk melakukan pembacaan/pelatihan/pembelajaran pada data latih untuk dapat memetakan dengan benar setiap vektor masukan ke label kelas keluarannya sehingga di akhir proses pelatihan, model sudah dapat melakukan pemetaan dengan benar semua data latih ke label kelas keluarannya. Setelah proses pelatihan tersebut selesai, maka model (biasanya berupa bobot atau sejumlah nilai kuantitatif tertentu) disimpan sebagai memori, sedangkan semua data latihnya dibuang. Proses prediksi dilakukan menggunakan model yang tersimpan dan tidak melibatkan data latih sama sekali. Cara ini mengakibatkan proses prediksi dapat berjalan dengan cepat, namun harus dibayar dengan proses pelatihan yang lama. Algoritma-algoritma klasifikasi yang masuk kategori ini di antaranya *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM), *Decision Tree, Bayesian* dan sebagainya.

## 2.3. Pohon Keputusan

Pohon keputusan (*decision tree*), dalam konteks penggalian data, merupakan struktur pohon dari aturan-aturan (yang sering disebut aturan asosiasi) (Olson & Shi, 2008). Pohon keputusan juga dapat diartikan sebagai pohon yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari masalah yang dimasukkan (Prasetyo, 2014). Pohon yang dibentuk tidak selalu berupa pohon biner. Jika semua fitur dalam data set menggunakan dua macam nilai

kategorikal maka bentuk pohon yang didapatkan berupa pohon biner. Jika dalam fitur berisi lebih dari dua macam nilai kategorikal atau menggunakan tipe numerik maka bentuk pohon yang didapatkan biasanya tidak berupa pohon biner. Istilah algoritma ini mengikuti metafora "pohon". Memiliki akar, yang merupakan titik perpecahan pertama dari atribut data untuk bangunan pohon keputusan. Ia juga memiliki daun, sehingga setiap jalur dari akar daun akan membentuk aturan yang mudah dimengerti(Tu, Shin, & Shin, 2009).

Proses pohon keputusan pada penggalian data meliputi pengumpulan variabel yang menurut analis mungkin berpengaruh dalam pembuatan keputusan dan analisis variabel-variabel tersebut atas kemampuannya memprediksi hasil akhir. Pohon keputusan sangat berguna untuk mendapatkan pemahaman lebih mendalam mengenai perilaku pelanggan dan mencari cara untuk menindaklanjuti hasil-hasilnya agar mendapatkan keuntungan tambahan(Olson & Shi, 2008). Generasi algoritma pohon keputusan juga tidak memerlukan informasi tambahan selain yang sudah terkandung dalam data pelatihan (misalnya, pengetahuan domain atau pengetahuan sebelumnya dari distribusi pada data atau kelas)(Rastogi & Shim, 2000). Pohon keputusan membutuhkan dua jenis data: pelatihan dan pengujian. Data pelatihan, yang biasanya lebih besar dari bagian data pengujian yang digunakan untuk membangun pohon. Semakin banyak data pelatihan yang dikumpulkan, maka semakin tinggi keakuratan hasil. Kelompok data yang lain, pengujian, digunakan untuk mendapatkan tingkat akurasi dan tingkat kesalahan klasifikasi pohon keputusan(Tu, Shin, & Shin, 2009). Metode ini relatif lebih unggul dibandingkan algoritma jaringan saraf tiruan dan genetika karena menyediakan aturan-aturanyang dapat dipakai ulang sehingga menjelaskan keimpulan dari model (Michie, 1998).

Banyak contoh penerapan pohon keputusan dalam penggalian data untuk bisnis, termasuk mengklasifikasi permohonan pinjaman, pemilihan pelanggan yang potensial dan penilaian pelamar kerja. Pohon keputusan memberikan cara untuk mengimplementasikan pendekatan-pendekatan sistem berbasis aturan. Pohon keputusan mempunyai tiga pendekatan klasik, yaitu:

a. Pohon klasifikasi, digunakan untuk melakukan prediksi ketika ada data baru yang belum diketahui label kelasnya. Pendekatan ini yang paling banyak dilakukan;
b. Pohon regresi, ketika hasil prediksi dianggap sebagai nilai nyata yang mungkin akan didapatkan;
c. CART (atau C&RT), ketika masalah klasifikasi dan regresi digunakan bersama-sama.

### 2.3.1. Aturan Pohon Keputusan

Meskipun terdapat data yang bersifat kategorik, data tersebut berpotensi memiliki banyak aturan. Metode pohon keputusan mampu mengidentifikasikan aturan-aturan yang berguna dalam memprediksi hasil akhir. Efektivitas aturan diukur dalam hal konfidensi (*confidence*) dan dukungan (*support*). Konfidensi merupakan derajat akurasi suatu aturan, sementara dukungan adalah derajat di mana kondisi-kondisi tertentu terjadi dalam data. Dukungan (*support*) untuk suatu aturan asosiasi menunjukkan bagian dari data yang memenuhi sekelompok atribut pada aturan asosiasinya. Tingkat konfidensi (*confidence levels*) dan tingkat dukungan (*support levels*) minimal dapat ditentukan untuk mendapatkan aturan-aturan yang diidentifikasi oleh metode pohon keputusan (atau aturan asosiasi lain).

Agar suatu aturan memiliki daya tarik, maka aturan itu harus menunjukkan sesuatu yang baru dan bermanfaat (memiliki tingkat konfidensi dan dukungan yang cukup tinggi). Sebagai contoh, sebuah toko yang menjual kebutuhan sehari-hari yang menerapkan proses penggalian data yang mendapati bahwa telur dan daging sapi dibeli secara bersamaan pada tingkat konfidensi 0,9 dan tingkat dukungan 0,2 mungkin tidak terkesan. Toko ini sudah mengetahui hal tersebut sebelum penggalian data dilakukan. Daya tarik (*interestingness*) adalah bahwa analisis penggalian data mendapati sesuatu yang tidak terduga (penemuan pengetahuan). Memang berguna untuk memastikan hipotesis bahwa telur dan daging sapi terjual bersama-sama. Tetapi lebih berguna untuk menemukan informasi bahwa selai *blackberry* dan telur juga terjual bersama-sama. Informasi

seperti ini dapat menyebabkan toko tersebut menata kembali tampilan barangnya dan atau program promosinya.

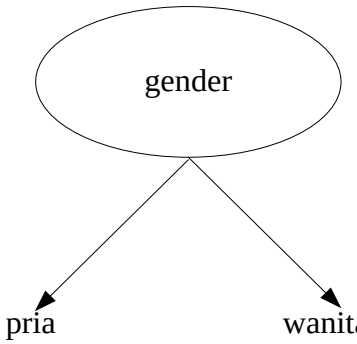
Algoritma-algoritma induksi aturan (*rule-induction*) telah dikembangkan untuk memproses data kategorik (dan juga data kontinu) secara otomatis. Untuk menggunakan pendekatan ini, dibutuhkan hasil akhir yang jelas (Dhar & Stein, 1997 dalam buku Olson & Shi, 2008). Induksi aturan bekerja dengan cara menelusuri data untuk mencari pola dan hubungan. Data dapat dibagi-bagi ke dalam berbagai kategori yang spesifik. Pembelajaran mesin (*machine learning*) dimulai tanpa asumsi sama sekali, hanya memandang data *input* dan hasil akhir. Penilaian yang dikembangkan oleh pakar tidak dipertimbangkan, yang tampaknya tidak efisien, tetapi berarti subjektivitas manusia dapat dihilangkan. Algoritma partisi rekursif memisahkan data (data asli, bukan data yang telah dikelompokkan) ke dalam bagian-bagian yang semakin lama semakin kecil untuk selanjutnya dibuat pohon keputusan.

### 2.3.2. Algoritma C4.5

Algoritma C4.5 diperkenalkan oleh Quinlan(1994) sebagai versi perbaikan dari ID3. Dalam ID3, induksi pohon keputusan hnya bisa dilakukan pada fitur bertipe kategorikal (nominal atau ordinal), sedangkan tipe numerik (interval atau rasio) tidak dapat digunakan. Perbaikan yang membedakan algoritma C4.5 dengan ID3 adalah dapat menangani fitur dengan tipe numerik, melakukan pemotongan (*pruning*) pohon keputusan dan penurunan (*deriving*) set aturan. Algoritma C4.5 juga menggunakkan kriteria gain dalam menentukan fitur yang menjadi pemecah node pada pohon yang diinduksi.

Hal terpenting dalam induksi pohon keputusan adalah bagaimana menyatakan syarat pengujian pada node. Ada tiga kelompok penting dalam syarat pengujian node:

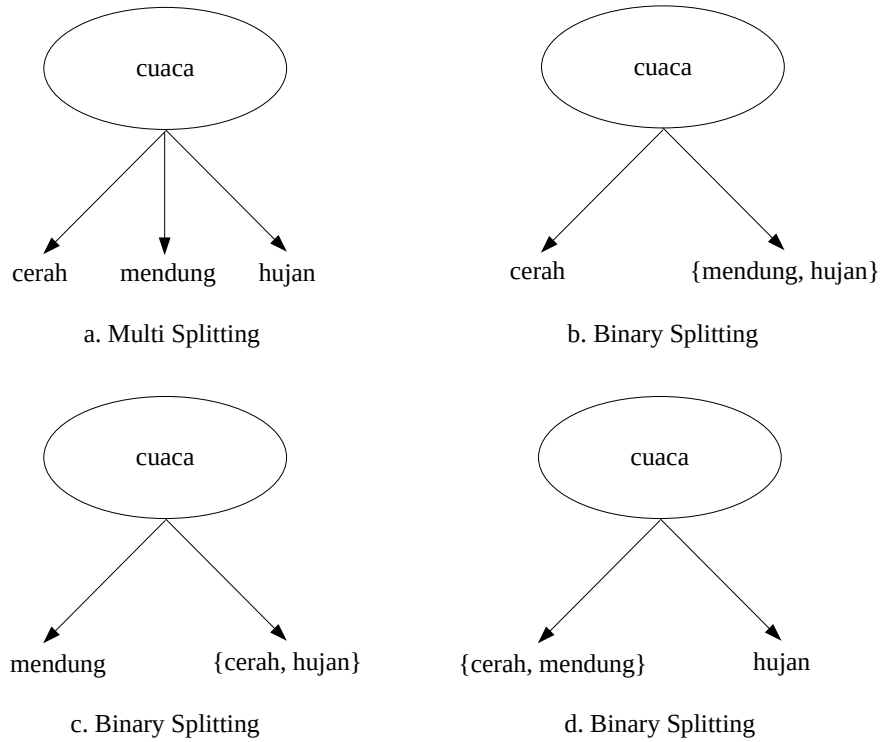
a. Fitur biner: hanya mempunyai dua nilai berbeda disebut dengan fitur biner. Syarat pengujian ketika fitur ini menjadi node (akar maupun internal) hanya mempunyai dua pilihan cabang. Contoh pemecahannya disajikan pada gambar 2.2;



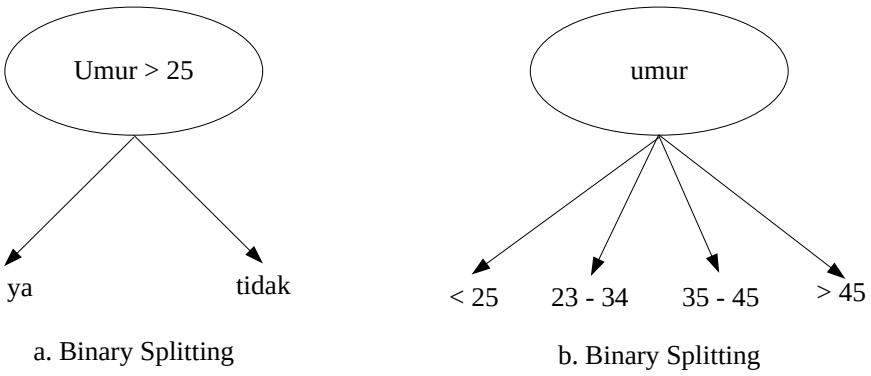

Gambar 2.2. Syarat pengujian fitur biner (Prasetyo, 2014)

b. Fitur bertipe kategorikal: untuk fitur ini (nominal atau ordinal) bisa mempunyai beberapa nilai yang berbeda. Contohnya adalah fitur ‘cuaca’ memiliki tiga nilai berbeda dan ini bisa mempunyai banyak kombinasi syarat pengujian pemecahan. Secara umum terdapat dua, yaitu pemecahan biner (*binary splitting*) dan *multi splitting*. Kombinasinya disajikan pada gambar 2.3. Untuk pemecahan yang hanya membolehkan pemecahan biner, seperti CART, maka akan memberikan kemungkinan jumlah kombinasi pemecahan sebanyak $2^{k-1}$-1, dimana *k* adalah jumlah nilai berbeda dalam fitur tersebut;
c. Fitur bertipe numerik: syarat pengujian dalam node (akar maupun internal) dinyatakan dengan pengujian perbandingan ($A < v$) atau ($A \geq v$) dengan hasil biner, atau untuk multi dengan hasil berupa jangkauan nilai dalam bentuk $v_i \leq A < v_{i+1}$, untuk $i = 1, 2, \ldots k$. Untuk kasus pemecahan biner, maka algoritma akan memeriksa semua kemungkinan posisi pemecahan *v* dan memilih posisi *v* yang terbaik. Untuk cara multi, maka algoritma harus memeriksa kemungkinan jangkauan nilai kontinyu. Contoh pemecahan pada fitur numerik disajikan pada gambar 2.4;

Algoritma C4.5 menggunakan kriteria *entropy* dalam pemilihan cabang pemecah. Kriteria ini didasarkan pada pemilihan titik pemecahan yang memaksimalkan informasi gain (pengurangan *entropy* maksimal). Nilai minimal 0 ketika semua data pada node tersebut dimiliki oleh satu kelas, hal ini mengimplikasikan paling informatif. Sebelum menghitung rasio perolehan, perlu menghitung dulu nilai informasi dalam satuan bits dari suatu kumpulan objek.




Gambar 2.3. Syarat pengujian fitur bertipe kategorikal (Prasetyo, 2014)




Gambar 2.4. Syarat pengujian fitur bertipe numerik (Prasetyo, 2014)

Cara menghitungnya dilakukan dengan menggunakan konsep *entropy*.*Entropy* dapat dihitung menggunakan persamaan berikut:

$$E(s) = -\sum_{i=1}^{m} p(\omega_i|s) log_2\, p(\omega_i|s) \qquad (\mathbf{2.1})$$

$p(\omega_i|s)$ adalah proporsi kelas ke-*i* dalam semua data latih yang diproses di node *s*. $p(\omega_i|s)$ didapatkan dari jumlah semua baris data dengan label kelas *i* dibagi jumlah baris semua data. Sementara *m* adalah jumlah nilai berbeda dalalm data.

*Entropy* digunakan untuk menentukan yang manakah node yang akan menjadi pemecah data latih berikutnya. Nilai *entropy* yang lebih tinggi akan meningkatkan potensi klasifikasi. Yang perlu diperhatikan adalah jika *entropy* untuk node bernilai 0 berarti semua data vektor berada pada label kelas yang sama dan node tersebut menjadi daun yang berisi keputusan (label kelas). Yang juga perlu diperhatikan dalam perhitungan *entropy* adalah jika salah satu dari elemen $\omega_i$ jumlahnya 0 maka *entropy* dipastikan 0 juga. Jika proporsi semua elemen $\omega_i$ sama jumlahnya maka dipastikan *entropy* dipastikan bernilai 1.

*Gain* digunakan untuk memperkirakan pemilihan fitur yang tepat untuk menjadi pemecah pada node tersebut. *Gain* sebuah fitur ke-*j* dihitung menggunakan persamaan berikut:

$$G(s,j) = E(s) - \sum_{i=1}^{n} p(v_i|s) \times E(s_i) \qquad (\mathbf{2.2})$$

$p(v_i|s)$ adalah proporsi nilai *v* muncul pada kelas dalam node. *E(*$s_i$*)* adalah *entropy* komposisi nilai *v* dari kelas ke-*j* dalam data ke-*i* node tersebut. *N* adalah jumlah nilai berbeda dalam node. Kriteria yang paling banyak digunakan untuk memilih fitur sebagai pemecah dalam algoritma C4.5 adalah rasio gain, yang diformulasikan oleh persamaan berikut:

$$Rasio\ Gain(s,j) = \frac{Gain(s,j)}{SplitInfo(s,j)} \qquad (\mathbf{2.3})$$

Persamaan 2.3 menyatakan nilai rasio gain pada fitur ke-*j*. *SplitInfo*(*s*,*j*), didapat dari:

$$SplitInfo(s,j) = -\sum_{i=1}^{k} p(v_i|\ s)\ log_2\ p(v_i\ |s) \qquad (\mathbf{2.4})$$

*k* menyatakan jumlah pemecahan.

## 2.4. Pemilihan Fitur

Prasetyo (2014) mengungkapkan bahwa salah satu fase penting dalam pemrosesan awal dalam penggalian data adalah pemilihan fitur yang nantinya akan diproses dalam metode penggalian data. Dalam aplikasi-aplikasi yang menerapkan penggalian data biasanya menggunakan sejumlah pemodelan fitur yang dikombinasikan dengan harapan memberikan akurasi kinerja yang baik. Jumlah fitur yang banyak pasti berimbas pada komputasi yang mahal dan kompleks. Akan tetapi, jumlah fitur yang banyak ternyata tidak selalu menjamin kinerja yang baik. Penggunaan dua fitur yang mempunyai kondisi diskriminan dua kelas yang dengan baik tentu akan memberikan kinerja sistem lebih baik daripada penggunaan banyak fitur tetapi tidak memberikan diskriminasi dua kelas atau lebih dengan baik. Dengan kata lain, fitur yang dipilih haruslah fitur yang mempunyai korelasi dalam mendiskriminasi kelas-kelas yang diproses. Hal ini sangat penting yang harus diperhatikan dalam klasifikasi adalah sifat generalisasi yang dibangun oleh klasifikator, di mana semakin tinggi rasio jumlah data latih terhadap jumlah parameter bebas maka akan semakin baik sifat generalisasi klasifikator yang dihasilkan.

Pekerjaan utama dalam pemilihan fitur adalah jika diberikan sejumlah fitur sebagai kandidat fitur yang digunakan, maka bagaimana cara memilih fitur yang paling penting di antara kandidat tersebut sehingga dapat mengurangi jumlahnya, dan pada saat yang sama memungkinkan memberikan diskriminasi kelas dengan baik. Fase ini tentu sangat kritis. Jika fitur yang dipilih yang dipilih memiliki kekuatan diskriminasi yang kecil, akibatnya desain klasifikator yang dibentuk mempunyai kinerja yang buruk. Sebaliknya, jika fitur yang kaya informasi diskriminasi saja yang dipilih, maka desain klasifikator yang dibentuk menjadi sangat sederhana. Dengan kata lain, yang harus diusahakan dalam pemilihan fitur adalah mengarah pada jarak perbedaan antar kelas yang besar dan variasi dalam kelas kecil. Ini berarti bahwa fitur harus mempunyai nilai perbedaan yang jauh dalam kelas berbeda dan nilainya dekat dalam kelas yang sama. Pengujian fitur biasanya tidak hanya dilakukan pada fitur dalam dimensi saat itu, adakalanya kasus data yang non-linear akan sulit menemukan diskriminasinya sehingga diperlukan transformasi fitur dari dimensi yang lama ke dimensi yang baru

dengan harapan dalam dimensi yang baru (relatif lebih tinggi) mampu mentransformasi yang asalnya non-linear menjadi linear.

Fitur yang menjadi elemen setiap vektor mempunyai jenis beragam, misalnya pada data mengenai fisik manusia ada fitur tinggi badan yang menggunakan nilai yang sifatnya kuantitatif. Fitur ini mempunyai tipe numerik yang dapat diperbandingkan satu dengan lain. Sementara fitur warna kulit menggunakan nilai yang sifatnya kualitatif sehingga tidak bisa dilakukan perbandingan. Umumnya tipe fitur ada dua, yaitu kategorikal (kualitatif) dan numerik (kuantitatif). Terdapat empat sifat penting yang dimiliki fitur secara umum:

a. *Distinctness*, meliputi sama dengan (=) dan tidak sama dengan (≠);
b. *Order*, meliputi lebih kecil (<), lebih kecil atau sama dengan (≤), lebih besar (>) dan lebih besar atau sama dengan (≥);
c. *Addition*, meliputi penjumlahan (+) dan pengurangan (-);
d. *Multiplication*, meliputi perkalian (*) dan pembagian (/).

Dari keempat sifat tersebut, maka dapat diturunkan empat tipe fitur, yaitu nominal, ordinal, interval dan rasio. Tabel 2.3 menyajikan penjelasan keempat fitur dan kaitannya dengan sifat – sifat di atas.

Tabel 2.3 Tipe Fitur

| **Tipe Fitur** | | **Penjelasan** | **Contoh** |
|---|---|---|---|
| Kategorikal (Kualitatif) | Nominal | Nilai fitur bertipe nominal memberikan nilai berupa nama, dengan nama inilah sebuah fitur membedakan dirinya pada vektor yang satu dengan yang lain (=, ≠) | Kode pos, nomor KTP, nomor induk mahasiswa, jenis kelamin |
| | Ordinal | Nilai fitur bertipe ordinal mempunyai nilai berupa nama yang mempunyai arti informasi yang terurut (<, ≤, >, ≥) | Predikat kelulusan (*cumlaude*, sangat memuaskan, memuaskan), suhu (dingin, normal, panas) |

| Tipe Fitur | | Penjelasan | Contoh |
|---|---|---|---|
| Numerik (Kuantitatif) | Interval | Nilai fitur di mana perbedaan di antara dua nilai mempunyai makna yang berarti (+, -) | Tanggal, suhu (dalam celcius atau fahrenheit) |
| | Rasio | Nilai fitur di mana perbedaan di antara dua nilai dan rasio dua nilai mempunyai makna yang berarti (*, /) | Umur, panjang, tinggi, rata-rata |

Sumber : Prasetyo, 2014

Berdasarkan angka nilai, fitur juga dapat dibedakan menjadi dua, yaitu diskret dan kontinyu. Sebuah fitur dapat bernilai diskret jika mempunyai nilai dalam himpunan jumlah yang terbatas. Jenis ini bisa ditemui pada fitur kategorikal yang hanya mempunyai beberapa variasi nilai, misalnya suhu. Suhu hanya mempunyai tiga kemungkinan nilai (dingin, normal, panas). Contoh lain yang paling sederhana adalah jenis kelamin, hanya pria dan wanita. Nilai ini kadang direpresentasikan dengan nilai biner seperti ya/tidak, benar/salah, pria/wanita atau 0/1. Sementara fitur yang bernilai kontinyu akan mempunyai jangkauan nilai real. Variabel panjang atau tinggi biasanya nilainya menggunakan representasi *floating point* (*real*). Akan tetapi meskipun menggunakan representasi real, tetap digunakan ukuran presisi jumlah angka dibelakang koma.

### 2.4.1. Transformasi Fitur

Transformasi fitur (FT) adalah cara lain untuk menangani pemilihan fitur yang heterogen. Metode transformasi menyatukan format dataset dan memungkinan algoritma pemilihanfitur konvensional mampu menangani dataset yang sifatnya heterogen. Proses ini dilakukan untuk mengubah fitur non-numerik menjadi fitur numerik yang tergantung pada fitur non-numerik asli itu sendiri. Transformasi fitur yang diusulkan berupa UFT (*unsupervised feature transformation*) merupakan metode untuk menemukan numerik $\tilde{X}$ yang menggantikan fitur non-numerik $X$, dan $\tilde{X}$ memenuhi kondisi mutual informasi (MI) ($\tilde{X}$; $X$) = H($X$). Kondisi ini menjadikan MI antara $\tilde{X}$ yang ditransformasi dan

*Xoriginal* menjadi sama dengan *entropy* dari *Xoriginal.* Kondisi ini sangat penting karena tidak mengubah informasi asli, ketika fitur non-numerik diubah menjadi fitur numerik (Wei, Chow, & Chan, 2015a). Transformasi tidak bergantung pada label kelas. Hal ini penting, karena bias yang muncul pada label kelas dapat dikurangi.

Diasumsikan bahwa $\tilde{X}$ adalah numerik pengganti untuk fitur non-numerik $X = \{x_i | \, i = 1, \dots, n\}$**,** maka

$$p(\tilde{x}, x = x_i) = p_{\tilde{x}|x=x_i}(\tilde{x}) p_i \qquad (\mathbf{2.8})$$

kemudian $p_i = p_x(x = x_i)$, maka

$$MI(\tilde{x}; x) = \sum_{i=1}^{n} p_i \left[ \int_{\tilde{x}|x=x_i} p_{\tilde{x}|x=x_i}(\tilde{x}) log\left(p_{\tilde{x}|x=x_i}(\tilde{x})\right) d\tilde{x} - \int_{\tilde{x}|x=x_i} p_{\tilde{x}|x=x_i}(\tilde{x}) \log\left(p_{\tilde{x}}(\tilde{x})\right) d\tilde{x} \right] \qquad (\mathbf{2.9})$$

untuk setiap subkelompok fitur non-numerik yang asli, digunakan kelompok nilai numerik yang mematuhi aturan *Gaussian* untuk menggantikannya. Selain itu, distribusi total penggantian nilai numerik juga diasumsikan dengan distribusi *Gaussian*. Kemudian, $x^* = (x - \mu)/\sigma$ digunakan untuk menormalkan data yang berubah. Dengan demikian, parameter μ dan σ (Mo & Huang, 2012; Sebban & Nock, 2002) dapat dihilangkan (Wei, Chow, & Cha, 2015b) dan hasil akhirnya adalah:

$$\mu^* = \left[ (n - i) - \sum_{k=1}^{i} (n - k) p_k \right] \sqrt{\left(1 - \sum_i p_i^3\right) \div \sum_{i \neq j} p_i p_j (i - j)^2} \quad i \in \{1, \dots, n\} \qquad (\mathbf{2.10})$$

$$\sigma^* = p_i \qquad i \in \{1, \dots, n\} \qquad (\mathbf{2.11})$$

berdasarkan solusi di atas, UFT dapat diformulasikan sebagai berikut:

**Algortima** : UFT

***Input*** :dataset D, yang memiliki fitur beragam $f_j, j \in \{1, \dots m\}$

***Output*** : mengubah dataset D’ dengan fitur numerik asli

***for****j = 1* ***to*** *m* ***do***

***if*** fitur $f_j$ adalah non-numerik maka

*n = size (unique($f_j$ ));*

$\{s_i | i = 1, \dots, n\}$adalah set nilai non-numerik dalam fitur $f_j$

$p_i$ adalah kemungkinan $s_i$

***for*** *i=1* ***to*** *n* ***do***

$$\mu^* = \left[(n - i\,) - \sum_{k=1}^{i}(n-k)p_k\right]\sqrt{\left(1 - \sum_i p_i^3\right) \div \sum_{i \neq j} p_i p_j (i-j)^2}$$

$\sigma_i = p_i$

gunakan distribusi *Gaussian* $\mathcal{N}(\mu_i, \sigma_i)$ untuk menghasilkan data numerik dan menggantikan nilai-nilai yang sama untuk $s_i$ di fitur $f_j$

***end for***

***end if***

***end for***

Dengan menggunakan UFT yang dijelaskan di atas, setiap nilai fitur non-numerik dapat digantikan oleh sekelompok nilai-nilai numerik. Pada saat yang sama, substitusi numerik dapat mempertahankan informasi asli yang terkandung dalam fitur non-numerik.

### 2.4.2. Pemilihan Fitur BerdasarkanMutual Informasi (MI)

Kondisi karakteristik optimal seringkali diartikan dengan tingkat minimal kesalahan klasifikasi (Peng, Long, & Ding, 2005). Tingkat minimal kesalahan memerlukan maksimal keterkaitan dari target kelas pada distribusi data di ruang bagian $R^m$ (dan sebaliknya). Dalam permasalahan pemilihan fitur, fitur yang relevan memiliki informasi penting berkaitan dengan kelas, sebaliknya fitur yang tidak relevan berisi sedikit informasi yang berkaitan dengan kelas(Li, Xie, & Goh, 2009; Kwak & Choi, 2002). Tujuan dari pemilihan fitur adalah mencari fitur yang berisi sebanyak mungkin informasiyang berkaitan dengan kelas. Berdasarkan tujuan tersebut, teori informasi Shannon menemukan cara yang mudah untuk mengukur informasi dari variabel acak dengan menggunakan entropi dan mutual

informasi. Dalam teori probablitas dan teori informasi, mutual informasi adalah jenis metode untuk analisis korelasi, yang digunakan untuk mengukur ketergantungan antara variabel acak (Han & Ren, 2015).

Nilai entropi *H(X)* merupakan ukuran ketidakpastian dari variabel acak *X*. Untuk variabel acak diskret *X*, dengan *p(x)*, nilai entropi dari *X* adalah:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \ (\mathbf{2.5})$$

Basislog yang digunakan adalah basis 2 dan entropi dalam satuan bits. Nilai join entropi dari *X* dan *Y* dengan join pdf *p(x,y)* adalah:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \epsilon Y} p(x,y) \log p(x,y) \ (\mathbf{2.6})$$

Ketika variabel-variabel tertentu dikenal dan yang lain tidak dikenal, ketidakpastian diukur dengan entropi bersyarat:

$$H(X|Y) = -\sum_{x \in X} p(x) H(Y|X = x)$$

$$= -\sum_{x \in X} \sum_{y \epsilon Y} p(x,y) \log p(x|y) \ (\mathbf{2.7})$$

Oleh karena itu, join entropi dan entropi bersyarat dapat dibuat hubungan sebagai berikut:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) (\mathbf{2.8})$$

Informasi yang ditemukan bersama oleh dua variabel acak penting dan didefinisikan sebagai mutual informasi antara dua variabel:

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \qquad (\mathbf{2.9})$$

Jika mutual informasi antara variabel dengan label kelas bernilai besar, berarti kedua variabel sangat berelasi. Jika mutual informasi bernilai nol, berarti kedua variabel bersifat independen. Mutual informasi dan entropi dapat dihubungkan sebagai berikut:

$$I(X;Y) = H(X) - H(X|Y),$$

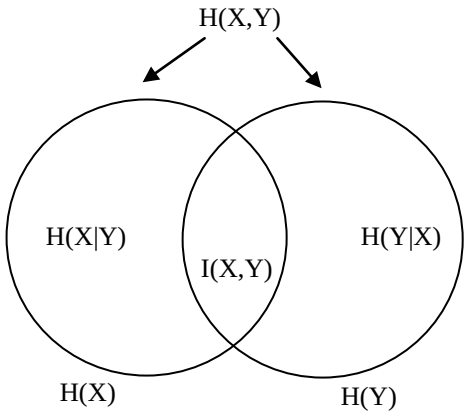
$$I(X;Y) = H(Y) - H(Y|X),$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \qquad (\mathbf{2.10})$$

$$I(X;Y) = I(Y;X),$$

$$I(X;X) = H(X)$$

Hubungan antara mutual informasi dan entropi ditunjukkan pada gambar 2.5. Mutual informasi sesuai dengan persimpangan atau singgungan informasi X dengan informasi Y.




Gambar 2.5 Hubungan antara mutual informasi dan entropi

Secara formal, mutual informasi membandingkan peluang pengamatan *x* dan *y* secara bersama dengan peluang pengamatan *x* dan *y* secara bebas. Jika terdapat hubungan yang kuat antara pengamatan *x* dan *y*, maka *p*(*x*,*y*) akan lebih besar dari *p*(*x*).*p*(y) dan sebagai akibatnya *I(x,y)*>> 0. Apabila tidak ada hubungan keterikatan antara pengamatan *x* dan *y* maka *p(x,y)* ≈ *p(x).p(y)* dan *I(x,y)*≈ 0.

### 2.4.3. *Dynamic* Mutual Informasi

Metode pengukuran informasi telah banyak digunakan sebagai ukuran seleksi fitur algoritma, dan kinerjanya telah dibuktikan oleh banyak penelitian. Walaupun informasi yang disajikan metrik memiliki representasi yang berbeda, tetapi menurut teori entropi atau teori mutual informasi, mereka dibangun atas dasar teori probabilitas. Dengan demikian, derajat metrik karakteristik korelasi, apakah itu metrik informasi atau metrik probabilitas, yang ditandai dalam sampel perlu dihitung terlebih dahulu distribusi probabilitas dari suatu dataset. Secara detail algoritma *dynamic* mutual informasi adalah sebagai berikut:

***Input*** **:** dataset latih *T= D(F, C)*

***Output*****:** fitur terpilih *S*

(1) Inisialisasi parameter : F = *F; S=Ø;* $D_u$ *= D;* $D_l$ *= Ø;*

(2) ***Repeat***

(3) **For** tiap fitur *f*∈*F***do**

(4) Hitung nilai mutual informasi *I(C;f)* pada $D_{u;}$

(5) **If** *I(C;f)* = 0 **then** *F = F – {f};*

(6) Pilih fitur *f* yang memiliki nilai tertinggi *I(C;f);*

(7) *S = S U{f}; F = F\{f};*

(8) Memperoleh*instances* label baru $D_l$yang diperoleh dari D*u*;

(9) Hapus dari $D_{u,}$ yaitu$D_u = D_u \backslash D_l$;

(10) **Until** *F = Ø* atau $|D_u| = I_T$

Dalam metode mutual informasi tradisional, nilai mutual informasi diperkirakan pada seluruh sampling, hal ini tidak dengan tepat mewakili relevansi dari fitur-fitur. Sedangkan pada *dynamic* mutual informasi mampu mendapatkan nilai mutual informasi yang tepat dari kandidat fitur.Metode ini juga mampu mengurangi redundansi dan data yang tidak relevan serta mampu memberikan mutual informasi antara subset fitur dengan label kelas secara langsung (Liu, Sun, Liu, & Zhang, 2009).Hal ini dikarenakan dalam proses perhitungan nilai mutual informasi pada *dynamic* mutual informasi dilakukan estimasi nilai mutual informasi dari tiap fitur ($f_i$) dalam *F* dengan label kelas *C*dan hasilnya akan diurutkan secara *descending* dan kandidat fitur yang memiliki nilai mutual informasi tertinggi akan dipilih dan ditambahkan dalam subset *S*. Tujuan dari tahapan ini adalah untuk mencegah perhitungan ulang dalam memperkirakan nilai mutual informasi dari fitur di tahap berikutnya. Setelah itu, akan dilanjutkan ke kandidat fitur berikutnya. Prosedur ini diulang sampai tidak ada lagi kandidat fitur di *F*. Strategi pencarian yang digunakan adalah *sequential forward search*. Pada setiap iterasi akan dilakukan perhitungan nilai mutual informasi $I(f_i, C)$ untuk semua fitur dalam kandidat *F*.

Sebelum mengidentifikasi fitur pada tiap tahapan, pertama kali algoritma memperkirakan nilai mutual informasi dari kandidat fitur *f* dengan label kelas *C*.

Prosedur rekursi ini mirip dengan langkah pertama dari algoritma mutual informasi lain, yang mengambil fitur dengan nilai mutual informasi tertinggi yang diperkirakan pada seluruh sampel ruang *D*.Selama tahap perhitungan, fitur ($f_i$) yang memiliki nilai mutual informasi nol (0) maka akan dihapus dari *F*(Liu, Sun, Liu, & Zhang, 2009). Dalam kondisi seperti ini, distribusi probabilitas dari fitur diacak secara penuh dan itu tidak akan memberikan kontribusi untuk prediksi kasus dalam $D_u$. Setelah itu fitur yang memiliki nilai mutual informasi tertinggi akan dipilih.

## 2.5. Pengukuran Kinerja dan Evaluasi

Setelah melalui tahapan praproses, selanjutnya akan dilakukan beberapa tahapan pada klasifikasi, yaitu pengukuran kinerja dan evaluasi. Hasil akhir yang menjadi keluaran dari model sistem harus ditentukan dengan suatu ukuran. Sehingga dapat dibandingkan model manakah yang paling baik.

### 2.5.1. *Cross Validation*

Validasi silang (*cross validation*) merupakan suatu metode statistik yang digunakan untuk menganalisa dan mengukur keakuratan hasil percobaan pada data independen. Metode ini membagi sebuah data menjadi beberapa subdata satu digunakan untuk mengkonfirmasi kebenaran subdata yang lain.

*K-fold cross validation* adalah salah satu metode *cross validation* yang membagi data menjadi *k* subdata. Salah satu subbagian data dijadikan sebagai *validator* dan *testing*, sedangkan *k-1* data digunakan sebagai pelatihan. Proses di atas dilakukan berulang sebanyak *k* kali untuk setiap subbagian data. hasil dari pengujian adalah rata-rata dari *k* kali pengujian pada data tersebut. Teknik validasi silang yang memiliki akurasi paling baik menggunakan *10 fold cross validation* (Witten, Frank, & Hall, 2011).

### 2.5.2. Akurasi

Sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua set data dengan benar. Akan tetapi, tidak dapat dipungkiri bahwa kinerja suatu sistem tidak bisa bekerja 100% benar (Prasetyo, 2014). Oleh

karena itu, sebuah sistem klasifikasi juga harus diukur kinerjanya. Umumnya cara mengukur kinerja klasifikasi menggunakan matriks *confusion*, seperti yang dapat dilihat pada tabel 2.4 berikut ini.

Tabel 2.4 Matriks Confusion

| | ***Actual Positive*** | ***Actual Negative*** |
|---|---|---|
| ***Predict Positive*** | a(Tp) | b(Fn) |
| ***Predict Negative*** | c(Fp) | d(Tn) |

Sumber : Arifiyanti, 2015

Keterangan:

Tp (*True Positive*) : Jumlah klasifikasi yang benar dari data positif

Fp (*False Positive*) : Jumlah klasifikasi yang salah dari data negatif

Fn (*False Negative*) : Jumlah klasifikasi yang salah dari data positif

Tn (*True Negative*) : Jumlah klasifikasi yang benar dari data negatif

Berdasarkan matriks *confusion*, maka dapat diketahui jumlah data dari masing-masing kelas yang diprediksi secara benar yaitu (TP dan TN) dan data yang diklasifikasi secara salah yaitu (FP dan FN). Dengan mengetahui jumlah data yang diklasifikasikan secara benar maka dapat diketahui akurasi hasil prediksi, dan dengan mengetahui jumlah data yang diklasifikasikan secara salah maka dapat diketahui laju *error*dari prediksi yang dilakukan. Untuk menghitung akurasi digunakan formula sebagai berikut:

$$Akurasi = \frac{Tp + Tn}{Tp + Fp + Fn + Tn} \qquad (\mathbf{2.13})$$

### 2.5.3. *Precision* dan *Recall*

Dalam bidang pencarian informasi, *precision* (disebut juga *positive prediction value*) merupakan metrik untuk mengukur kinerja sistem dalam mendapatkan data yang relevan. Sementara *recall* (disebut juga sensitivitas) merupakan metrik untuk mengukur kinerja sistem dalam mendapatkan data relevan yang terbaca (dalam bidang pencarian informasi). Dalam bidang penggalian data, *precision* adalah jumlah data yang *true positive* (jumlah data positif yang dikenali secara benar sebagai positif) dibagi dengan jumlah data yang

dikenali sebagai positif, sedangkan *recall* adalah jumlah data yang *true positive* dibagi dengan jumlah data yang sebenarnya positif (*true positive + true negative*).

Mengacu pada tabel 2.4, berikut persamaan yang digunakan untuk menghitung *precision*:

$$Precison = \frac{Tp}{Tp + Fp} \qquad (\mathbf{2.14})$$

Persamaan yang digunakan untuk menghitung *recall* adalah:

$$Recall = \frac{Tp}{Tp + Fn} \qquad (\mathbf{2.15})$$

### 2.5.4. *F-Measure*

*F-measure* merupakan parameter tunggal ukuran keberhasilan *retrieval* yang menggabungkan *precision* dan *recall* (Rijsbergen, 1979 dalamArifiyanti, 2015).

$$F - Measure = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \qquad (\mathbf{2.16})$$

Keterangan:

$\beta$ = parameter kepentingan relatif aspek *precision* dan *recall*

$P$ = nilai *precision*

$R$ = nilai *recall*

Jika nilai $\beta > 1$, maka akan memberikan bobot kepentingan *recall* lebih tinggi daripada *precision*. Jika nilai $\beta = 2$ maka bobot *recall* dua kali lebih besar daripada *precision.* Jika nilai $\beta = 0{,}5$ maka bobot *precision* dua kali lebih besar daripada *recall*. Tetapi jika *recall* dan *precision* memiliki bobot yang sama, maka $\beta = 1$ dan parameter *F-measure* dituliskan pada persamaan 2.14.

$$F - Measure = \frac{2PR}{P + R} \qquad (\mathbf{2.17})$$

Suatu sistem klasifikasi dinyatakan efektif jika hasil perhitungan menunjukkan ketepatan (*precision*) yang tinggi sekalipun *recall*-nya rendah.

## 2.6. Manajemen Hubungan Pelanggan

Manajemen hubungan pelanggan atau yang dikenal dengan istiah CRM merupakan suatu pengelolaanyang melibatkan semua aspek hubungan pelanggan dengan organisasi untuk meningkatkan loyalitas dan retensi pelanggan serta keuntungan organisasi(Baltzan & Phillips, 2009). Sebagai organisasi yang mulai bermigrasi dari organisasi yang berfokus pada produk tradisional terhadap organisasi berorientasi pada pelanggan, mereka mengakui bahwa pelanggan mereka adalah ahli, bukan hanya sebagai pembangkit pendapatan. Organisasi yang cepat menyadari bahwa tanpa pelanggan, mereka tidak akan ada dan itu sangat penting, mereka melakukan segala sesuatu yang mereka bisa untuk memastikan kepuasan pelanggan mereka. Di zaman ketika diferensiasi produk sulit, CRM adalah salah satu aset paling berharga bagi perusahaan. Semakin cepat sebuah perusahaan mencakup CRM yang lebih baik, akan sulit bagi pesaing untuk mencuri pelanggan yang setia.Fokus CRM harus selalu berdasarkan pengalaman pelanggan, karena harapan pelanggan sering berubah(Motiwalla & Thompson, 2012). Agrawal (2003)memaparkan bahwa kebanyakan definisi CRM selalu berkaitan dengan*relationship* (hubungan antara pelanggan dan perusahaan) dan teknologi informasi, sehingga CRM dapat didefinisikan sebagaipenggunaan teknologiinformasi pada proses bisnis yang bertujuanuntukmembangun hubungan jangkapanjang dan saling menguntungkan dengan pelanggandalam rangka usaha retensi pelanggan, peningkatan nilai dan keuntungan.Perusahaan dapat memahami perilaku dan kebiasaan dalam berbelanja dan menggunakan produk atau layanan, sehingga harapan dari pelanggan mengenai produk dapat diusahakan dan dipenuhi oleh perusahaan.

*Customer Relationship Management* (CRM) telah menjadi strategi terkemuka di lingkungan bisnis yang sangat kompetitif, CRM dapat dilihat sebagai upaya manajerial untuk mengelola interaksi bisnis dengan pelanggan yang menggabungkan proses bisnis dan teknologi yang berusaha untuk memahami pelanggan perusahaan (Kim, Suh, & Hwang, 2003). Perusahaan menjadi semakin sadar akan banyak manfaat yang diberikan oleh CRM. Beberapa manfaat CRM menurut (Jutla, Craig, & Bodorik, 2001; Stone, Woodcock, & Wilson, 1996)adalah sebagai berikut:

a. Peningkatan retensi dan loyalitas pelanggan;
b. Profitabilitas pelanggan yang lebih tinggi;
c. Nilai penciptaan untuk pelanggan;
d. Kustomisasi produk dan layanan;
e. Proses yang lebih singkat, produk-produk dan layanan berkualitas tinggi.

CRM yang efektif adalah tentang memperoleh, menganalisis dan berbagi pengetahuan tentang pelanggan perusahaan (Jutla, Craig, & Bodorik, 2001).Sebagai kompetisi dan meningkatnya biaya untuk menarik pelanggan baru, perusahaan semakin memfokuskan upaya strategis pada retensi pelanggan(Jones, Mothersbaugh, & Beatty, 2000).Pendapatandapat ditingkatkan dengan menerapkan program retensi. Lamanya hubungan pelanggan mempengaruhi profitabilitas perusahaan(Reichheld & Sasser, 1990). Pelanggan yang bertahan lama akan menghasilkan pendapatan dan marjin yang lebih tinggi daripada pelanggan baru. Membuat pelanggan baru tertarik atas penjualan membutuhkan tindakan yang sangat mahal (Athanassopoulos, 2000). Upaya biaya periklanan serta promosi dan penjualan adalah biaya yang signifikan tetapi diperlukan untuk menarik pelanggan (Zeithaml & Berry, 1996)dan membangun hubungan baru (Athanassopoulos, 2000). Selain itu, pelanggan baru sering tidak menguntungkan untuk beberapa waktu.

Dalam semua kasus, bagaimanapun perilaku beralih didefinisikan sebagai pembelotan secara total (Buckinx & Poel, 2005). Dalam dunia industri, sangat mudah untuk mengamati terjadinya pembelotan: pelanggan benar-benar merusak hubungan mereka dengan perusahaan. Pembeli biasanya tidak menunjukkan tanda-tanda pembelotan, semuanya terjadi tiba-tiba. Mereka beralih ke toko lain untuk melakukan transaksi pembelian, yaitu mereka menunjukkan pembelotan parsial. Ada bahaya nyata bahwa setelah beberapa saat, mereka akan beralih sepenuhnya ke pesaing. Jadi dalam pembelotan parsial jangka panjang dapat menyebabkan pembelotan secara keseluruhan.Meminimalkan jumlah pelanggan yang melakukan pembelotan dapat meningkatan penjualan dan secara signifikan dapat menghemat biaya pemasaran(Kim, Song, & Kim, 2005). Hal ini karena pelanggan setia cenderung menghasilkan arus kas yang lebih besar dan meningkatkan keuntungan; dan mereka kurang sensitif terhadap kenaikan harga,

mereka menghasilkan pesan kata dari mulut yang positif dan mereka tidak memiliki biaya perolehan (Reichheld, 1996; Ng & Liu, 2000).

Loyalitas pelanggan tergantung pada kemampuan perusahaan dalam mengelola hubungan pelanggan dengan baik (Duygu & KIRMACI, 2012). Dengan membuat nilai bagi pelanggan, perusahaan akan mendapatkan loyalitas pelanggan. Sebuah kuote statistik menyatakan bahwa dibutuhkan sepuluh kali uang dan usaha untuk menarik pelanggan baru daripada yang dilakukannya untuk mempertahankan yang sudah ada(www.innovationpei.com). Beberapa alasan munculnya manajemen hubungan pelanggan adalah:

a. Biaya pemasaran melalui media masa yang semakin mahal;
b. Pangsa pelanggan memperoleh keuntungan, bukan pangsa pasar;
c. Konsep kepuasan pelanggan dan loyalitas pelanggan menjadi lebih penting;
d. Nilai hubungan pelanggan memperoleh keuntungan;
e. Pemasaran *one-to-one* memperoleh keuntungan;
f. Kompetisi intensif dan perkembangan teknologi komunikasi.

Ngai, Xiu, & Chau(2009)menjelaskan bahwa *framework* CRM dapat dibagi ke dalam dua kelompok, yaitu operasional dan analitikal.Operasional berhubungan dengan proses bisnis, sedangkan analitikal berkaitan dengan analisis karakteristikdan perilaku pelanggan. Penggalian data digunakanuntuk menemukan pola yang tersembunyi dari data pelanggan kemudiandilakukan analisis yang sesuai dengan tujuan dan kebutuhan.Berdasarkan penelitian yang dilakukan pada 900 jurnal internasional,yang terdiri dari jurnal yang membahas CRM dan jurnalyang membahas penggalian data, serta artikel, konferensi, tesis, desertasi, buku teks dan *paper*, maka disimpulkan bahwa CRM mencakup banyak aspek.Hung, Yen, & Wang (2006)mengelompokkan aplikasi CRMberdasarkanteknik penggalian data yang digunakan, dan mengelompokkan teknik penggalian data berdasarkan model penggalian data,yang meliputi *association* (asosiasi),*estimation* (estimasi),*classification* (klasifikasi), *prediction* (prediksi), *segmentation* (segmentasi). Teknik-teknik penggalian data yang dapat digunakan adalah *clustering, decision tree,neural network,genetic algorithm, associations rule*.

### 2.6.1. Loyalitas Pelanggan

Loyalitas pelanggan merupakan dorongan perilaku untuk melakukan pembelian secara berulang-ulang dan untuk membangun kesetiaan pelanggan terhadap suatu produk/jasa yang dihasilkan oleh badan usaha tersebut membutuhkan waktu yang lama melalui suatu proses pembelian yang berulang-ulang tersebut(Olson 1993, dalam Musanto, 2004). Loyalitas pelanggan didefinisikan juga sebagai sikap yang menampilkan hubungan antara pelanggan dan bisnis atau perusahaan (Udo, Bagchi, & Kirs, 2010). Berdasarkan perspektif perilaku, loyalitas pelanggan digambarkan sebagai pembelian berulang, jumlah waktu pelanggan membeli produk yang sama atau jasa dari penjual atau penyedia yang sama. Perilaku berulang atau frekuensi pembelian dan berbicara tinggi mengenai perusahaan kepada orang lain diambil sebagai ukuran loyalitas kepada perusahaan (Rabinovich & Bailey, 2004). Mengetahui loyalitas pelanggan sangat bermanfaat bagi perusahaan dan pelanggan (Behjati, Nahich, & Othaman, 2012).

Pelanggan (*Customer*) berbeda dengan konsumen (*Consumer*), seorang dapat dikatakan sebagai pelanggan apabila orang tersebut mulai membiasakan diri untuk membeli produk atau jasa yang ditawarkan oleh badan usaha(Musanto, 2004). Kebiasaan tersebut dapat dibangun melalui pembelian berulang-ulang dalam jangka waktu tertentu, apabila dalam jangka waktu tertentu tidak melakukan pembelian ulang maka orang tersebut tidak dapat dikatakan sebagai pelanggan tetapi sebagai seorang pembeli atau konsumen.Pelanggan cenderung menjadi loyal jika mereka mengembangkan hubungan pribadi dengan penjual. Pelanggan yang secara teratur membeli dari orang yang sama datang bergantung pada bantuan orang itu dalam membuat keputusan pembelian berikutnya(Griffin, 2002). Zeithaml & Berry (1996) berpendapat bahwa tujuan akhir keberhasilan perusahaan menjalin hubungan relasi dengan pelanggan adalah membentuk loyalitas yang kuat. Indikator dari pengukuran loyalitas yang kuat adalah mengatakan hal-hal yang positif tentang perusahaan, merekomendasikan perusahaan untuk seseorang yang memerlukan saran, mendorong teman dan keluarga untuk melakukan bisnis dengan perusahaan, mengingat bahwa perusahaan adalah pilihan pertama, melakukan lebih banyak bisnis dengan perusahaan dalam beberapa tahun ke depan.

Griffin (2002) juga mengatakan bahwa konsep loyalitas pelanggan melibatkan perilaku pelanggan lebih dari sikap. Ketika pelanggan setia, mereka menunjukkan perilaku pembelian secara non-random dari waktu ke waktu dan mereka memiliki bias tertentu tentang apa yang mereka beli dan dari siapa mereka membelinya. Loyalitas berkonotasi kondisi beberapa durasi dan mensyaratkan bahwa tindakan pembelian terjadi tidak kurang dari dua kali.Konsumen yang loyal berarti konsumen yang melakukan pembelian secara berulang-ulang terhadap merek produk tertentu dan tidak mudah terpengaruhi oleh karakteristik produk, harga dan kenyamanan penggunanya atau berbagai atribut lain yang ditawarkan oleh produk merek alternatif (Nugraha, 2014).Salah satu indikator terbaik untuk mengukur loyalitas adalah kesediaan untuk menyampaikan suatu produk dari perusahaan kepada teman, keluarga dan kolega, karena pengorbanan konsumen dalam membuat rekomendasi ini tanpa ada yang membayarnya (Winarso, 2010). Karakteristik pelanggan yang loyal dari produk *fast moving consumer goods* antara lain:

a. Melakukan pembelian secara teratur,
b. Tidak membeli di luar lini produk,
c. Menolak produk lain, dan
d. Menunjukkan kekebalan daya tarik pesaing (tidak mudah terpengaruh oleh daya tarik produk sejenis dari pesaing).

Menurut Nugraha (2014), parameter yang digunakan dalam pengukuran loyalitas pelanggan terhadap salah satu merek mie instan adalah:

a. Mie instan merek “X” merupakan pilihan utama ketika ingin mengonsumsi mie instan,
b. Menyarankan kepada orang lain atau pelanggan lain untuk membeli mie instan merek “X”,
c. Sering membeli mie instan merek “X”,
d. Tidak membeli mie instan merek lain, meskipun harganya lebih rendah.

Sedangkan dalam penelitian ini parameter yang digunakan dalam pengukuran loyalitas pelanggan berdasarkan penjelasan di atas adalah melakukan

pembelian secara berulang-ulang, merekomendasikan kepada teman, kolega dan keluarga, memberikan komentar yang positif terhadap merk mie instan yang biasa dikonsumsi dan meskipun pernah membeli merk lain namun akan kembali pada merk sebelumnya.

Loyalitas pelanggan memiliki dampak yang luar biasa pada keuntungan bisnis(Oliver, 1999) dan untuk kelangsungan hidup bisnis dan pengembangan, serta membuka jalan bagi suatu organisasi untuk mencapai keunggulan kompetitif yang berkelanjutan (Gronroos, 2009) dan(Gummesson, 2008). Menjaga pelanggan dalam relasi jangka panjang merupakan tantangan bagi praktisi bisnis dan tetap belum diteliti oleh para ahli (misalnya, Gronroos, 2009). Persepsi konsumen dan evaluasi nilai yang diterima dalam hubungan yang berkelanjutan dapat memiliki dampak yang pasti tentang niat mereka untuk terus tinggal, atau meninggalkan, sebuah organisasi (Ravald & Gronroos, 1996). Dalam jangka panjang pelanggan yang loyal akan cenderung untuk membeli lebih banyak produk dan kurang menuntut perusahaan (Rabinovich & Bailey, 2004). Jika terjadi kesalahan kecil, pelanggan loyal akan dengan mudah mengabaikan masalah, mereka tidak akan keberatan jika harga naik dan juga akan memberikan review kata dari mulut ke mulut mengenai perusahaan dan produk kepada orang lain. Meningkatkan loyalitas dapat membawa penghematan biaya untuk perusahaan dalam setidaknya enam bidang(Griffin, 2005):

a. Biaya pemasaran menjadi berkurang (biaya pengambilalihan pelanggan lebih tinggi daripada biaya mempertahankan pelanggan);
b. Biaya transaksi menjadi lebih rendah, seperti negosiasi kontrak dan pemrosesan order;
c. Biaya perputaran pelangganmenjadi berkurang (lebih sedikit pelanggan hilang yang harus digantikan);
d. Keberhasilan*cross-selling*menjadi meningkat, menyebabkan pangsa pelanggan yang lebih besar;
e. Pemberitaan dari mulut ke mulut menjadi lebih positif, dengan asumsi para pelanggan yang loyal juga merasa puas;
f. Biaya kegagalan menjadi menurun (pengurangan pengerjaan ulang, klaim garansi, dan sebagainya).

Menciptakan pelanggan yang setia adalah jantung pada setiap bisnis. Seperti dalam buku *Marketing Management*(Kotler & Keller, 2012) menyebutkan bahwa ahli pemasaran Don Peppers dan Martha Rogers mengatakan: “Satu-satunya nilai perusahaan adalah nilai yang berasal dari pelanggan yang Anda miliki sekarang dan yang akan Anda miliki di masa depan. Bisnis yang sukses itu dengan cara mendapatkan, menjaga, dan memperbanyak pelanggan. Pelanggan adalah satu-satunya alasan Anda untuk membangun pabrik, mempekerjakan karyawan, membuat jadwal pertemuan, atau terlibat dalam kegiatan bisnis. Tanpa pelanggan, “Anda tidak memiliki bisnis”. Akibatnya, pelanggan yang loyal adalah dasar alami untuk keuntungan jangka panjang dan pertumbuhan bagi perusahaan (Rabinovich & Bailey, 2004).

*(lembar ini sengaja dikosongkan)*

# BAB 3
# METODOLOGI PENELITIAN

Bab ini akan menguraikan tahap-tahap yang akan dilakukan pada penelitian. Secara umum, tahapan penelitian dapat dilihat pada gambar 3.1 Tahapan identifikasi masalah, perumusan masalah, dan penentuan tujuan, batasan penelitian, dankontribusipenelitiantelah dijelaskan pada bab pendahuluan. Sedangkantahap studi literatur yang mencakup landasan teori dan kajian pustaka telah dijelaskan pada bab 2. Pada bab 3 ini dijelaskan mengenai tahapan dalam penyusunan penelitian mulai daripenyiapan data, praproses data, klasifikasi loyalitas pelanggan, skenario uji coba dan penyusunan kesimpulan dan saran pengembangan penelitian lebih lanjut.

Pada gambar 3.3 dijelaskan bahwa metodologi penelitian terdiri dari dari beberapa tahap, yaitu identifikasi masalah, studi literatur, perumusan masalah, penetapan tujuan, batasan dan kontribusi penelitian, penyiapan data, praproses data, klasifikasi loyalitas pelanggan, uji coba dan analisa hasil serta penyusunan kesimpulan dan saran pengembangan penelitian lebih lanjut.

## 3.1. Penyiapan Data

Data yang digunakan berupa hasil penyebaran kuisonerdari para pelanggan *fast moving consumer goods* terhadap beberapa merek mie instan yang menduduki peringkat TOP *Brand Awards* Fase 1 Tahun 2016. Daftar merek mie instan tersebut ditunjukkan pada gambar 3.1.

| MEREK | TBI | TOP |
|---|---|---|
| Indomie | 78.7% | TOP |
| Mi Sedaap | 12.5% | TOP |
| Sarimi | 3.6% | |
| Supermi | 3.0% | |

Gambar 3.1 Daftar Merek Mie Instan

(Sumber: TOP Brand Awards Fase 1 Tahun 2016)

Berdasarkan data yang diperoleh dari *World Instant Noodles Association* (data yang di-*update* pada bulan Mei tahun 2015), negara Indonesia menempati peringkat kedua sebagai negara dengan permintaan atau mengonsumsi mie instant terbanyak. Data tersebut ditunjukkan pada Gambar 3.2.

Updated on May 13, 2015

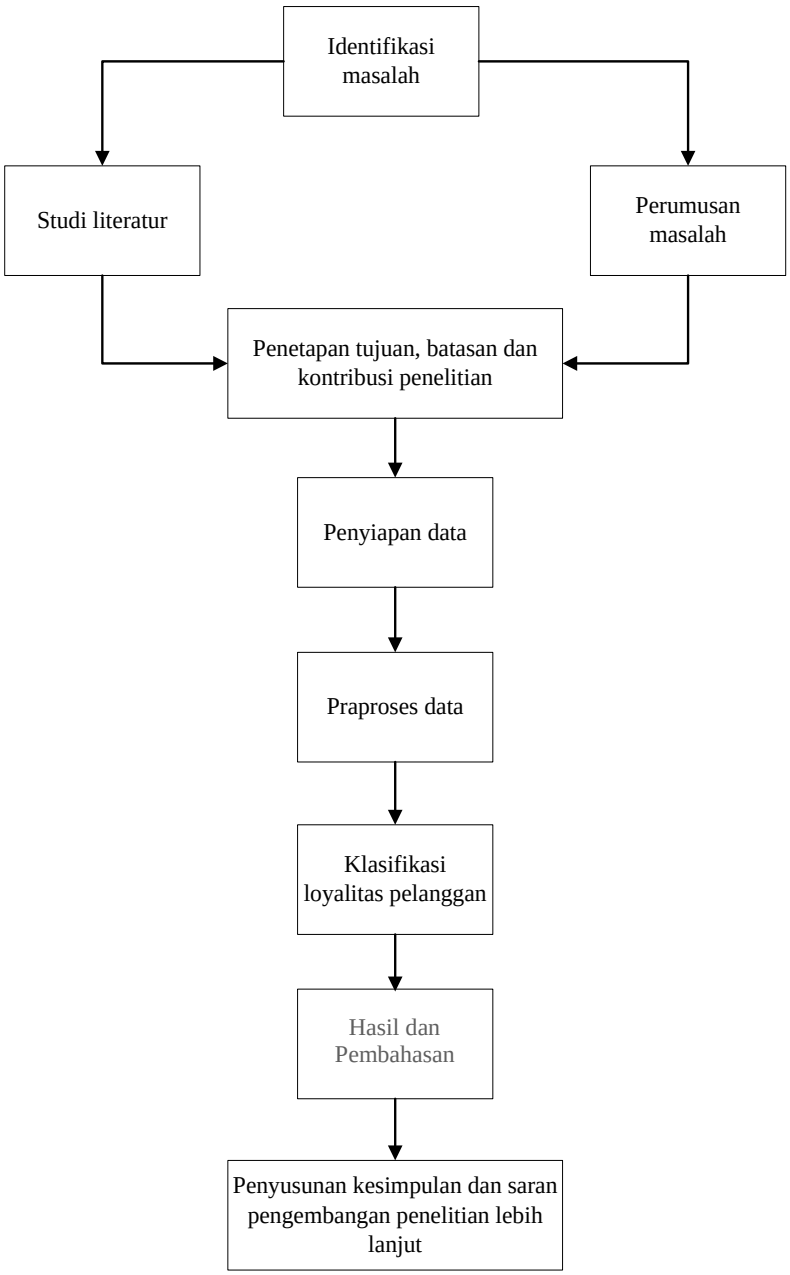
| | Country / Region | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|
| 1 | China / Hong Kong | 42,300 | 42,470 | 44,030 | 46,220 | **44,400** |
| 2 | Indonesia | 14,400 | 14,530 | 14,750 | 14,900 | **13,430** |
| 3 | Japan | 5,290 | 5,510 | 5,410 | 5,520 | **5,500** |
| 4 | India | 2,940 | 3,530 | 4,360 | 4,980 | **5,340** |
| 5 | Vietnam | 4,820 | 4,900 | 5,060 | 5,200 | **5,000** |
| 6 | USA | 4,180 | 4,270 | 4,340 | 4,350 | **4,280** |
| 7 | Republic of Korea | 3,410 | 3,590 | 3,520 | 3,630 | **3,590** |
| 8 | Thailand | 2,710 | 2,880 | 2,960 | 3,020 | **3,070** |
| 9 | Philippines | 2,700 | 2,840 | 2,720 | 2,720 | **2,800** |
| 10 | Brazil | 2,000 | 2,140 | 2,320 | 2,480 | **2,360** |
| 11 | Russia | 1,900 | 2,060 | 2,090 | 2,120 | **1,940** |
| 12 | Nigeria | 1,180 | 1,260 | 1,340 | 1,430 | **1,520** |
| 13 | Malaysia | 1,220 | 1,320 | 1,300 | 1,350 | **1,340** |

Gambar 3.2 Data Permintaan Instant Noodles secara Global

(Sumber: *World Instant Noodles Association/ WINA)*

Dengan banyaknya jumlah permintaan mie instant di negara Indonesia, menuntut para pelaku industri mie instan bersaing secara ketat. Salah satu strategi yang digunakan untuk bertahan di masyarakat adalah dengan mempertahankan pelanggan yaitu dengan mengetahui faktor-faktor yang mempengaruhi loyalitas pelanggan. Dan salah satu metode untuk mengetahui faktor-faktor tersebut adalah dengan melakukan prediksi menggunakan klasifikasi pohon keputusan. Variabel atau fitur yang akan digunakan dalam membentuk klasifikasi pohon keputusan pada penelitian ini berupa data demografis (Buckinx & Poel, 2005; Chu, Tsai, & Ho, 2007; Hung, Yen, & Wang, 2006; Kim Y. , 2006; Tsai & Chen, 2010), data psikografis atau perilaku pelanggan (Kim, Song, & Kim, 2005; Kim Y. , 2006; Larivie`re & Poel, 2005; Tsai & Chen, 2010), data transaksi (Tsai & Chen, 2010; Kim, Jung, Suh, & Hwang, 2006; Kim, Song, & Kim, 2005; Hung, Yen, & Wang, 2006; Buckinx & Poel, 2005; Chu, Tsai, & Ho, 2007; Cheng, Chiu, Cheng, & Wu, 2012), data produk atau barang (Aktepe, Ersoz, & Toklu, 2015) dan promosi

(Rygielski, Wang, & Yen, 2002; Buckinx & Poel, 2005; Chu, Tsai, & Ho, 2007). Tabel 3.1 akan menerangkan mengenai bagian fitur dari masing – masing variabel atau fitur.




Gambar 3.3 Diagram alur metodologi penelitian

Tabel 3.1. Daftar Bagian Fitur (Atribut) dari Masing – Masing Fitur atau Variabel

| Data Demografis | Data Psikografis atau Perilaku Pelanggan | Data Transaksi | Data Produk atau Barang | Promosi |
|---|---|---|---|---|
| Usia | Pindah Merek | Jumlah beli | Tampilan Produk | Media Promosi |
| Jenis Kelamin | Kepuasan Harga | Jumlah pengeluaran | Merek | Rekomendasi |
| Alamat | Konsumsi Merek Lain | Jarak pembelian | | Komentar |
| Status Pernikahan | Lama Konsumsi | Tempat Pembelian | | |
| Pekerjaan | Alasan mengonsumsi | | | |
| Pendidikan | Alasan pindah merek | | | |
| Status Tinggal | Perilaku Kembali | | | |
| | Rata-Rata konsumsi | | | |
| | Jumlah sekali konsumsi | | | |
| | Kepuasan Merek | | | |

## 3.2. Praproses Data

Data yang telah terkumpul, baik data yang digunakan untuk tahap latih maupun tahap uji perlu melalui tahap praproses sebelum data-data tersebut diolah lebih lanjut. Pada tahap praproses data dilakukan pemilihan fitur optimum atau yang mendekati optimum dengan label kelas menggunakan mutual informasi, agar pohon keputusan yang dibangun memiliki kinerja akurasi yang tinggi. Tahapan pemilihan fitur dilakukan guna mengurangi jumlah fitur, menghilangkan relevan, redundansi, atau *noise*, dan membawa efek langsung untuk aplikasi yaitu mempercepat algoritma penggalian data, meningkatkan kinerja pertambangan seperti akurasi prediksi dan hasil yang komprehensif.

Penelitian ini menggunakan pemilihan fitur menggunakan metode filter, metode tersebut dapat menghemat waktu dan mampu menangani masalah *over-fitting* yang disebabkan oleh klasifikasi ketergantungan pada *clasifier*(Zhang, Li,

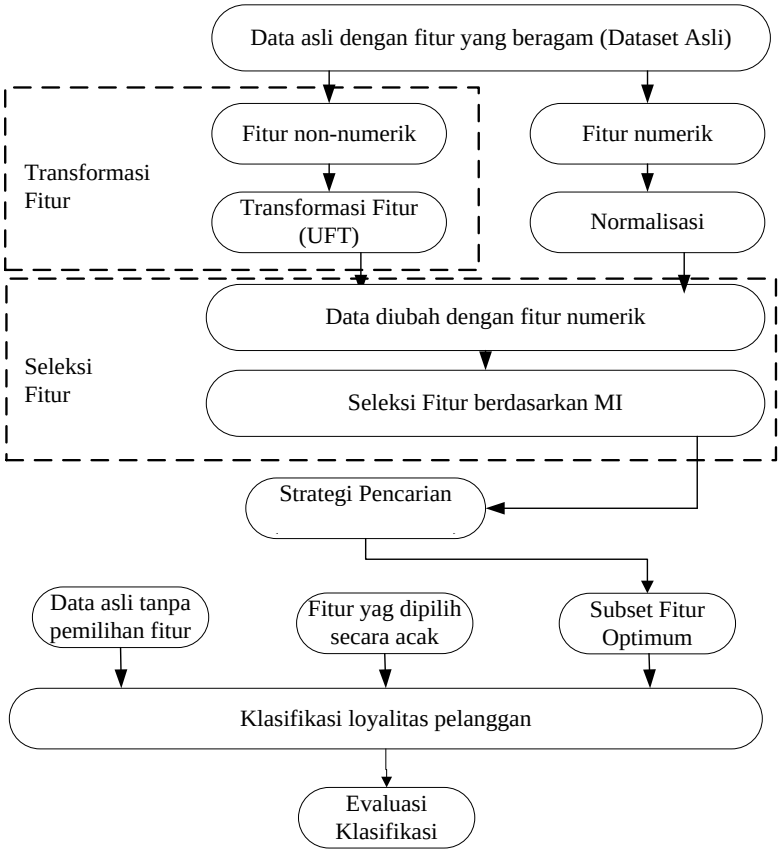
Scarf, & Ball, 2011). Metode gabungan transformasi fitur dengan algortima pemilihan fitur *filter* digunakan dalam tahapan ini, selain dapat mengubah fitur non numerik dengan tepat tanpa adanya bias dalam label kelas, tetapi juga dapat memelihara keunggulan algoritma pemilihan fitur *filter* untuk dataset dengan fitur yang beragam (Wei, Chow, & Chan, 2015a). Gambar 3.4 menjelaskan tentang alur praproses data.

Pada gambar 3.4 dijelaskan bahwa dalam tahap praproses digunakan metode transformasi fitur (*unsupervised transformation feature*) yang dapat mengubah fitur non-numerik menjadi fitur numerik(Wei, Chow, & Chan, 2015a). Fitur non-numerik yang terdapat dalam dataset diubah terlebih dahulu menjadi fitur numerik. Sementara itu, fitur numerik asli dalam dataset harus dinormalisasikan untuk mengurangi perbedaan skala data. Data ditransformasikan dengan fitur numerik terpadu yang dapat digunakan untuk pemilihan fitur. Pemilihan fitur menggunakan algoritma pemilihan fitur *filter* yaitu yaitu metode pemilihan fitur berdasarkanmutual informasi,metode tersebut dapat dipergunakan untuk memperkirakan mutual informasi antara fitur dan label kelas(Kwak & Choi, 2002). Hasil dari pencarian fitur optimum akan digunakan untuk klasifikasi loyalitas pelanggan. Kemudian, akan dilakukan perbandingan hasil kinerja klasifikasi berdasarkan penggunaan seluruh fitur, fitur terpilih menggunakan metode *dynamic* mutual informasi, fitur yang terpilih menggunakan metode *p-value* dan fitur yang dipilih berdasarkan perkiraan peneliti.

### 3.3. Klasifikasi Loyalitas Pelanggan

Berdasarkan latar belakang yang telah dipaparkan sebelumnya, penelitian ini menerapkan model pohon keputusan untuk menganalisis loyalitas pelanggan terhadap merek produk dalam penggunaan barang-barang konsumen yang bergerak cepat dengan menambahkan transformasi fitur dan pemilihan fitur guna akurasi kinerja yang lebih baik. Pohon keputusan adalah sebuah diagram alir yang mirip dengan struktur pohon, di mana setiap *internal node* menotasikan atribut atau subset fitur yang akan diuji, setiap cabangnya merepresentasikan hasil dari atribut tes tersebut dan *leaf node* merepresentasikan kelas-kelas tertentu atau distribusi dari kelas-kelas (Mandasari & Tama, 2011). Struktur dari pohon

keputusan dibangun dari tiga tipe simpul, yaitu simpul *root*, simpul perantara dan simpul *leaf* (Badriyah & Rahmawati, 2006). Simpul *leaf* memuat suatu keputusan akhir atau kelas target untuk suatu pohon keputusan. Simpul *root* adalah titik awal dari suatu pohon keputusan. Setiap simpul perantara berhubungan dengan suatu pernyataan atau pengujian.




Gambar 3.4 Alur Praproses Data

Algoritma C4.5 adalah salah satu metode untuk membuat pohon keputusan berdasarkan data pelatihan yang telah disediakan(Hartama, 2011). Algoritma C4.5merupakan pengembangan dari ID3.Beberapa pengembangan yang dilakukan pada C4.5adalah sebagai antara lain bisa mengatasi *missing value*, bisa mengatasi data kontinyu dan *pruning*. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut (Astuti, 2011):

a. Pilih atribut sebagai *root;*
b. Buat cabang untuk masing-masing nilai;
c. Bagi kasus dalam cabang;

d. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Proses pada pohon keputusan adalahmengubah bentuk data (tabel) menjadimodelpohon,mengubah model pohon menjadi aturan,dan menyederhanakan aturan(Hartama, 2011). Berikut adalah algoritma pembentukan pohon keputusan C4.5:

*Input* : sampel *training*, label *training*, atribut

a. Membuat simpul akar untuk pohon yang akan dibuat;
b. Jika semua sampel positif, berhenti dengan suatu pohon dengan satu simpul akar, beri tanda (+);
c. Jika semua sampel negatif, berhenti dengan suatu pohon dengan satu simpul akar, beri tanda (-);
d. Jika atribut kosong, berhenti dengan suatu pohon dengan suatu simpul akar, dengan label sesuai nilai terbanyak yang ada pada label *training*;
e. Untuk yang lain, mulai
    1) A ----- atribut yang mengklasifikasikan sampel dengan hasil terbaik (berdasarkan *gain* rasio);
    2) Atribut keputusan untuk simpul akar -----A;
    3) Untuk setiap nilai, $v_i$yang mungkin untuk A;
        a) Tambahkan cabang di bawah akar yang berhubungan dengan A=$v_i$*;*
        b) Tentukan sampel S$v_i$ sebagai subset dari sampel yang mempunyai nilai $v_i$untuk atribut A;
        c) Jika sampel S$v_i$ kosong;
            (1) Di bawah cabang tambahkan simpul daun dengan label = nilai yang terbanyak yang ada pada label *training;*
            (2) Yang lain tambah cabang baru di bawah cabang yang sekarang C4.5 (sampel *training*, label *training*, atribut –[A]);
        d) Berhenti

Pada setiap *node* dari pohon, C4.5 memilih satu atribut data yang paling efektif dalam membagi himpunan dari sampel ke subset diperkaya dalam satu kelas atau yang lain. Kriteria adalah keuntungan informasi dinormalisasi (perbedaan entropi) hasildaripemilihan atribut untuk membelah data. Atribut

dengan*information gain* tertinggidinormalisasi dipilih untuk membuat keputusan. Algoritma C4.5 kemudian *recurses* padasublists lebih kecil. Algoritma ini memiliki beberapa kasus dasar. Semua sampel dalam daftar ini termasuk ke dalam kelas yang sama. Ketika ini terjadi, itu hanya menciptakan *node* daun untuk pohon keputusan dalam memilih kelastersebut. Tidak ada fitur yang memberikan keuntungan informasi. Dalamhal ini, pohon keputusanC4.5 menciptakan *node* lebih tinggi dengan menggunakan nilai yang diharapkan dari kelas. Sekali lagi, pohon keputusan C4.5 menciptakan *node*yang lebih tinggi menggunakan nilai yang diharapkan (Quinlan, 1993).

Mengubah pohon atau *tree* yang dihasilkan dalam beberapa aturan. Jumlah aturan sama dengan jumlah *path* yang mungkin dapat dibangun dari *root* sampai *leaf node* (Hartama, 2011). Pengetahuan yang diperoleh dari pohon keputusan dapat direpresentasikan dalam bentuk aturan klasifikasi IF – THEN. Nilai suatu atribut akan menjadi bagian *anticendent* (bagian IF), sedang daun (*leaf*) pada pohon keputusan akan menjadi *consequent* (THEN). Aturan seperti ini sangat membantu manusia dalam memahami model klasifikasi, terutama jika ukuran pohon keputusan terlalu besar.

### 3.4. Skenario Uji Coba dan Analisis Hasil

Pada bagian ini akan dijelaskan mengenai skenario uji coba, skenario uji coba merupakan rencana uji coba. Sehingga analisis dari uji coba yang dilakukan dapat menjawab rumusan masalah dan tujuan yang telah ditetapkan sebelumnya.Penelitian ini berfokus pada pemilihan fitur pada tahap praproses terhadap metode klasifikasi. Pada penelitian ini akan dilakukan tiga skenario uji coba. Skenario pertama membahas tentang transformasi fitur heterogen menjadi fitur homogen dan pemilihan fitur yang paling terkait dengan label kelas dalam analisis loyalitas pelanggan *fast moving consumer goods* berdasarkan nilai mutual informasi antara fitur dengan label kelas.

Skenario kedua adalah melakukan klasifikasi loyalitas pelanggan menggunakan metode pohon keputusanberdasarkan:

a. Menggunakan seluruh fitur dari data aslinya (sebelum dilakukan pemilihan fitur);

b. Menggunakan metode pemilihan fitur yang diusulkan;
c. Menggunakan metode pemilihan fitur *p-value;*
d. Menggunakan fituryang dipilih berdasarkan perkiraan peneliti.

Tujuan dari uji coba pemilihan fitur ini adalah untuk mencari jumlah fitur yang dapat meningkatkan kinerja klasifikasi dalam memprediksi loyaitas pelanggan FMCG. Uji coba skenario kedua menggunakan metode pohon keputusan yang mampu mengidentifikasikan aturan-aturan yang berguna dalam memprediksi hasil akhir. Model prediksi diolah dengan menggunakan Weka dan divalidasi dengan metode 10 *fold cross validation.* Berdasarkan skenario uji coba yang telah dijelaskan di atas, maka dilakukan analisis perbandingan tingkat akurasi untuk klasifikasi loyalitas pelanggan. Evaluasi yang dilakukan menggunakan nilai akurasi, presisi, *recall* dan *f-measure.*

Sedangkan skenario ketiga adalah merepresentasikan hasil dari klasifkasi loyalitas pelanggan yang memiliki model terbaik dalam aturan/ rule IF-THEN. Syarat-syarat model pohon keputusan terbaik (Firat, 2009) adalah sebagai berikut:

a. Model pohon keputusan yang memiliki jumlah aturan yang paling banyak. Semakin banyak jumlah aturan yang diperoleh, penanganan data juga lebih bervariasi. Apabila ditemukan beberapa iterasi yang menghasilkan *rule* yang sama, *rule* tersebut tidak dapat digunakan karena *rule* dihasilkan dari penanganan data yang sama sehingga tidak bervariasi;
b. Model pohon keputusan yang memiliki tingkat akurasi tertinggi;
c. Model yang mencakup semua kelas target yang mungkin muncul dalam tes set.

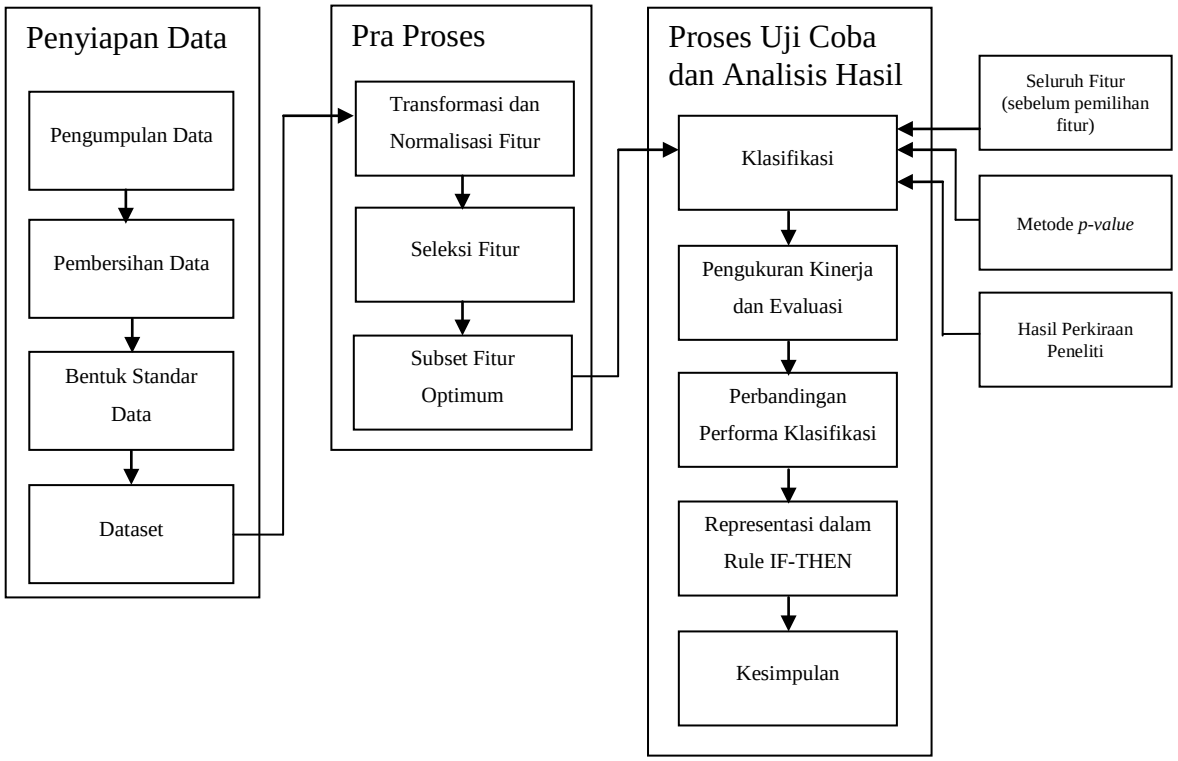
### 3.5. Penyusunan Kesimpulan dan Saran Pengembangan Penelitian Lebih Lanjut

Tahapan ini dilakukan untuk pembuatan kesimpulan pada penelitian yang telah dilakukan, kesimpulan pada penelitian ini akan merangkum jawaban dari permasalahan dan menjawabtujuan yang telah ditentukan pada awal penelitian mulai dirancang. Selain itu, saran pengembangan terhadap penelitian ini juga disusun pada tahap ini.

*(lembar ini sengaja dikosongkan)*

# BAB 4
# HASIL DAN PEMBAHASAN

Bab ini menjelaskan mengenai proses-proses penelitian yang telah dilakukan antara lain penyiapan data, praproses data, klasifikasi loyalitas pelanggan dan uji coba dengan melakukan perbandingan terhadap kinerja berdasarkan data yang diperoleh menggunakan aplikasi atau alat teknik penggalian dataserta analisis hasil dari uji coba yang telah didapatkan. Secara detail proses yang telah dilakukan ditunjukkan pada gambar 4.1.Gambar 4.1 menunjukkan tahapan yang dilakukan dalam melakukan analisis loyalitas pelanggan secara detail. Kotak menunjukkan tahapan yang dilakukan dan garis menunjukkan arah dari setiap tahapan.




Gambar 4.1. Tahapan Analisis Loyalitas Pelanggan

## 4.1. Penyiapan Data

Data yang digunakan untuk melakukan klasifikasi loyalitas pelanggan adalah berupa informasi dari penduduk Lampung yang pernah mengonsumsi produk mie instan. Kuisoner yang disebarkan berupa daftar pertanyaan-pertanyaan yang berhubungan dengan permasalahan dalam penelitian. Data yang diperoleh dari hasil penyebaran kuisoner tidak dapat digunakan secara langsung untuk penelitian, sehingga perlu dilakukan proses penyiapan data agar sesuai dengan metode yang akan digunakan. Proses penyiapan data meliputi pengumpulan data, pembersihan data dan transformasi data sehingga menghasilkan suatu dataset.

### 4.1.1. Pengumpulan Data

Dalam penelitian ini yang dijadikan populasi adalah pengguna atau konsumen mie instan di wilayah Lampung yang jumlahnya tidak diketahui dengan pasti. Karena populasi yang tidak diketahui jumlahnya dan penelitian ini tidak mungkin dapat mempelajari semua yang ada pada populasi, maka diambil sebagian sampel untuk diteliti yang mewakili populasi tersebut. Sampel penelitian ini adalah sebagian penduduk yang pernah mengonsumsi produk mie instan di wilayah Lampung. Jumlah sampel yang diharapkan 100% mewakili populasi adalah jumlah anggota populasi itu sendiri(Putra, Suprayogi, & Kahar, 2013). Untuk jumlah populasi yang terlalu banyak akan diambil untuk dijadikan sampel dengan harapan jumlah sampel yang diambil dapat mewakili populasi yang ada.Dikarenakan jumlah populasi tidak diketahui, maka untuk menentukan jumlah sampel minimum menggunakan rumus *Lemeshow*(Lemeshow, Hosmer Jr, Klar, & Lwanga, 1990), persamaan 4.1 menunjukkan formula dari *Lemeshow*:

$$n = \frac{P(1-P)(Z_{1-\propto/2})^2}{d^2} (\mathbf{4.1})$$

Keterangan:

n : jumlah sampel

P : *maximal estimation* (0,5)

Z : mewakili jumlah kesalahan standar dari *mean* (secara khusus, jika tingkat kepercayaan sebesar 95% maka kesalahan standar dari proporsi populasi adalah 1,96)

d : limit dari error atau presisi absolut (dapat dibuat sekecil mungkin jika ingin meningkatkan ukuran sampel)

Berdasarkan formula pada persamaan 4.1, dianjurkan menggunakan nilai P sebesar 0,5 di rumus untuk ukuran sampel yang cukup mewakili dalam pengamatan(Lemeshow, Hosmer Jr, Klar, & Lwanga, 1990), maka dapat dilakukan penghitungan untuk jumlah minimal sampel:

Z : 95% maka nilainya 1,96

P : 0,5

d : 5%

$$n = \frac{0{,}5(1-0{,}5)(1{,}96)^2}{(0{,}05)^2} \quad = 384{,}16 = 384$$

Sampel yang digunakan dalam penelitian ini yaitu sebanyak 384 responden. Dari penyebaran kuisoner diperoleh 386 responden, 284 responden menyatakan loyal dan 102 menyatakan tidak loyal terhadap merek mie instan yang biasa dikonsumsi. Proses pengumpulan data merupakan tahap awal dari penyiapan data yang bertujuan untuk memperoleh data yang dibutuhkan untuk penelitian. Setelah data diperoleh, selanjutnya data-data tersebut disalin ke dalam *file spreadsheet*. Tujuan dari penyalinan ini adalah memastikan bahwa semua data dan informasi yang berguna telah disalin ke dalam tabel. Contoh hasil pengumpulan data yang telah disalin dalam tabel *spreadsheet* ditunjukkan pada tabel 4.1. Pada tabel 4.1 ditunjukkan contoh hasil pengumpulan data yang telah disalin dalam tabel *spreadsheet*, tabel tersebut terdiri dari beberapa kolom yang disebut dengan fitur.

Tabel 4.1 Contoh Hasil Pengumpulan Data

| Usia | Alamat | Status Pernikahan | Merek | Media | Lama Konsumsi | Jarak Pembelian | Pindah Merek | Alasan Pindah Merek | Jumlah Beli | Jumlah Pengeluaran | Kepuasan Pembelian | Loyal |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24 | Lampung Timur | Belum Menikah | Mie Sedaap | Televisi | <=10 tahun | >= 1 bulan | Pernah | Coba - Coba | > 5 - 10 bungkus | 25000 | Puas | Ya |
| 17 | Bandar Lampung | Belum Menikah | Indomie | Radio | <=10 tahun | >= 1 bulan | Pernah | Coba - Coba | 1 - 5 bungkus | 2000 | Kurang Puas | Tidak |
| 21 | Bandar Lampung | Belum Menikah | Mie Sedaap | Televisi | <=10 tahun | < 1 minggu | Pernah | Coba - Coba | 1 - 5 bungkus | 5000 | Puas | Ya |
| 22 | Pesawaran | Menikah | Mie Sedaap | Televisi | <=10 tahun | < 1 minggu | Pernah | Kualitas rasa lebih enak | 1 - 5 bungkus | 2500 | Sangat Puas | Ya |
| 22 | Lampung Timur | Belum Menikah | Indomie | Televisi | > 10 - 20 tahun | < 1 minggu | Pernah | Kualitas rasa lebih enak | 1 - 5 bungkus | 5000 | Puas | Ya |
| 31 | Bandar Lampung | Menikah | Indomie | Televisi | > 10 - 20 tahun | >= 1-2 minggu | Pernah | Lebih banyak varian rasa | > 5 - 10 bungkus | 15000 | Puas | Ya |

Penelitian ini dilaksanakan dengan menggunakan metode *nonprobability sampling* dan teknik pengambilan sampel *convenience sampling*untuk menentukan siapa saja yang akan dijadikan responden. Dengan menggunakan metode ini, responden yang berhak mengisi kuisoner tergantung sepenuhnya pada kemudahan peneliti (Sekaran, 2003:66 dalam Abubakar, 2009). Teknik ini juga disebut dengan teknik *aksidental*. Menurut Sugiyono (2006 dalam Abubakar, 2009), sampel *aksidental* adalah teknik penentuan responden berdasarkan siapa saja yang secara kebetulan dipandang cocok sebagai sumber data maka akan diberikan kuisoner.*Nonprobability sampling* adalah pengambilan sampel yang tidak memberi peluang atau kesempatan yang sama bagi setiap unsur atau anggota populasi untuk dipilih menjadi sampel dalam penelitian.

Seperti yang sudah dibahas pada Bab 3, fitur yang akan digunakan pada penelitian ini terdiri dari data demografis pelanggan, data psikografis atau perilaku pelanggan, data transaksi atau barang dan data promosi. Jumlah keseluruhan fitur adalah sebanyak dua puluh enam (26) fitur dan satu (1) fitur sebagai label kelas yang digunakan untuk menganalisis loyalitas pelanggan terhadap merek produk *fast moving consumer goods*. Tabel 4.2. akan menjelaskan mengenai deskripsi dari masing-masing fitur.

Tabel 4.2. Deskripsi masing-masing fitur

| Fitur | Nama Fitur | Deskripsi |
|---|---|---|
| f1 | Jenis Kelamin | Jenis kelamin pelanggan |
| f2 | Usia | Usia pelanggan (pada saat mengisi kuisoner) |
| f3 | Alamat | Alamat asal pelanggan berdasarkan kota atau kabupaten (Propinsi Lampung terdiri dari 13 kabupaten dan 2 kota) |
| f4 | Status Pernikahan | Status perkawinan pelanggan |
| f5 | Status Tinggal | Status tinggal pelanggan (seperti: kos, ikut orang tua, rumah sendiri, dsb) |
| f6 | Pekerjaan | Pekerjaan pelanggan pada saat mengisi kuisoner |
| f7 | Pendidikan | Pendidikan pelanggan pada saat mengisi kuisoner |
| f8 | Merek | Merek mie instan yang paling sering atau biasa dikonsumsi (dibagi menjadi 4, yaitu Indomie, Mie Sedaap, Sarimie dan Supermie) |
| f9 | Tempat Pembelian | Tempat pelanggan biasa mendapatkan produk mie instan atau melakukan transaksi pembelian |

| Fitur | Nama Fitur | Deskripsi |
|---|---|---|
| f10 | Media Promosi | Media promosi sebagai sumber informasi pelanggan tentang merek produk |
| f11 | Alasan mengonsumsi | Alasan pelanggan memilih merek tertentu |
| f12 | Lama Konsumsi | Durasi mengonsumsi produk mie instan |
| f13 | Jarak pembelian | Jarak pembelian sebelumnya dengan pembelian berikutnya |
| f14 | Jumlah beli | Rata-rata jumlah beli pelanggan dalam satu kali transaksi |
| f15 | Rata-Rata konsumsi | Rata-rata mengonsumsi produk mie instan merek tertentu dalam kurun waktu satu bulan |
| f16 | Tampilan Produk | Tampilan merek produk yang menarik |
| f17 | Kepuasan Harga | Kepuasan pelanggan terhadap harga yang ditawarkan dari merek produk yang biasa dikonsumsi |
| f18 | Kepuasan Merek | Kepuasan pelanggan terhadap kualitas merek produk yang biasa dikonsumsi |
| f19 | Rekomendasi | Kegiatan merekomendasikan ke orang lain untuk menggunakan merek produk yang sama |
| f20 | Komentar | Memberikan komentar positif terhadap merek produk |
| f21 | Jumlah pengeluaran | Rata-rata jumlah pengeluaran (uang yang dihabiskan) untuk membeli produk mie instan merek tertentu dalam satu kali transaksi |
| f22 | Jumlah sekali konsumsi | Jumlah dalam satu kali mengonsumsi (dalam satuan bungkus) |
| f23 | Pindah Merek | Perilaku pelanggan dalam berpindah merek |
| f24 | Alasan pindah merek | Alasan pelanggan berpindah ke merek lain |
| f25 | Konsumsi Merek Lain | Rata-rata jumlah konsumsi merek lain dalam waktu satu bulan |
| f26 | Perilaku Kembali | Untuk mengidentifikasi apakah pelanggan pernah berpindah merek atau tidak |
| f27 | Loyalitas Pelanggan | Label Kelas yang menyatakan apakah pelanggan akan loyal atau tidak dengan merek yang biasa dikonsumsi |

### 4.1.2. Bentuk Standar Data

Selanjutnya setelah data mengalami proses pembersihan maka data ditransfer kedalambentuk standar. Bentuk standar adalah bentuk data yang akan diakses oleh*tools data mining*, dalam penelitian ini menggunakan *tools* WEKA. Bentuk standar yang dapat diakses oleh WEKA adalah *file* dalam format .arff

(*attribute-relation file format*) atau .csv (*comma separated value*). Gambar 4.2 menunjukkan contoh bentuk standar *file* bertipe .arff.

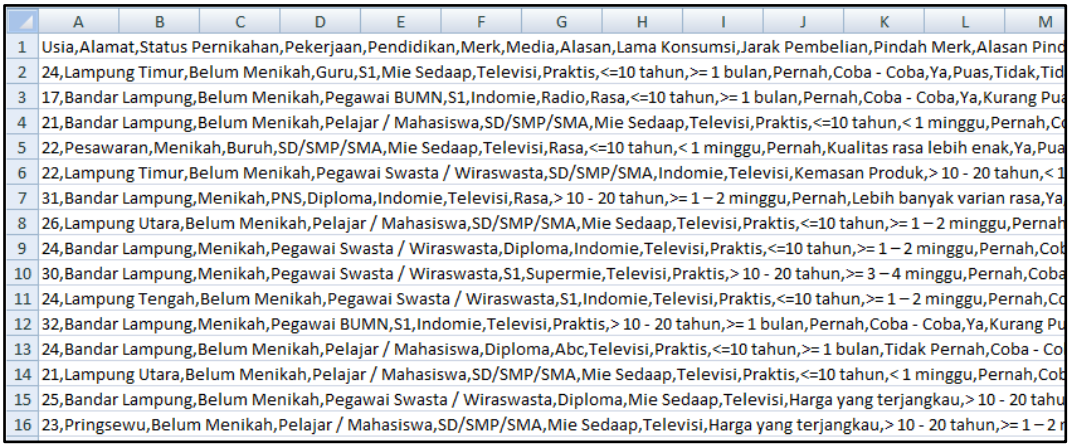
```
@attribute 'Jumlah Pengeluaran' numeric
@attribute 'Jumlah Sekali Konsumsi' numeric
@attribute 'Jenis Kelamin' {Pria,Wanita}
@attribute 'Status Tinggal' {'Ikut Orang Tua',Kos,'Rumah
Sendiri','ikut saudara','rumah kantor'}
@attribute 'Rata-Rata Konsumsi' numeric
@attribute 'Tempat Pembelian'
{Supermarket,Pasar,Warung,Toko,Indomart}
@attribute 'Konsumsi Merk Lain' numeric
@attribute 'Tampilan Produk' {Tidak,Ya}
@attribute 'Kepuasan Merk' {Puas,'Kurang Puas','Sangat Puas'}
@attribute class Loyal {Ya,Tidak}

@data
24,'Lampung Timur','Belum Menikah',Guru,S1,'Mie
Sedaap',Televisi,Praktis,'<=10 tahun','>= 1
bulan',Pernah,'Coba - Coba',Ya,Puas,Tidak,Tidak,5-
10,25000,1,Pria,'Ikut Orang Tua',5,Supermarket,2,Tidak,Puas,Ya
17,'Bandar Lampung','Belum Menikah','Pegawai
BUMN',S1,Indomie,Radio,Rasa,'<=10 tahun','>= 1
```

Gambar 4.2 Bentuk Standar file .arff

Pada gambar 4.2 ditunjukkan bentuk standar file dengan format .arff, *file* .arff adalah sebuah gile teks ASCII yang berisi daftar *instances* dalam sekumpulan atribut(Pranatha, 2012). Data dalam format .arff harus memenuhi syarat sebagai berikut:

a. Data dipisahkan dengan koma, dengan kelas sebagai atribut terakhir;
b. Bagian *header* diawali dengan @relation;
c. Tiap atribut ditandai dengan @attribute. Tipe-tipe data dalam Weka adalah *numeric* (*real* atau *integer*), nominal, *string* dan *date;*
d. Bagian data diawali dengan @data.

Sedangkan gambar 4.3 menunjukkan contoh bentuk standar *file* bertipe .csv yang dapat juga diakses oleh Weka. Pada gambar 4.2 dan 4.3 menunjukkan bentuk standar *file* .arff dan .csv, tampilan dari kedua bentuk tersebut berbeda dengan tampilan di *spreadsheet*. Tampilan antar fitur dalam *spreadsheet* dipisahkan oleh kolom, sedangkan dalam *file* .arff dan .csv, antar fitur dipisahkan oleh koma (,).



| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Usia,Alamat,Status Pernikahan,Pekerjaan,Pendidikan,Merk,Media,Alasan,Lama Konsumsi,Jarak Pembelian,Pindah Merk,Alasan Pind | | | | | | | | | | | | |
| 2 | 24,Lampung Timur,Belum Menikah,Guru,S1,Mie Sedaap,Televisi,Praktis,<=10 tahun,>= 1 bulan,Pernah,Coba - Coba,Ya,Puas,Tidak,Tid | | | | | | | | | | | | |
| 3 | 17,Bandar Lampung,Belum Menikah,Pegawai BUMN,S1,Indomie,Radio,Rasa,<=10 tahun,>= 1 bulan,Pernah,Coba - Coba,Ya,Kurang Pua | | | | | | | | | | | | |
| 4 | 21,Bandar Lampung,Belum Menikah,Pelajar / Mahasiswa,SD/SMP/SMA,Mie Sedaap,Televisi,Praktis,<=10 tahun,< 1 minggu,Pernah,Co | | | | | | | | | | | | |
| 5 | 22,Pesawaran,Menikah,Buruh,SD/SMP/SMA,Mie Sedaap,Televisi,Rasa,<=10 tahun,< 1 minggu,Pernah,Kualitas rasa lebih enak,Ya,Pua | | | | | | | | | | | | |
| 6 | 22,Lampung Timur,Belum Menikah,Pegawai Swasta / Wiraswasta,SD/SMP/SMA,Indomie,Televisi,Kemasan Produk,> 10 - 20 tahun,< 1 | | | | | | | | | | | | |
| 7 | 31,Bandar Lampung,Menikah,PNS,Diploma,Indomie,Televisi,Rasa,> 10 - 20 tahun,>= 1 – 2 minggu,Pernah,Lebih banyak varian rasa,Ya | | | | | | | | | | | | |
| 8 | 26,Lampung Utara,Belum Menikah,Pelajar / Mahasiswa,SD/SMP/SMA,Mie Sedaap,Televisi,Praktis,<=10 tahun,>= 1 – 2 minggu,Pernah | | | | | | | | | | | | |
| 9 | 24,Bandar Lampung,Menikah,Pegawai Swasta / Wiraswasta,Diploma,Indomie,Televisi,Praktis,<=10 tahun,>= 1 – 2 minggu,Pernah,Cob | | | | | | | | | | | | |
| 10 | 30,Bandar Lampung,Menikah,Pegawai Swasta / Wiraswasta,S1,Supermie,Televisi,Praktis,> 10 - 20 tahun,>= 3 – 4 minggu,Pernah,Coba | | | | | | | | | | | | |
| 11 | 24,Lampung Tengah,Belum Menikah,Pegawai Swasta / Wiraswasta,S1,Indomie,Televisi,Praktis,<=10 tahun,>= 1 – 2 minggu,Pernah,Co | | | | | | | | | | | | |
| 12 | 32,Bandar Lampung,Menikah,Pegawai BUMN,S1,Indomie,Televisi,Praktis,> 10 - 20 tahun,>= 1 bulan,Pernah,Coba - Coba,Ya,Kurang Pu | | | | | | | | | | | | |
| 13 | 24,Bandar Lampung,Belum Menikah,Pelajar / Mahasiswa,Diploma,Abc,Televisi,Praktis,<=10 tahun,>= 1 bulan,Tidak Pernah,Coba - Co | | | | | | | | | | | | |
| 14 | 21,Lampung Utara,Belum Menikah,Pelajar / Mahasiswa,SD/SMP/SMA,Mie Sedaap,Televisi,Praktis,<=10 tahun,< 1 minggu,Pernah,Cob | | | | | | | | | | | | |
| 15 | 25,Bandar Lampung,Belum Menikah,Pegawai Swasta / Wiraswasta,Diploma,Mie Sedaap,Televisi,Harga yang terjangkau,> 10 - 20 tahu | | | | | | | | | | | | |
| 16 | 23,Pringsewu,Belum Menikah,Pelajar / Mahasiswa,SD/SMP/SMA,Mie Sedaap,Televisi,Harga yang terjangkau,> 10 - 20 tahun,>= 1 – 2 r | | | | | | | | | | | | |

Gambar 4.3 Bentuk Standar file .csv

## 4.2. Lingkungan Uji Coba

Lingkungan uji coba pada penelitian ini terdiri dari dua jenis perangkat, yaitu perangkat keras dan perangkat lunak. Spesifikasi perangkat pengujian yang digunakan untuk menguji sistem dijelaskan dalam tabel 4.3 dan tabel 4.4. Tabel 4.3 menjelaskan mengenai spesifikasi perangkat keras.

Tabel 4.3 Spesifikasi Lingkungan Uji Coba - Perangkat Keras

| **Perangkat Keras** | **Spesifikasi** |
|---|---|
| Jenis | Notebook |
| Processor | Intel(R) Atom(TM) CPU N2800 @ 1.86GHz (4 CPUs), ~1.9GHz |
| RAM | 2 GB |

Selain perangkat keras, dalam menguji sistem juga digunakan perangkat lunak. Tabel 4.4 menjelaskan mengenai spesifikasi perangkat lunak yang digunakan.

Tabel 4.4 Spesifikasi Lingkungan Uji Coba - Perangkat Lunak

| **Perangkat Lunak** | **Spesifikasi** |
|---|---|
| Sistem Operasi | Windows 7 |
| *Tools* | ✓ Matlab 2015<br>✓ Weka |

## 4.3. Pelaksanaan dan Hasil Uji Coba

Sub bab ini menjelaskan tentang pelaksanaan dan hasil uji coba dari skenario uji coba yang telah dijelaskan pada sub bab 3.4.

### 4.3.1. Uji Coba Pemilihan Fitur

Pada bagian ini, uji coba pemilihan fitur dilakukan dengan melakukan transformasi fitur non-numerik menjadi fitur numerik, pemilihan fitur menggunakan metode filter yang megestimasi nilai mutual informasi antara fitur dengan label kelas dengan metode*dynamic* mutual informasi.

#### 4.3.1.1. Transformasi Fitur

Transformasi fitur merupakan cara lain untuk mengatasi pengubahan data heterogen menjadi data homogen dalam tahapan praproses yang bertujuan untuk mengubah fitur non-numerik menjadi fitur numerik sebelum dilakukan proses pemilihan fitur. Transformasi yang diusulkan mampu meminimalkan kehilangan informasi, hal ini dikarenakan transformasi fitur tersebut merupakan hubungan analitis antara mutual informasi dan nilai entropi dari masing-masing fitur. Kondisi tersebut dapat mempertahankan informasi dari fitur aslinya ketika fitur non-numerik diubah menjadi fitur numerik. Transformasi fitur ini hanya bergantung pada fitur non-numerik asli itu sendiri dan dapat menghindari bias informasi yang diberikan oleh label kelas. Transformasi fitur ini dilakukan dengan cara fitur non-numerik diubah menjadi fitur numerik yang didasarkan pada aturan *gaussian*. Diasumsikan bahwa sekelompok fitur numerik menggantikan fitur non-numerik dan disubstitusi menggunakan aturan *gaussian*. Aturan *gaussian* dipilih karena:

a. Dapat menggambarkan secara umum distribusi probabilitas dari data numerik,
b. Mampu menyederhanakan ekspresi dari MI dan entropi,
c. Bila dibandingkan dengan metode distribusi lainnya yang menggambarkan data numerik (seperti distribusi *uniform*), parameter distribusi *gaussian* lebih mudah untuk memperkirakan rentang data tanpa pengetahuan sebelumnya.

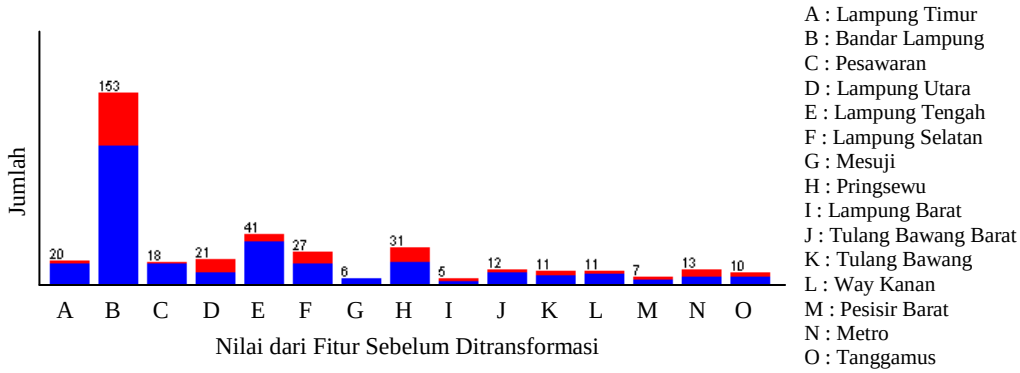
Berdasarkan prosedur transformasi pada gambar 4.4, setiap nilai dari fitur non-numerik dapat digantikan oleh sekelompok nilai-nilai numerik.Pada saat yang sama, substitusi numerik dapat tetap menjaga informasi asli yang terdapat dalam fitur non-numerik. Usulan UFT dapat digunakan untuk tahapan praproses data dan

membuat fitur heterogen menjadi fitur homogen dalam bentuk numerik. Gambar 4.5dan gambar 4.6 menunjukkan nilai distribusi dari salah satu fitur yaitu fitur 'Alamat' sebelum dilakukan transformasi dan setelah dilakukan transformasi.

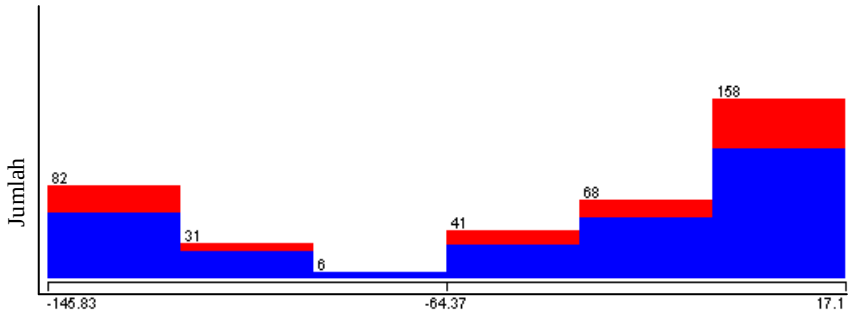
Berikut adalah potongan *source code* yang digunakan untuk mengubah fitur non-numerik menjadi fitur numerik:

```
for pIterasi=1:size(si,1)
      pi(pIterasi)  = size(find(strcmp(si{pIterasi}, cols)),1)/rowNum;
    end
    tmp   = zeros(0,0);
for i=1:length(si)
     sigPi3  = sigPi3+power(pi(i),3);
if (j~=i)
       sigPij  = sigPij+(pi(i)*pj*power((i-j),2));
else
       sigPij  = sigPij+(pi(i)*pj);
    end
for k=1:i
      sigNKP = sigNKP+((n-k)*pi(k));
    end
myui   = ((n-i)-sigNKP)*sqrt((1-sigPi3)/sigPij);
chg    = normrnd(myui,pi(i));
```

Gambar 4.4 Potongan *Source code* UFT




Gambar 4.5 Nilai Distribusi Fitur Alamat Sebelum ditransformasi




Nilai dari Fitur Setelah Ditransformasi

Gambar 4.6 Distribusi Fitur Alamat Setelah ditransformasi

Berdasarkan gambar 4.5 dan 4.6 terlihatbahwa distribusi *gaussian* digunakan untuk mengganti grup nilai fitur non-numerik menjadi fitur numerik dengan rentang nilai -145.83sampai 17.1.Tabel 4.5dan tabel 4.6 menunjukkan contoh fitur asli dan fitur hasil transformasi dari dataset pelanggan.

Tabel 4.5Contoh Fitur Asli

| Alamat | Status Perkawinan | Pekerjaan | Pendidikan | Merek | Media Promosi | Alasan Mengonsumsi |
|---|---|---|---|---|---|---|
| Lampung Timur | Belum Menikah | Tenaga pengajar / Dosen | S1 | Mie Sedaap | Televisi | Praktis |
| Bandar Lampung | Belum Menikah | Pegawai BUMN | S1 | Indomie | Radio | Rasa |
| Bandar Lampung | Belum Menikah | Pelajar / Mahasiswa | SD/ SMP/ SMA | Mie Sedaap | Televisi | Praktis |
| Pesawaran | Menikah | Pegawai Swasta / Wiraswasta | SD/ SMP/ SMA | Mie Sedaap | Televisi | Rasa |
| Lampung Timur | Belum Menikah | Pegawai Swasta / Wiraswasta | SD/ SMP/ SMA | Indomie | Televisi | Kemasan Produk |
| Bandar Lampung | Menikah | PNS | Diploma | Indomie | Televisi | Rasa |
| Lampung Utara | Belum Menikah | Pelajar / Mahasiswa | SD/ SMP/ SMA | Mie Sedaap | Televisi | Praktis |
| Bandar Lampung | Menikah | Pegawai Swasta / Wiraswasta | Diploma | Indomie | Televisi | Praktis |

Tabel 4.6 Contoh Fitur Hasil Transformasi

| Alamat | Status Perkawinan | Pekerjaan | Pendidikan | Merk | Media Promosi | Alasan Mengonsumsi |
|---|---|---|---|---|---|---|
| -45.7724 | -0.06227 | -13.075 | 0.532025 | -2.54194 | 0.016748 | -2.16714 |
| 17.09824 | -0.06227 | 18.94196 | 0.532025 | 1.616909 | 23.99101 | -7.32413 |
| 17.09824 | -0.06227 | -5.77296 | -3.8105 | -2.54194 | 0.016748 | -2.16714 |
| -113.306 | -1.15309 | 3.825943 | -3.8105 | -2.54194 | 0.016748 | -7.32413 |
| -45.7724 | -0.06227 | 3.825943 | -3.8105 | 1.616909 | 0.016748 | 2.540283 |
| 17.09824 | -1.15309 | 32.87412 | 5.028887 | 1.616909 | 0.016748 | -7.32413 |
| -64.2947 | -0.06227 | -5.77296 | -3.8105 | -2.54194 | 0.016748 | -2.16714 |
| 17.09824 | -1.15309 | 3.825943 | 5.028887 | 1.616909 | 0.016748 | -2.16714 |

Berdasarkan tabel 4.5 dan tabel 4.6 didapatkan informasi bahwa setelah melalui tahapan transformasi fitur, nilai-nilai numerik menggantikan fitur yang berisi nilai non-numerik. Sebagai contoh, pada fitur Alamat nilai '-45.7724' menggantikan nilai 'Lampung Timur', nilai '17.09824' menggantikan nilai 'Bandar Lampung', nilai '-113.306' menggantikan nilai 'Pesawaran', nilai '-64.2947' menggantikan nilai 'Lampung Utara'.

#### 4.3.1.2. PemilihanFitur Berdasarkan Mutual Informasi

Setelah melalui tahapan transformasi fitur, tahapan selanjutnya adalah melakukan pemilihan fitur berdasarkan mutual informasi. Mutual informasi dipilih sebagai kriteria pemilihan karena mampu mengukur keterkaitan antara dua variabel secara umum tanpa memperhatikan distribusi fitur. Pemilihan fitur merupakan salah satu tahapan praproses data untuk memilih fitur yang berpengaruh dalam model klasifikasi. Masing-masing fitur dilakukan penghitungan nilai mutual informasi terhadap label kelas dengan pendekatan join mutual informasi. Prosedur metode pemilihan fitur berdasarkan mutual informasi dapat diringkaspada gambar 4.7.

```
D = load('DataPelangganTransformasi.csv');
header = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
24 25 26];
c_index = size(D,2);
FF = D(:,1:c_index-1);
C = D(:,c_index);
%Cari nilai MI Maximum
FF = round(FF);
for i=1:c_index-1
   Fi_c(i) = mutualInformation(abs(FF(:,i)), C);
end
[val,index]= sort(Fi_c,'descend');
index
dataset= D(:,index(1));
for j=2:size(index,2)
  dataset = [dataset D(:,index(i))];
end
dataset = cat(2,dataset, D(:,c_index));
```

Gambar 4.7 Prosedur pemilihan fitur berdasarkan mutual informasi

Proses pemilihan fitur berdasarkan nilai mutual informasi ini dilakukan dengan menghitung hubungan atau keterkaitan dari tiap fitur terhadap label kelas menggunakan prinsip entropi dan mencari nilai maksimal antar fitur dengan label kelas menggunakan prinsip join mutual informasi. Jika nilai mutual informasi lebih dari 0 ($I(x,y) > 0$) maka terdapat hubungan yang kuat antara fitur dengan label kelas. Tetapi jika nilai mutual informasi kurang dari atau sama dengan 0 ($I(x,y) \leq 0$) maka tidak ada hubungan keterikatanantara fitur dengan label kelas,*x* dinotasikan sebagai fiitur ke-*i* dan *y* adalah label kelas. Perhitungan mutual informasi bertujuan untuk mencari fitur-fitur yang paling terkait dengan label kelas atau memiliki keterkaitan informasi yang paling banyak dengan label kelas.

Berikut adalah salah satu contoh perhitungan nilai mutual informasi antara fitur 'Merek' dengan label kelas "Loyal'. Berikut langkah-langkah untuk menghitung nilai mutual informasi:

a. Menghitung nilai distribusi dan hasil dari perhitungan probabilitas tiap nilai fitur dengan nilai kelas dapat dilihat pada tabel 4.7.
b. Menghitung nilai entropi tiap fitur (*H(X)*) dan label kelas (*H(Y)*), entropi merupakan paramater yang menyatakan kandungan informasi dengan menggunakan persamaan:

   Nilai entropi fitur *x:*

   $$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

   *H(X)* = -(0.736 * log (0.736) + 0.264 * log (0.264))

   *H(X)* = 0.8335

   Sedangkan nilai entropi label kelas *y:*

   $$H(Y) = -\sum_{y \in Y} p(y) \log p(y)$$

   *H(Y)* = -(0.448 * log (0.448) + 0.503 * log (0.503) + 0.031 * log (0.031) + 0.018 * log (0.018))

   *H(Y)* = 1.2789

Maka nilai entropi dari fitur *x* adalah sebesar 0.8335 bits dan nilai entropi label kelas *y* adalah sebesar 1.2789 bits.

Tabel 4.7 Hasil Nilai Distribusi dan Probabilitas

| **Nilai Fitur "Merk"** | **Nilai Distribusi** | | **Nilai Probabilitas** | | ***P(y)*** |
|---|---|---|---|---|---|
| | **Ya (*y1*)** | **Tidak (*y2*)** | ***p*(*x*,*y1*)** | ***p*(*x*,*y2*)** | |
| Mie Sedaap | 125 | 48 | 0.324 | 0.124 | 0.448 |
| Indomie | 147 | 47 | 0.381 | 0.122 | 0.503 |
| Supermie | 6 | 6 | 0.016 | 0.016 | 0.031 |
| Sarimie | 6 | 1 | 0.016 | 0.003 | 0.018 |
| ***P(x)*** | | | 0.736 | 0.264 | |

c. Menghitung nilai join entropi antara variabeldengan kelas, sehingga join entropi antara fitur *x* dan label kelas *y* adalah

$$H(X,Y) = -\sum_{x \in X} \sum_{y \epsilon Y} p(x,y) \log p(x,y)$$

*H(X,Y)* = -(0.324 * log (0.324) + 0.381 * log (0.381) + 0.016 * log (0.016) + 0.016 * log (0.016) + 0.124 * log (0.124) + 0.122 * log (0.122) + 0.016 * log (0.016) + 0.003 * log (0.003))

*H(X,Y)* = 2.1039

Berdasarkan perhitungan di atas, maka nilai join entropi antara fitur *x* dan label kelas *y* adalah sebesar 2.1039 bits.

d. Melakukan perhitungan nilai mutual informasi berdasarkan nilai entropi yang telah dihitung sebelumnya, nilai mutual informasi dapat dihitung dengan menggunakan formula:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

*I(X;Y)*= 0.8335 + 1.2789– 2.1039 = 0.0085

Maka nilai mutual informasi antara fitur 'Merek' dan label kelas 'Loyal' adalah sebesar 0.0085. Hal tersebut berarti bahwa fitur 'Merek' memiliki keterkaitan dengan label kelas sebesar 0.0085 bits. Tabel 4.8 menunjukkan

nilai mutual informasi dari setiap fitur terhadap label kelas yang sudah diurutkan.

Tabel 4.8 Nilai MI antara fitur dengan label kelas

| Fitur | *I(x,y)* | Fitur | *I(x,y)* |
|---|---|---|---|
| Jumlah Pengeluaran | 0.0754 | Media Promosi | 0.0065 |
| Rata-Rata Konsumsi | 0.0559 | Kepuasan Harga | 0.0051 |
| Usia | 0.0523 | Jumlah Sekali Konsumsi | 0.005 |
| Alamat | 0.0497 | Kepuasan Merek | 0.0045 |
| Alasan Berpindah Merek | 0.0204 | Status Tinggal | 0.004 |
| Konsumsi Merek Lain | 0.0192 | Tempat Pembelian | 0.0016 |
| Alasan Mengonsumsi | 0.0167 | Pindah Merek | 0.001 |
| Pendidikan | 0.0152 | Status Pernikahan | 0 |
| Pekerjaan | 0.0113 | Perilaku Kembali | 0 |
| Merek | 0.0085 | Rekomendasi | 0 |
| Jumlah Beli | 0.0083 | Komentar | 0 |
| Jarak Pembelian | 0.008 | Jenis Kelamin | 0 |
| Lama Konsumsi | 0.0076 | Tampilan Produk | 0 |

Berdasarkan tabel 4.8 dapat diketahui bahwa dari 26 fitur, fitur ‘Jumlah Pengeluaran’ memiliki nilai mutual informasi tertinggi yang artinya bahwa fitur ‘Jumlah Pengeluaran’ merupakan fitur yang paling berkaitan dengan label kelas. Sedangkan enam fitur dari hasil perhitungan mutual informasi memiliki nilai nol (0) yang berarti bahwa fitur tersebut tidak memiliki informasi yang berkaitan dengan label kelas, fitur tersebut antara lain status pernikahan, perilaku kembali, rekomendasi, komentar, jenis kelamin dan tampilan produk. Setelah diketahui nilai mutual informasi masing-masing fitur dengan label kelas, selanjutnya mencari nilai maksimal dari proses perhitungan mutual informasi $I(f_i;C)$, hasil pencarian nilai maksimal tersebut dilakukan perangkingan untuk membentuk subset fitur.Hasil dari perangkingan menggunakan metode pemilihan fitur berdasarkan skenario uji coba disajikan dalam Tabel 4.9.

Tabel 4.9 Hasil Rangking Pemilihan Fitur Menggunakan Beberapa Metode

| Metode Pemilihan Fitur | Fitur yang Terpilih |
|---|---|
| Sebelum dilakukan Pemilihan Fitur | f1, **f2**, **f3**, f4, f5, **f6**, **f7**, **f8**, f9, f10, f11, f12, f13, f14, f15, f16, f17, f18, f19, f20, f21, f22, f23, f24, f25, f26 |

| Metode Pemilihan Fitur | Fitur yang Terpilih |
|---|---|
| *Dynamic* Mutual Informasi | f21, f15, **f2**, **f3**, f24, f25, f11, **f7**, **f6**, **f8**, f14, f13, f12, f10, f17, f22, f18, f5, f9, f23, f4, f26, f19, f20, f1, f16 |
| *P-Value* | **f6**, f26, f17, f10, f1, **f8**, **f7**, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9, f15, f20, f4, f22, f2, f23 |
| Perkiraan | **f2**, f25, **f3**, f4, **f6**, f15, f21, **f7**, **f8**, f10, f11, f18, f12, f13, f24, f1, f23, f26, f5, f16, f9, f17, f14, f22, f19, f20 |

Fitur yang terpilih menggunakan metode *dynamic* mutual informasi, *p-value* dan perkiraan peneliti terdapat lima fitur yang selalu terpilih, jika jumlah fitur dibatasi sebanyak sepuluh fitur. Fitur-fitur tersebut adalah usia (f2), alamat (f3), pekerjaan (f6), pendidikan (f7) dan merek (f8). Hal tersebut berarti bahwa kelima fitur tersebut relevan terhadap loyalitas pelanggan.

### 4.3.2. Uji Coba Klasifikasi Loyalitas Pelanggan

Uji coba klasifikasi loyalitas pelanggan dilakukan menggunakan metode pohon keputusanyaitu algoritma C4.5dengan evaluasi *k-fold cross validation* dan hasil uji coba berupa nilai akurasi, presisi, *recall* dan *f-measure*. Algoritma pohon keputusansangat berguna untuk mendapatkan pemahaman lebih mendalam mengenai perilaku pelanggan dan mencari cara untuk menindaklanjuti hasil-hasilnya agar mendapatkan keuntungan tambahan (Olson & Shi, 2008). Metode ini relatif lebih unggul dibandingkan algoritma jaringan saraf tiruan dan genetika karena menyediakan aturan-aturan yang dapat dipakai ulang sehingga menjelaskan keimpulan dari model (Michie, 1998).

Ada beberapa tahapan dalam membuat sebuah pohon keputusan dalam agoritma C4.5 Larose (dalamAndriani, 2012)yaitu:

a. Mempersiapkan data
b. Menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai gain dari masing-masing atribut, nilai gain yangpaling tinggi yang akan menjadi akar pertama.
c. Ulangi langkah kedua dan langkah ketiga hingga semua *record* terpatisi.

d. Proses partisi pohon keputusan akan berhenti saat :
    2) Semua*record* dalam simpul N mendapat kelas yang sama.
    3) Tidak ada atribut di dalam*record* yang dipartisi lagi
    4) Tidak ada *record* di dalam cabang yang kosong

Dalam evaluasi *k-fold cross-validation*, data pengujian dipisah secara acak ke dalam *k* himpunan bagian yang *mutually exclusive* atau "*folds* (lipatan)", $D_1$,$D_2$,..., $D_k$ , yang masing-masing kurang lebih berukuran sama. Pelatihan dan pengujian dilakukan sebanyak *k* kali. Pada iterasi ke-*i*, partisi *Di* digunakan sebagai data tes, dan partisi sisanya digunakan bersama untuk melatih model. Dalam iterasi pertama, yaitu himpunan bagian $D_2$, ..., $D_k$secara bersama bertindak sebagai data pelatihan untuk memperoleh model pertama, yang diuji pada $D_1$; iterasi kedua dilatih pada himpunan bagian $D_1$, $D_3$, ..., $D_k$dan diuji pada $D_2$; dan seterusnya. Dalam penelitian ini digunakan *10-fold crossvalidation*, yang berarti dataset pelatihan dan pengujian dilakukan sebanyak 10 kali.

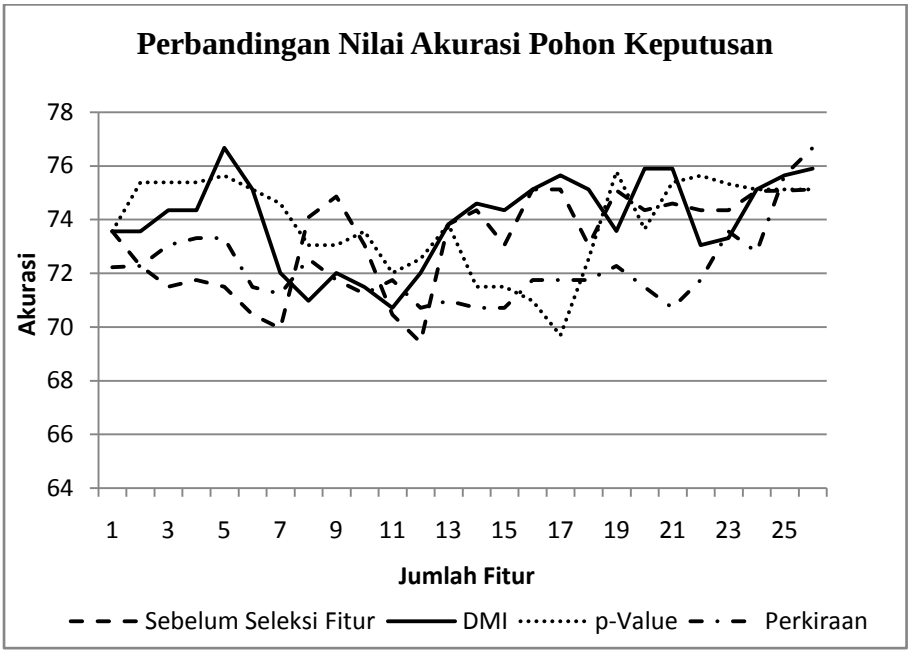
### 4.3.3. Uji Coba Perbandingan Performa Penggunaan Fitur

Pada uji coba perbandingan performa dilakukan dengan mengkombinasikan seluruh jumlah fitursebelum diterapkan metode pemilihan fitur, setelah menerapkan metode pemilihan fitur *dynamic* mutual informasi, metode pemilihan fitur*p-Value*danfitur yang dipilih berdasarkan perkiraan peneliti. Kombinasi fitur tersebut digunakan dalam membangun model klasifikasi pohon keputusan. Dataset dalam uji coba ini dibagi menjadi empat, yaitu:

a. Dataset sebelum dilakukan pemilihan fitur;
b. Dataset hasil rangking dari metode pemilihan fitur *dynamic* mutual informasi;
c. Dataset hasil rangking dari metode pemilihan fitur *p-value*;
d. Dataset hasil rangking dari pemilihan fitur berdasarkan perkiraan peneliti.

Uji coba ini dilakukan untuk mengetahui performa dari pengklasifikasi, yang dapat menghasilkan prediksi dengan nilai kesalahan terkecil. Gambar 4.8 menunjukkan hasil perbandingan nilai akurasi dari beberapa metode pemilihan fitur berdasarkan model prediksi pohon keputusan.Reduksi data dengan pemilihan

fitur biasanya meningkatkan performa model prediksi, karena fitur yang tidak relevan terhadap target klasifikasi telah berkurang. Hasil uji coba menunjukkan bahwa tidak selalu terjadi kenaikan tingkat akurasi, tetapi menurunkan tingkat akurasi. Hal ini juga terjadi pada penelitian Dinakaran dan Thangaiah (2013) yang menerapkan metode pemilihan fitur pada model prediksi, namun menurunkan tingkat akurasi dari pohon keputusan.Berdasarkan gambar 4.8 dapat dijelaskan bahwa dengan menggunakan metode pemilihan fitur berdasarkan *dynamic* mutual informasi dapat memberikan nilai akurasi yang lebih baik dengan menggunakan lima fitur terpilih bila dibandingkan dengan melakukan pemilihan fitur berdasarkan p-*value*, perkiraan peneliti dan menggunakan seluruh fitur sebelum dilakukan pemilihan fitur.




Gambar 4.8 Hasil Perbandingan Akurasi Pohon Keputusan

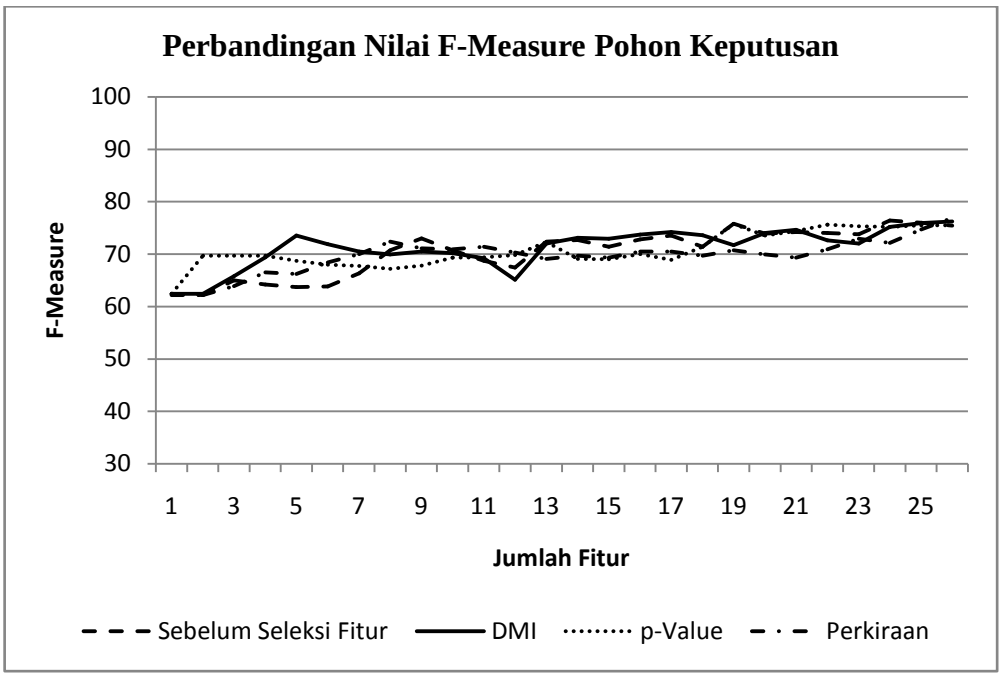
Perbedaan nilai akurasi juga terlihat pada penggunaan seluruh fitur berdasarkan masing-masing metode seleksi fitur. Hal ini dikarenakan metode pohon keputusan menggunakan struktur hierarki untuk pembelajaran *supervised*(Ariadni & Arieshanti, 2010). Proses dari algoritma pohon keputusan dimulai dari *root node* hingga *leaf node* yang dilakukan secara rekursif. Pohon dibangun dengan cara membagi data secara rekursif hingga tiap bagian terdiri dari

data yang berasal dari kelas yang sama. Bentuk pemecahan (*split*) yang digunakan untuk membagi data tergantung dari jenis atribut yang digunakan dalam *split.*Dalam melakukan pemisahan obyek (*split*) dilakukan tes terlebih dahulu terhadap atribut dengan mengukur tingkat ketidakmurnian pada sebuah simpul (*node*). Pada algoritma C.45 menggunakan rasio perolehan (*gain ratio*). Sebelum menghitung rasio perolehan, perlu menghitung dulu nilai informasi dalam satuan bits dari suatu kumpulan objek dengan menggunakan konsep entropi. Hal tersebut juga mengakibatkan perbedaan jumlah *leaf* dan ukuran *tree* yang dihasilkan dari klasifikasi pohon keputusan dengan menggunakan seluruh fitur, baik sebelum dilakukan seleksi fitur maupun setelah dilakukan perangkingan fitur menggunakan beberapa metode. Perbedaan tersebut dapat dilihat pada tabel 4.10.

Tabel 4.10 Perbedaan jumlah *leaf* dan ukuran *tree* klasifikasi Pohon Keputusan

| **Metode Seleksi Fitur** | **Jumlah *Leaf*** | **Ukuran *Tree*** |
|---|---|---|
| Sebelum Seleksi Fitur | 43 | 85 |
| *Dynamic* Mutual Informasi | 43 | 85 |
| *P-Value* | 45 | 89 |
| Perkiraan | 45 | 89 |

Sedangkan gambar 4.9 menunjukkan hasil perbandingan *f-measure* dari beberapa metode pemilihan fitur berdasarkan model prediksi pohon keputusan.




Gambar 4.9 Perbandingan *f-measure* Pohon Keputusan

Berdasarkan gambar 4.9 dapat diketahui bahwa perbedaan nilai *f-measure* yang signifikan terjadi pada penggunaanlima fitur dengan menggunakan metode *dynamic* mutual informasi.Nilai performa klasifikasi metodepohon keputusantertinggi adalah dengan menerapkan pemilihan fitur *dynamic*mutual informasi pada lima fitur yang terpilih, yaitu sebesar 76.68% untuk nilai akurasi dan 73.5% untuk nilai *f-measure*. Hasil perbedaan nilai performa klasifikasi dari masing-masing metode pemilihan fitur dapat dilihat pada lampiran B. Berdasarkan lampiran B, apabila dibandingkan nilai performa sebelum dan setelah teknik pemilihan fitur menggunakan *dynamic* mutual informasi, *p-value* dan pemilihan fitur berdasarkan perkiraan peneliti diimplementasikan, maka didapatkan informasi bahwa pada penerapan lima fitur yang terpilih menggunakan teknik pemilihan fitur *dynamic* mutual informasi memiliki nilai akurasi, presisi, *recall* dan *f-measure* yang lebih tinggi bila dibandingkan dengan penggunaan fitur sebelum diterapkan metode pemilihan fitur, metode pemilihan fitur *p-value* dan fitur yang dipilih berdasarkan perkiraan peneliti.

### 4.3.4. Uji Coba Representasi Klasifikasi dalam Rule IF-THEN

Setelah melalui tahapan transformasi fitur, pemilihan fitur, klasifikasi loyalitas pelanggan dan perbandingan performa klasifikasi, didapatkan sebuah pohon keputusan yang memiliki model terbaik dari proses klasifikasidalam bentuk model prediksi maupun dalam bentuk aturan/ *rule* IF-THEN. Berdasarkan syarat-syarat model pohon keputusan terbaik seperti yang telah disebutkan pada bab metodologi penelitian, maka model terbaik dari hasil klasifikasi dan perbandingan performa klasifikasi yaitu model yang menerapkan metode pemilihan fitur *dynamic* mutual informasi dengan jumlah fitur sebanyak lima. Dari hasil pengolahan dan uji coba menggunakan pohon keputusan pada dataset dihasilkan penyusunan informasi dalam bentuk *tree* seperti yang ditunjukkanGambar 4.10. Berdasarkan gambar 4.10 diketahui bahwa rata-rata konsumsi merupakaan *root* dari *tree*. Dari 386 data dengan lima fitur terpilih, 296 data (76.68%) dapat diklasifikasikan dengan benar, sedangkan 90 data (23.32%) salah diklasifikasikan. Hal ini ditunjukkan pada tabel 4.11.

Tabel 4.11*Confusion Matrix*pohon keputusandengan *10-fold cross validation*

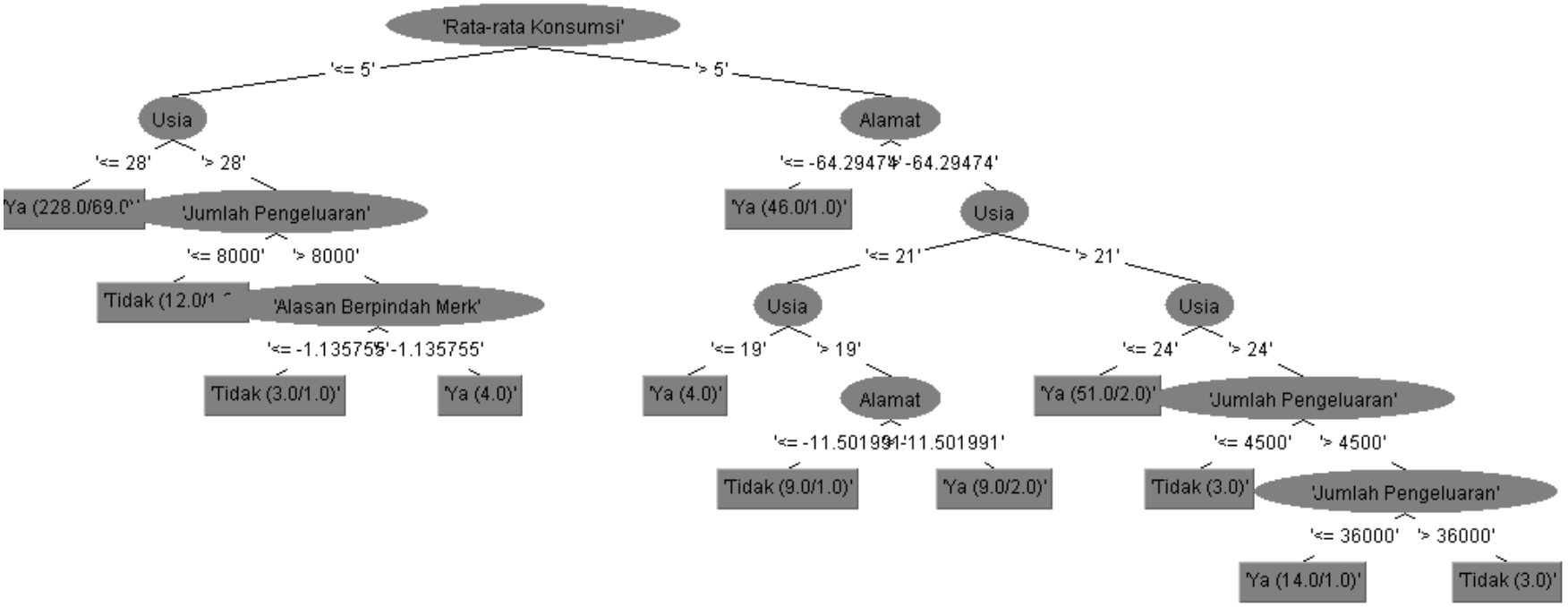
| Hasil Pengujian | Loyalitas Pelanggan | |
|---|---|---|
| | Loyal | Tidak Loyal |
| Positif | 266 | 18 |
| Negatif | 72 | 30 |

Dari tabel 4.11di atas dapat dijelaskan bahwa jumlah data pengujian untuk pelanggan yang diduga loyal adalah sebanyak 284, dimana 266 pelanggan *true-positive/ TP)* terprediksi dengan benar bahwa pelanggan akan loyal, sedangkan 18 pelanggan salah diprediksi (*false-positive/ FP)* oleh pengklasifikasi pohon keputusanC4.5, dimana sebenarnya pelanggan tersebut tidak loyal. Pengujian pada pelanggan yang diduga tidak loyal menunjukkan 30 pelanggan (*true-negatif/ TN*) dikenali dengan benar bahwa pelanggan tidak loyal, sebaliknya terdapat 72 pelanggan (*false-negatif/ FN*) salah diprediksi sebagai pelanggan yang tidak loyal.

Berdasarkan hasil klasifikasi pohon keputusan yang memiliki tingkat akurasi tertinggi, dapat diketahuifaktor-faktor relevan yang mempengaruhi performa klasifikasi pohon keputusanloyalitas pelanggan dengan mengimplementasikan transformasi fitur dan metode pemilihan fitur *dynamic* mutual informasi. Faktor-faktor tersebut antara lain rata-rata konsumsi, usia, jumlah pengeluaran, alasan berpindah merek dan alamat.

Selain dalam bentuk gambar, hasil model prediksi dari pohon keputusanjuga dapat disajkan dalam bentuk aturan/*rule* IF-THEN. Bentuk aturan IF-THEN ini merupakan hasil pendeskripsian dari bentuk gambar yang dihasilkan oleh pengklasifikasi pohon keputusan. Bentuk aturan ini mempermudah bagi para pengambil keputusan dalam membaca hasil prediksi. Berikut adalah contoh beberapa aturan/*rule* yang dihasilkan dari model prediksi loyalitas pelanggan menggunakan pohon keputusan:

JIKA rata-rata konsumsi <= 5 DAN usia <= 28 MAKA 'Loyal'
JIKA rata-rata konsumsi <= 5 DAN usia > 28 DAN jumlah pengeluaran <= 8000 MAKA 'Tidak Loyal'
JIKA rata-rata konsumsi <= 5 DAN usia > 28 DAN jumlah pengeluaran > 8000 DAN alasan berpindah <= -1.135755/ kualitas rasa lebih enak, lebih banyak varian rasa, promosi (beli satu gratis satu/ potongan harga/ dsb), produk lebih mudah didapat, harga lebih murah, kemasan yang lebih menarik MAKA 'Tidak Loyal'
JIKA rata-rata konsumsi <= 5 DAN usia > 28 DAN jumlah pengeluaran > 8000 DAN alasan berpindah > -1.135755/ coba – coba MAKA 'Loyal'
JIKA rata-rata konsumsi > 5 DAN alamat <= -64.29474/ Pesawaran, Lampung Utara, Mesuji, Pringsewu, Tulang Bawang, Tulang Bawang Barat, Way Kanan, Metro, Tanggamus, Pesisir Barat MAKA 'Loyal'
JIKA rata-rata konsumsi > 5 DAN alamat > -64.29474/ Lampung Timur, Bandar Lampung, Lampung Tengah, Lampung Selatan, Lampung Barat DAN usia <= 21 MAKA 'Loyal'
JIKA rata-rata konsumsi > 5 DAN alamat > -64.29474/ Lampung Timur, Bandar Lampung, Lampung Tengah, Lampung Selatan, Lampung Barat DAN usia > 21 DAN jumlah pengeluaan <= 4500 MAKA 'Tidak Loyal'




Gambar 4.10 Hasil Klasifikasi Pohon Keputusan

**LAMPIRAN A**

**KUISONER PENELITIAN**
**PEMILIHAN FITUR UNTUK KLASIFIKASI LOYALITAS PELANGGAN**
***FAST MOVING CONSUMER GOODS***

Petunjuk pengisian:

a. Bacalah terlebih dahulu pertanyaan sebelum anda menjawab dan jawablah semua pertanyaan tersebut dengan sejujurnya.
b. Pilihlah salah satu jawaban dan berilah tanda ( √ ) pada kolom jawaban yang telah disediakan.

**Bagian I : Tahap *Screening***

1. Apakah anda tahu tentang produk mie instan?
   ☐ Ya ☐ Tidak
2. Apakah anda pernah membeli dan mengkonsumsi produk mie instan?
   ☐ Pernah ☐ Tidak Pernah

**Bagian II : Identifikasi Pelanggan**

1. Nama Lengkap :
2. Jenis Kelamin : ☐ Pria ☐ Wanita
3. Usia : Tahun
4. Alamat (Kabupaten/kota) : _____
5. Status Perkawinan : ☐ Menikah ☐ Belum Menikah
6. Status Tinggal : ☐ Kos
   ☐ Ikut Orang Tua
   ☐ Rumah Sendiri
   ☐ .....(lainnya)
7. Pekerjaan : ☐ Pelajar / Mahasiswa

PNS ☐

Pegawai ☐MN

☐ Pegawai Swasta / Wirasawasta

☐ .....(lainnya)

8. Pendidikan terakhir saat ini:
   - ☐ SD, SMP, SMA
   - ☐ Diploma
   - ☐ S1
   - ☐ S2
9. Merek mie instan apakah yang paling sering anda konsumsi?
   - ☐ Indomie
   - ☐ Mie Sedaap
   - ☐ Supermie
   - ☐ Sarimie
   - ☐ .....(lainnya)
10. Dimanakah anda biasa membeli mie instan?
   - ☐ Warung
   - ☐ Pasar
   - ☐ Toko Serba Ada (Indomaret, Alfamart, dsb)
   - ☐ Supermarket
   - ☐ .....(lainnya)
11. Darimanakah anda mengetahui tentang produk mie instan?
   - ☐ Koran
   - ☐ Majalah
   - ☐ Televisi
   - ☐ Radio
   - ☐ Teman / Keluarga
   - ☐ .....(lainnya)
12. Ketertarikan anda terhadap produk mie instan karena?
   - ☐ Rasa
   - ☐ Kemasan Produk
   - ☐ Promosi (Diskon, Beli satu gratis satu, dll)
   - ☐ Mudah Didapat
   - ☐ Harga yang terjangkau
   - ☐ Praktis
   - ☐ Kandungan Gizi
   - ☐ .....(lainnya)

13. Sudah berapa lama anda mengkonsumsi produk mie instan?
    - ☐ ≤ 10 tahun
    - ☐ 10 – 20 tahun
    - ☐ 20 – 30 tahun
    - ☐ 30 tahun
14. Jarak pembelian pertama dengan berikutnya
    - ☐ 1 minggu
    - ☐ ≥ 1 – 2 minggu
    - ☐ ≥ 3 – 4 minggu
    - ☐ ≥ 1 bulan
15. Berapakah jumlah mie instan yang biasa anda beli dalam satu kali transaksi?
    - ☐ 1 – 5 bungkus
    - ☐ 10 bungkus
    - ☐ – 15 bungkus
    - ☐ – 20 bungkus
    - ☐ ≥ 1 dus
16. Rata – rata, berapa kali anda mengonsumsi mie instan dalam satu bulan?

    _____
17. Apakah tampilan produk dari mie instan yang biasa anda konsumsi menarik perhatian anda?
    - ☐ Ya
    - ☐ Tidak
18. Apakah anda merasa puas dengan harga yang ditawarkan oleh merek instan yang biasa anda konsumsi?
    - ☐ Sangat Puas
    - ☐ Puas
    - ☐ Kurang Puas
    - ☐ Tidak Puas
    - ☐ Sangat Tidak Puas

19. Apakah anda merasa puas dengan kualitas merek instan yang biasa anda konsumsi?
    - ☐ Sangat Puas
    - ☐ Puas
    - ☐ Kurang Puas
    - ☐ Tidak Puas
    - ☐ Sangat Tidak Puas
20. Apakah anda akan merekomendasikan kepada orang lain untuk mengonsumsi mie instan dengan merek yang biasa anda konsumsi?
    - ☐ Ya
    - ☐ Tidak
21. Apakah anda akan berkomentar positif atau memberikan pujian tentang produk mie yang biasa anda konsumsi kepada orang lain?
    - ☐ Ya
    - ☐ Tidak
22. Berapakah jumlah uang yang anda habiskan untuk membeli produk mie instan dalam satu kali transaksi?

    _____
23. Berapa bungkus mie instan yang anda konsumsi dalam satu kali konsumsi?
    - ☐ 1 bungkus
    - ☐ 2 bungkus
    - ☐ 3 bungkus
    - ☐ 4 bungkus
    - ☐ 5 bungkus
24. Apakah anda pernah mencoba mie instan merek lain (selain yang biasa anda konsumsi)?
    - ☐ Pernah
    - ☐ Tidak Pernah
25. Jika Pernah, apa yang menyebabkan anda berpindah atau menggunakan mie instan merek lain?
    - ☐ Harga lebih murah
    - ☐ Kualitas rasa lebih enak

- ☐ Promosi (Beli satu gratis satu, potongan harga, dsb)
- ☐ Produk lebih mudah didapat
- ☐ Kemasan yang lebih menarik
- ☐ Lebih banyak varian rasa
- ☐ Coba – Coba
- ☐ .....(lainnya)

26. Jika pernah, rata – rata berapa kali anda mengonsumsi mie instan merek lain dalalm rentan waktu satu bulan?

    _____

27. Setelah anda mie instan merek lain, apakah anda akan kembali pada merek mie instan yang biasa anda konsumsi?
    - ☐ Ya
    - ☐ Tidak
28. Apakah anda akan tetap setia mengonsumsi mie instan dengan merek yang biasa anda konsumsi?
    - ☐ Ya
    - ☐ Tidak

.....,.....

(.....)

*(lembar ini sengaja dikosongkan)*

## LAMPIRAN B

Lampiran ini mencakup hasil pengukuran performa klasifikasi yang diolah menggunakan *software* Weka secara detail. Nilai hasil pengukuran yang disajikan antara lain berupa nilaiakurasi, *precision, recall,* dan *f-measure.*

### B.1 Hasil Klasifikasi Sebelum Seleksi Fitur

Hasil uji coba membangun model prediksi klasifikasi dengan algoritma Pohon keputusan menggunakan seluruh fitur sebelum diterapkannya metode pemilihan fitur akan disajikan sebagai berikut:

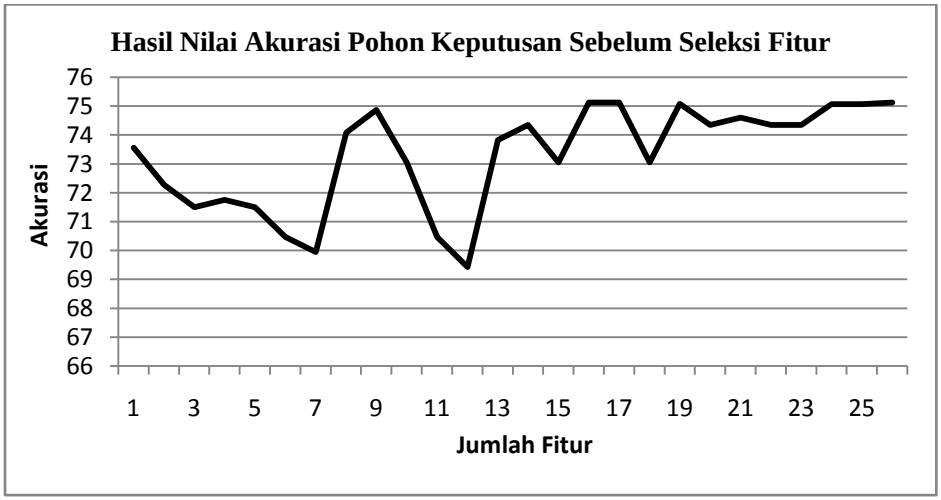
| Fitur | Jumlah Fitur | Akurasi (%) | Presisi (%) | Recall (%) | F-Measure (%) |
|---|---|---|---|---|---|
| f1 | 1 | 73.57 | 54.1 | 73.6 | 62.4 |
| f1, f2 | 2 | 72.28 | 57.7 | 72.3 | 62.2 |
| f1, f2, f3 | 3 | 71.5 | 64.1 | 71.5 | 65 |
| f1, f2, f3, f4 | 4 | 71.76 | 63.2 | 71.8 | 64.2 |
| f1, f2, f3, f4, f5 | 5 | 71.5 | 62.2 | 71.5 | 63.7 |
| f1, f2, f3, f4, f5, f6 | 6 | 70.47 | 61.8 | 70.5 | 63.8 |
| f1, f2, f3, f4, f5, f6, f7 | 7 | 69.95 | 64.9 | 69.9 | 66.3 |
| f1, f2, f3, f4, f5, f6, f7, f8 | 8 | 74.09 | 70.7 | 74.1 | 70.7 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9 | 9 | 74.87 | 72.6 | 74.9 | 73 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10 | 10 | 73.06 | 70 | 73.1 | 70.7 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11 | 11 | 70.47 | 67.7 | 70.5 | 68.7 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12 | 12 | 69.43 | 66.3 | 69.4 | 67.4 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13 | 13 | 73.83 | 71.8 | 73.8 | 72.4 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14 | 14 | 74.35 | 72.2 | 74.4 | 72.7 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15 | 15 | 73.06 | 70.7 | 73.1 | 71.4 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16 | 16 | 75.13 | 72.5 | 75.1 | 72.8 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f17 | 17 | 75.13 | 73 | 75.1 | 73.5 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f17, f18 | 18 | 73.06 | 70.7 | 73.1 | 71.4 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f18, f19 | 19 | 76.42 | 75.4 | 76.4 | 75.8 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, | 20 | 74.35 | 73.5 | 74.4 | 73.9 |

| Fitur | Jumlah Fitur | Akurasi (%) | Presisi (%) | Recall (%) | F-Measure (%) |
|---|---|---|---|---|---|
| f11, f12, f13, f14, f15, f16, f18, f19, f20 | | | | | |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f18, f19, f20, f21 | 21 | 74.61 | 73.8 | 74.6 | 74.2 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f18, f19, f20, f21, f22 | 22 | 74.35 | 73.7 | 74.4 | 74 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f18, f19, f20, f21, f22, f23 | 23 | 74.35 | 73.4 | 74.4 | 73.8 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f18, f19, f20, f21, f22, f23, f24 | 24 | 76.68 | 76.1 | 76.7 | 76.4 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f18, f19, f20, f21, f22, f23, f24, f25 | 25 | 76.42 | 75.7 | 76.4 | 76 |
| f1, f2, f3, f4, f5, f6, f7, f8, f9, f10, f11, f12, f13, f14, f15, f16, f18, f19, f20, f21, f22, f23, f24, f25, f26 | 26 | 75.13 | 75.8 | 75.1 | 75.4 |

Tabel *confussion matrix*

| Hasil Pengujian | Loyalitas Pelanggan | |
|---|---|---|
| | Loyal | Tidak Loyal |
| Positif | 232 | 52 |
| Negatif | 44 | 58 |

Grafik hasil akurasi klasifikasi pohon keputusan sebelum menerapkan metode seleksi fitur:



## B.2 Hasil Klasifikasi Menggunakan Metode Pemilihan Fitur *Dynamic* Mutual Informasi

Hasil uji coba membangun model prediksi klasifikasi dengan algoritma pohon keputusan dan menggunakan metode pemilihan fitur *dynamic* mutual informasiakan disajikan sebagai berikut:

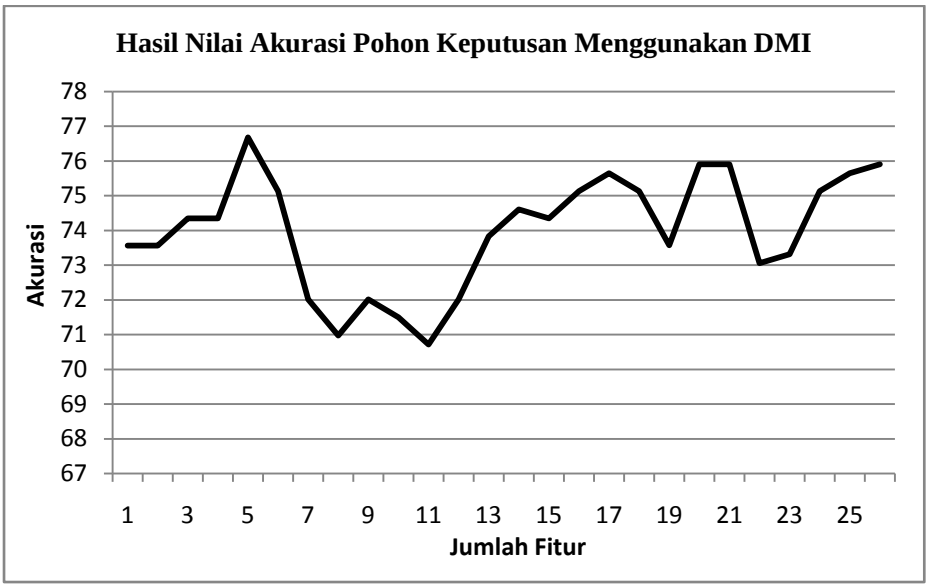
| Fitur | Jumlah Fitur | Akurasi (%) | Presisi (%) | Recall (%) | F-Measure (%) |
|---|---|---|---|---|---|
| f21 | 1 | 73.57 | 54.1 | 73.6 | 62.4 |
| f21, f15 | 2 | 73.57 | 54.1 | 73.6 | 62.4 |
| f21, f15, f2 | 3 | 74.35 | 71.8 | 74.4 | 65.8 |
| f21, f15, f2, f3 | 4 | 74.35 | 70.5 | 74.4 | 69.4 |
| f21, f15, f2, f3, f24 | 5 | 76.68 | 74.4 | 76.7 | 73.5 |
| f21, f15, f2, f3, f24, f25 | 6 | 75.13 | 72.2 | 75.1 | 71.9 |
| f21, f15, f2, f3, f24, f25, f11 | 7 | 72.02 | 69.7 | 72 | 70.5 |
| f21, f15, f2, f3, f24, f25, f11, f7 | 8 | 70.98 | 69.2 | 71 | 69.9 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6 | 9 | 72.02 | 69.7 | 72 | 70.5 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8 | 10 | 71.5 | 69.4 | 71.5 | 70.2 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14 | 11 | 70.72 | 68.3 | 70.7 | 69.2 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13 | 12 | 72.02 | 70.3 | 71 | 65.1 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12 | 13 | 73.83 | 71.4 | 73.8 | 72 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10 | 14 | 74.61 | 72.5 | 74.6 | 73.1 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17 | 15 | 74.35 | 72.3 | 74.4 | 72.9 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22 | 16 | 75.13 | 73.2 | 75.1 | 73.7 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18 | 17 | 75.65 | 73.7 | 75.6 | 74.2 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18, f5 | 18 | 75.13 | 73.1 | 75.1 | 73.6 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18, f5, f9 | 19 | 73.58 | 71.1 | 73.6 | 71.7 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18, f5, f9, f23 | 20 | 75.91 | 73.7 | 75.9 | 74 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18, f5, f9, f23, f4 | 21 | 75.91 | 74.2 | 75.9 | 74.6 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, | 22 | 73.06 | 72.2 | 73.1 | 72.6 |

| Fitur | Jumlah Fitur | Akurasi (%) | Presisi (%) | Recall (%) | F-Measure (%) |
|---|---|---|---|---|---|
| f14, f13, f12, f10, f17, f22, f18, f5, f9, f23, f4, f26 | | | | | |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18, f5, f9, f23, f4, f26, f19 | 23 | 73.32 | 71.1 | 73.3 | 72 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18, f5, f9, f23, f4, f26, f19, f20 | 24 | 75.13 | 75.3 | 75.1 | 75.2 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18, f5, f9, f23, f4, f26, f19, f20, f1 | 25 | 75.65 | 76.1 | 75.6 | 75.9 |
| f21, f15, f2, f3, f24, f25, f11, f7, f6, f8, f14, f13, f12, f10, f17, f22, f18, f5, f9, f23, f4, f26, f19, f20, f1, f23 | 26 | 75.91 | 76.5 | 75.9 | 76.2 |

Tabel *confussion matrix*

| Hasil Pengujian | Loyalitas Pelanggan | |
|---|---|---|
| | Loyal | Tidak Loyal |
| Positif | 234 | 50 |
| Negatif | 43 | 59 |

Grafik hasil akurasi klasifikasi pohon keputusan menggunakan metode seleksi fitur *Dynamic* Mutual Informasi (DMI):



### B.3 Hasil Klasifikasi Menggunakan Metode Pemilihan Fitur *P-Value*

Hasil uji coba membangun model prediksi klasifikasi dengan algoritma pohon keputusan dan menggunakan metode pemilihan fitur *p-value* akan disajikan sebagai berikut:

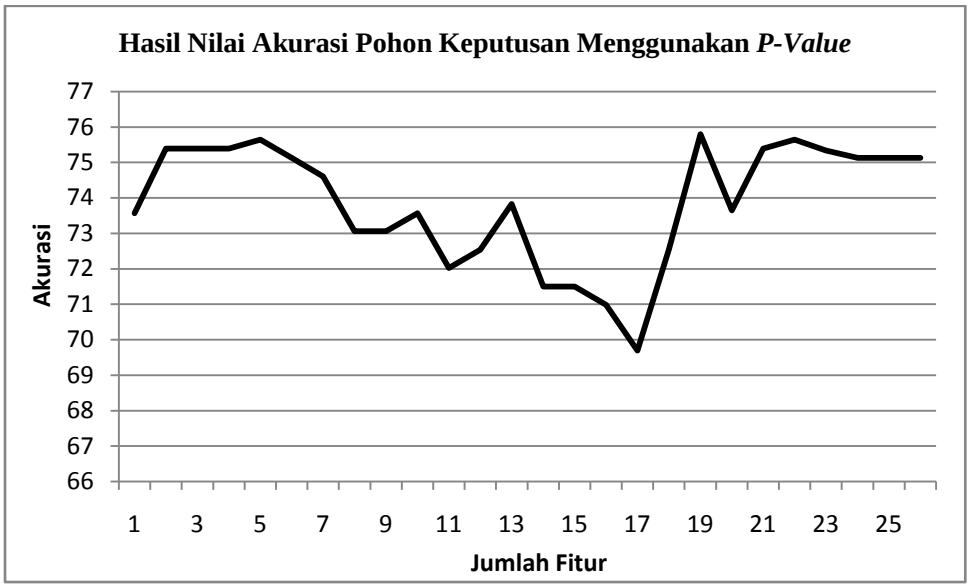
| Fitur | Jumlah Fitur | Akurasi (%) | Presisi (%) | Recall (%) | F-Measure (%) |
|---|---|---|---|---|---|
| f6 | 1 | 73.57 | 54.1 | 73.6 | 62.4 |
| f6, f26 | 2 | 75.39 | 72.8 | 75.4 | 69.7 |
| f6, f26, f17 | 3 | 75.39 | 72.8 | 75.4 | 69.7 |
| f6, f26, f17, f10 | 4 | 75.39 | 72.8 | 75.4 | 69.7 |
| f6, f26, f17, f10, f1 | 5 | 75.65 | 74.9 | 75.6 | 68.7 |
| f6, f26, f17, f10, f1, f8 | 6 | 75.13 | 73.2 | 75.1 | 68 |
| f6, f26, f17, f10, f1, f8, f7 | 7 | 74.61 | 71.3 | 74.6 | 67.7 |
| f6, f26, f17, f10, f1, f8, f7, f24 | 8 | 73.06 | 67.7 | 73.1 | 67.2 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16 | 9 | 73.06 | 68 | 73.1 | 67.8 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12 | 10 | 73.57 | 69.5 | 73.6 | 69.3 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11 | 11 | 72.02 | 68.5 | 72 | 69.4 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14 | 12 | 72.54 | 69.1 | 72.5 | 69.8 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3 | 13 | 73.83 | 71.7 | 73.8 | 72.3 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18 | 14 | 71.5 | 68.2 | 71.5 | 69.1 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5 | 15 | 71.5 | 68.1 | 71.5 | 69 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13 | 16 | 70.98 | 69.4 | 71 | 70 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25 | 17 | 69.69 | 68.3 | 69.7 | 68.9 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21 | 18 | 72.54 | 70.7 | 72.5 | 71.4 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19 | 19 | 75.8 | 76.5 | 77.7 | 75.7 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9 | 20 | 73.65 | 76.4 | 77.5 | 73.54 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9, f15 | 21 | 75.39 | 73.9 | 75.4 | 74.4 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, | 22 | 75.65 | 75.5 | 75.6 | 75.6 |

| Fitur | Jumlah Fitur | Akurasi (%) | Presisi (%) | Recall (%) | F-Measure (%) |
|---|---|---|---|---|---|
| f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9, f15, f20 | | | | | |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9, f15, f20, f4 | 23 | 75.34 | 75.3 | 75.4 | 75.3 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9, f15, f20, f4, f22 | 24 | 75.13 | 75.3 | 75.1 | 75.2 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9, f15, f20, f4, f22, f2 | 25 | 75.13 | 75.9 | 75.1 | 75.5 |
| f6, f26, f17, f10, f1, f8, f7, f24, f16, f12, f11, f14, f3, f18, f5, f13, f25, f21, f19, f9, f15, f20, f4, f22, f2, f23 | 26 | 75.13 | 75.9 | 75.1 | 75.5 |

Tabel *confussion matrix*

| Hasil Pengujian | Loyalitas Pelanggan | |
|---|---|---|
| | Loyal | Tidak Loyal |
| Positif | 231 | 53 |
| Negatif | 43 | 59 |

Grafik hasil akurasi klasifikasi pohon keputusan menggunakan metode seleksi fitur *p-value*:



### B.4 Hasil Klasifikasi Menggunakan Fitur Berdasarkan Perkiraan Peneliti

Hasil uji coba membangun model prediksi klasifikasi dengan algoritma pohon keputusan dan menggunakan metode pemilihan fitur berdasarkan perkiraan peneliti akan disajikan sebagai berikut:

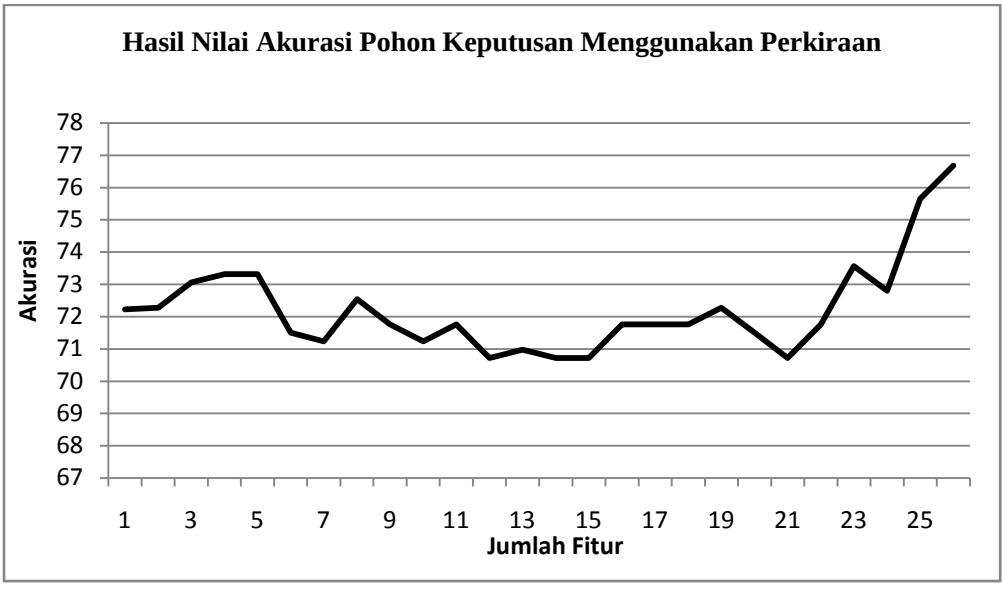
| Fitur | Jumlah Fitur | Akurasi (%) | Presisi (%) | Recall (%) | F-Measure (%) |
|---|---|---|---|---|---|
| f2 | 1 | 72.23 | 57.7 | 72.3 | 62.2 |
| f2, f25 | 2 | 72.28 | 57.7 | 72.3 | 62.2 |
| f2, f25, f3 | 3 | 73.06 | 65 | 73.1 | 63.9 |
| f2, f25, f3, f4 | 4 | 73.32 | 67.8 | 73.3 | 66.5 |
| f2, f25, f3, f4, f6 | 5 | 73.32 | 67.6 | 73.3 | 66.2 |
| f2, f25, f3, f4, f6, f15 | 6 | 71.5 | 67.5 | 71.5 | 68.4 |
| f2, f25, f3, f4, f6, f15, f21 | 7 | 71.24 | 69.2 | 71.2 | 70 |
| f2, f25, f3, f4, f6, f15, f21, f7 | 8 | 72.54 | 72.2 | 72.5 | 72.4 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8 | 9 | 71.76 | 70.6 | 71.8 | 71.1 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10 | 10 | 71.24 | 70.6 | 71.2 | 70.9 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11 | 11 | 71.76 | 71.1 | 71.8 | 71.4 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18 | 12 | 70.72 | 69.9 | 70.7 | 70.3 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12 | 13 | 70.98 | 68.1 | 71 | 69.1 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13 | 14 | 70.72 | 69 | 70.7 | 69.7 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24 | 15 | 70.72 | 68.5 | 70.7 | 69.3 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1 | 16 | 71.76 | 69.8 | 71.8 | 70.5 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23 | 17 | 71.76 | 69.8 | 71.8 | 70.5 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26 | 18 | 71.76 | 68.8 | 71.8 | 69.7 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26, f5 | 19 | 72.28 | 69.9 | 72.3 | 70.7 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26, f5, f16 | 20 | 71.5 | 69.1 | 71.5 | 69.9 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26, f5, f16, f9 | 21 | 70.72 | 68.5 | 70.7 | 69.3 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26, f5, f16, f9, f17 | 22 | 71.76 | 70.3 | 71.8 | 70.9 |

| Fitur | Jumlah Fitur | Akurasi (%) | Presisi (%) | Recall (%) | F-Measure (%) |
|---|---|---|---|---|---|
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26, f5, f16, f9, f17, f14 | 23 | 73.57 | 72.5 | 73.6 | 72.9 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26, f5, f16, f9, f17, f14, f22 | 24 | 72.8 | 71.6 | 72.8 | 72.1 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26, f5, f16, f9, f17, f14, f22, f19 | 25 | 75.65 | 74.3 | 75.6 | 74.7 |
| f2, f25, f3, f4, f6, f15, f21, f7, f8, f10, f11, f18,f12, f13, f24, f1, f23, f26, f5, f16, f9, f17, f14, f22, f19, f20 | 26 | 76.68 | 77.1 | 76.7 | 76.9 |

Tabel *confussion matrix*

| Hasil Pengujian | Loyalitas Pelanggan | |
|---|---|---|
| | Loyal | Tidak Loyal |
| Positif | 236 | 48 |
| Negatif | 42 | 60 |

Grafik hasil akurasi klasifikasi pohon keputusan menggunakan fitur hasil perkiraan peneliti:



## LAMPIRAN C

Lampiran ini mencakup *source code* transformasi fitur dan metode pemilihan fitur *dynamic* mutual informasi yang diolah menggunakan Matlab.

### C.1*Source Code* Transformasi Fitur

```
clear;clc;
data = readtable('datapelanggan.csv');
me=zeros(0,0);
transform   = zeros(0,0);
rowNum      = size(data,1);
colNum      = size(data,2);
for j=1:colNum-1
   if(iscellstr(data{1,j}))
n      = size(unique(data{:,j}),1);
cols   = data{:,j};
si     = unique(data{:,j});
      sigPij = 0;
sigPi3 = 0;
sigNKP = 0;
pj  = 1/colNum;
      for pIterasi=1:size(si,1)
 pi(pIterasi)   = size(find(strcmp(si{pIterasi}, cols)),1)/rowNum;
       end
 tmp    = zeros(0,0);
      for i=1:length(si)
sigPi3  = sigPi3+power(pi(i),3);
if (j~=i)
sigPij  = sigPij+(pi(i)*pj*power((i-j),2));
 else
            sigPij  = sigPij+(pi(i)*pj);
end
       for k=1:i
sigNKP = sigNKP+((n-k)*pi(k));
 end
myui    = ((n-i)-sigNKP)*sqrt((1-sigPi3)/sigPij);
chg     = normrnd(myui,pi(i));
replace = find(strcmp(si{i}, cols));
for z=1:size(replace,1)
tmp{replace(z),1} = chg;
end
end
transform  = cat(2,transform,tmp);

else
transform  = cat(2,transform,table2cell(data(:,j)));
```

```
end
end
transform  = cat(2,transform,table2cell(data(:,size(data,2))));
% for norm=1:size(transform,2)-1
%    ma = max([transform{:,norm}]);
%    mi = min([transform{:,norm}]);
%    for data=1:size(transform,1)
%  transform{data,norm}= (ma-transform{data,norm})/(ma-mi);
%    end
% end
%dataset
%end
```

## C.2 *Source Code* Menghitung Nilai Mutual Informasi

```
function z = mutualInformation(x, y)
% Compute mutual information I(x,y) of two discrete variables x and y.
% Written by Mo Chen (mochen80@gmail.com).

   assert(numel(x) == numel(y));
   n = numel(x);
   x = reshape(x,1,n);
   y = reshape(y,1,n);

   l = min(min(x),min(y));
   x = x-l+1;
   y = y-l+1;
   k = max(max(x),max(y));

   idx = 1:n;
%     Mx = sparse(idx,1,x,n,k,n);
   Mx = sparse(idx, x, 1,n,k,n);
   My = sparse(idx, y, 1,n,k,n);
   Pxy = nonzeros(Mx'*My/n); %joint distribution of x and y
   Hxy = -dot(Pxy,log2(Pxy+eps));

   Px = mean(Mx,1);
   Py = mean(My,1);

   % entropy of Py and Px
   Hx = -dot(Px,log2(Px+eps));
   Hy = -dot(Py,log2(Py+eps));

   % mutual information
   z = Hx + Hy - Hxy;
end
```

## C.3 *Source Code* Menghitung Nilai Entropi

```
function z = entropy(x)
% Compute entropy H(x) of a discrete variable x.
% Written by Mo Chen (mochen80@gmail.com).
   n = numel(x);
   x = reshape(x,1,n);
   [u,~,label] = unique(x);
   p = full(mean(sparse(1:n,label,1,n,numel(u),n),1));
   z = -dot(p,log2(p+eps));
end
```

## C.4 *Source Code* Menghitung *Conditional Entropy*

```
function z = conditionalEntropy (x, y)
% Compute conditional entropy H(x|y) of two discrete variables x and y.
% Written by Mo Chen (mochen80@gmail.com).
   assert(numel(x) == numel(y));
   n = numel(x);
   x = reshape(x,1,n);
   y = reshape(y,1,n);

   l = min(min(x),min(y));
   x = x-l+1;
   y = y-l+1;
   k = max(max(x),max(y));

   idx = 1:n;
   Mx = sparse(idx,x,1,n,k,n);
   My = sparse(idx,y,1,n,k,n);
   Pxy = nonzeros(Mx'*My/n); %joint distribution of x and y
   Hxy = -dot(Pxy,log2(Pxy+eps));

   Py = mean(My,1);
   Hy = -dot(Py,log2(Py+eps));

   % conditional entropy H(x|y)
   z = Hxy-Hy;
end
```

### C.5 *Source CodeDynamic* Mutual Informasi

```
clear;clc;
D = load('DataPelangganTransformasi.csv');
header = [1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26];
c_index = size(D,2);
FF = D(:,1:c_index-1);
C = D(:,c_index);
%Cari nilai MI Maximum
FF = round(FF);
for i=1:c_index-1
   Fi_c(i) = mutualInformation(abs(FF(:,i)), C);
end
[val,index]= sort(Fi_c,'descend');
index
dataset= D(:,index(1));
for j=2:size(index,2)
  dataset = [dataset D(:,index(i))];
end
dataset = cat(2,dataset, D(:,c_index));
```

### C.6 *Source CodeP-Value*

```
clear;clc;
D = load('DataPelangganTransformasi.csv');
c_index = size(D,2);
C = D(:,c_index);
F = D(:,1:c_index);
for i=1:size(F,2)-1
  [h,p,ci,stats] = ttest2(F(:,i),C);
%       result(i,2)=i;
%       result(i,1)=p;
result(i)=p;
end
```

```
result = result.';
[val,index] = sort(result,'descend');
dataset= D(:,index(1));
for j=2:size(index,1)
   dataset = [dataset D(:,index(i))];
end
dataset = cat(2,dataset, D(:,c_index));
```

# BAB 5
# KESIMPULAN DAN SARAN

Bab ini menjelaskan kesimpulan dari hasil dan pembahasan yang dilakukan sesuai dengan skenario uji coba. Selain itu, dalam bab ini juga dijelaskan saran kemungkinan pengembangan lebih lanjut dari penelitian yang telah dilakukan.

## 5.1. Kesimpulan

Berdasarkan hasil uji coba dan pembahasan, dapat disimpulkan sebagai berikut:

1. Sulitnya mengevaluasi fitur heterogen, seperti fitur numerik dan non-numerik secara bersamaan dapat diselesaikan dengan cara melakukan transformasi atau mengubah fitur yang heterogen tersebut menjadi fitur yang homogen. Transformasi fitur ini dilakukan dengan cara fitur-fitur dengan non-numerik diubah menjadi fitur numerik yang didasarkan pada aturan *gaussian*. Pengimplementasian metode pemilihan fitur *dynamic* mutual informasi dapat mempengaruhi tingkat performa algoritma klasifikasi pohon keputusan untuk memprediksi loyalitas pelanggan *fast moving consumer goods*. Hal tersebut dikarenakan fitur yang tidak relevan terhadap target klasifikasi telah berkurang. Metode pemilihan fitur tersebut didasarkan dengan nilai mutual informasi antara fitur dengan label kelas dengan cara melakukan perhitungan nilai entropi. Metode pemilihan fitur *dynamic* mutual informasi dengan memilih lima fitur pada rangking teratas menunjukkan hasil terbaik dalam penelitian ini.
2. Hasil fitur yang terpilih diklasifikasikan menggunakan metode pohon keputusan dengan *10-fold crossvalidation*. Dari hasil pengujian terhadap model prediksi klasifikasi pohon keputusandiperoleh faktor-faktor relevan yang mempengaruhi performa klasifikasi pohon keputusanloyalitas pelanggan. Peningkatan performa tersebut dapat dilihat pada pengimplementasian metode pemilihan fitur *dynamic* mutual informasi dan

penggunaan jumlah fitur sebanyak lima. Nilai akurasi, presisi, *recall* dan *f-measure* mengalami peningkatan bila dibandingkan dengan penggunaan seluruh fitur (sebelum dilakukan pemilihan fitur), metode pemilihan fitur *p-value*dan hasil perkiraan, masing-masing nilai tersebut secara berturut-turut adalah sebesar 76.68%, 74.4%, 76.7% dan 73.5%.

## 5.2. Saran

Saran-saran yang dapat diberikan berkaitan dengan hasil uji coba dan pembahasan pada penelitian ini adalah:

1. Perlu dilakukan penelitian lebih lanjut dengan menggunakan fitur-fitur yang lebih spesifik dan lengkap agar model yang diperoleh mampu mengidentifikasi faktor-faktor relevan yang mempengaruhi performa klasifikasi loyalitas pelanggan dapat lebih spesifik.
2. Pada penelitian ini menggunakan dataset loyalitas pelanggan yang diperoleh dari hasil penyebaran kuisoner dan memiliki ketidakseimbangan kelas. Hal tersebut mengakibatkan berkurangnya akurasi pada klasifikasi. Penelitian lebih lanjut diharapkan agar dapat menerapkan teknik-teknik yang digunakan untuk menangani ketidakseimbangan kelas tersebut, sehingga akurasi dari klasifikasi dapat lebih ditingkatkan.

# DAFTAR PUSTAKA


Abubakar. (2009). Pengukuran Persepsi Penumpang tentang Efektivitas Strategi Pencegahan Kejahatan TransJakarta. *Universitas Indonesia*.

Agrawal, M. (2003). Customer Relationship Management (CRM) & Corporate Renaissance. *Journal of Services Research*, 149 -17.

Aktepe, A., Ersoz, S., & Toklu, B. (2014). Customer Satisfaction and Loyalty Analysis with Classification Algorithms and Structural Equation Modeling. *Computers & Industrial Engineering*, 95-106.

AL-Nabi, D. L., & Ahmed, S. S. (2013). Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation). *Computer Engineering and Intelligent Systems*, 18 - 25.

Anderson, E. W., & Mittal, V. (2014). Strengthening the Satisfaction-Profit Chain. *Journal of Service Research*, 107-120.

Andriani, A. (2012). Penerapan Algoritma C4.5 Pada Program Klasifikasi Mahasiswa Dropout. *Seminar Nasional Matematika*, 139-147.

Ariadni, R., & Arieshanti, I. (2010). Implementasi Pohon Keputusan untuk Klasifikasi Data dengan Nilai Fitur yang Tidak Pasti. *Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember*.

Arifiyanti, A. A. (2015). Ekstraksi Fitur Pada Konten Jejaring Sosial Twitter Berbahasa Indonesia dalam Peningkatan Kinerja Klasifikasi Sentimen.

Ariwibowo, A. S. (2013). Metode Data Mining Untuk Klasifikasi Kesetiaan Pelanggan Terhadap Merek Produk. *Seminar Nasional Sistem Informasi Indonesia*.

Astuti, Y. A. (2011). *Analisis Perbandingan Teknik Support Vector Regression (SVR) dan Decision Tree C4.5 dalam Data Mining.* Medan: Universitas Sumatera Utara.

Athanassopoulos, A. D. (2000). Customer Satisfaction Cues To Support Market Segmentation and Explain Switching Behavior. *Journal of Business Research 47*, 191–207.

Badriyah, T., & Rahmawati, R. (2006). Alat Bantu Klasifikasi dengan Pohon Keputusan untuk Sistem Pendukung Keputusan. *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 1-4.

Baltzan, P., & Phillips, A. (2009). *Business Driven Information Systems Second Edition.* New York: Paul Ducham - Mc Graw Hill.

Behjati, S., Nahich, M., & Othaman, D. S. (2012). Interrelation between E-service Quality and E-satisfaction and Loyalty. *European Journal of Business and Management*, 75-86.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 245-271 .

Buckinx, W., & Poel, D. V. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, 252–268.

Chen, M.-S., Han, J., & Yu, P. S. (1996). Data Mining: An Overview from a Database Perspective. *IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6*, 866-883.

Cheng, C.-J., Chiu, S., Cheng, C.-B., & Wu, J.-Y. (2012). Customer lifetime value prediction by a Markov chain based datamining model: Application to an auto repair and maintenance company in Taiwan. *Transactions E: Industrial Engineering*, 849–855.

Chong, J. Y., & Wong, A. K. (1995). Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 641-651.

Chow, T. W., & Huang, D. (2005). Estimating Optimal Feature Subsets Using Efficient Estimation of High-Dimensional Mutual Information. *IEEE Transactions on Neural Networks*, 213-224.

Chu, B.-H., Tsai, M.-S., & Ho, C.-S. (2007). Toward a hybrid data mining model for customer retention. *Knowledge-Based Systems*, 703–718.

Dinakaran, & Thangaiah, R. J. (2013). Role of Attribute Selection in Classification Algorithm. *International Journal of Scientific & Engineering Research, Volume 4, Issue 6*, 67-71.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification Second Edition.* New York: Willey-Interscience.

Duygu, & KIRMACI, S. (2012). Customer Relationship Management and Customer Loyalty; a Survey in The Sector of Banking. *International Journal of Business and Social Science* .

Estevez, P. A., Tesmer, M., Perez, C. A., & Zurada, J. M. (2009). Normalized Mutual Information Feature Selection. *IEEE Transactions on Neural Networks, Vol. 20, No. 2*, 189-201.

Fayyad, U. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of The ACM, Vol. 39, No. 11*, 27-34.

Firat, R. (2009). Penerapan Teknik Klasifikasi Menggunakan Metode Fuzzy Decision Tree dengan Algoritma ID3 pada Data Diabetes. *Internetworking Indonesia Journal Vol. 1, No. 2* , 45.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research 3*, 1289-1305.

Griffin, J. (2005). *Customer Loyalty: How to Earn It, How to Keep It (Terjemahan).* Jakarta: Erlangga.

Griffin, J. (2002). *Customer Loyalty: How to Earn It, How to Keep It, New and Revised Edition.* USA.

Gronroos, C. (2009). Marketing as Promise Management: Regaining Customer Management. 351–359.

Gummesson, E. (2008). Customer Centricity: Reality or a Wild Goose Chase? 315–330.

Hall, M. A. (2000). Correlation-based Feature Selection for Machine Learning. *In Proceedings of the 17th Intl.Conf. Machine Learning*, 1-16.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* USA: Morgan Kaufmann Publishers.

Han, M., & Ren, W. (2015). Global Mutual Information-based Feature Selection Approach Using Single-Objective and Multi-Objective Optimization. *Neurocomputing*, 47-54.

Hartama, D. (2011). *Model Aturan Keterhubungan Data Mahasiswa Menggunakan Algoritma C4.5 untuk Meningkatkan Indeks Prestasi*. Medan: Universitas Sumatera Utara.

Hu, Q., Yu, D., Liu, J., & Wu, C. (2008). Neighborhood rough set based heterogeneous feature subset selection. *Information Science 178*, 3577–3594.

Huang, D., & Chow, T. W. (2005). Effective feature selection scheme using mutual information. *Neurocomputing*, 325–343.

Hung, S.-Y., Yen, D. C., & Wang, H.-Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 515–524.

Jones, M. A., Mothersbaugh, D. L., & Beatty, S. E. (2000). Switching Barriers and Repurchase Intentions in Services. *Journal of Retailing, Volume 76(2) pp. 259 –274*, 259-274.

Jutla, D., Craig, J., & Bodorik, P. (2001). Enabling and Measuring Electronic Customer Relationship Management Readiness. *Proceedings of the 34th Hawaii International Conference on System Sciences*, 1-10.

Kim, J., Suh, E., & Hwang, H. (2003). A Model for Evaluating the Effectiveness of CRM using the Balance Scorecard. *Journal of Interactive Marketing*, 5 - 19.

Kim, S.-Y., Jung, T.-S., Suh, E.-H., & Hwang, H.-S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert Systems with Applications*, 101–107.

Kim, Y. A., Song, S. H., & Kim, S. H. (2005). Strategies for preventing defection based on the mean time to defection and their implementations on a self-organizing map. *Expert Systems, Vol. 22, No. 5*, 265-278.

Kim, Y. (2006). Toward a successful CRM: variable selection sampling, and ensemble. *Decision Support Systems*, 542 – 553.

Kotler, P., & Keller, K. L. (2012). *Marketing Management 14E*. New Jersey: Prentice Hall.

Kurniawan, D. (2014, Oktober 16). *Kantar Worldpanel: Pertumbuhan Industri FMCG Indonesia Tertinggi di Asia*. Retrieved Oktober 31, 2015, from Gatra: http://www.gatra.com

Kwak, N., & Choi, C.-H. (2002a). Input Feature Selection by Mutual Information Based on Parzen Window. *IEEE Transactional Pattern Analitycal Matchematics Intelligence*, 1667-1671.

Kwak, N., & Choi, C.-H. (2002b). Input Feature Selection for Classification Problems. *IEEE Transactions on Neural Network, Vol. 13 No. 1*, 143-159.

Larivie`re, B., & Poel, D. V. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, 472–484.

Lee, T.-S., Chiu, C.-C., Chou, Y.-C., & Lu, C.-J. (2006). Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis* , 1113 – 1130.

Lemeshow, S., Hosmer Jr, D. W., Klar, J., & Lwanga, S. K. (1990). *Adequacy of Sample Size in Health Studies.* New York: John Wiley & Sons Ltd.

Li, B., Chow, T. W., & Tang, P. (2014). Analyzing rough set based attribute reductions by extension rule. *Neurocomputing 123*, 185-196.

Li, W. (1990). Mutual Information Functions Versus Correlation-Functions. *Journal of Statistical Physics Vol. 60*, 823-837.

Li, Y., Xie, M., & Goh, T. (2009). A Study of Mutual Information Based Feature Selection for Case Based Reasoning in Software Cost Estimation. *Expert System with Application*, 5921-5931.

Liu, H., & Setiono, R. (1997). Feature Selection via Discretization. *IEEE Transactions on Knowledge and Data Engineering, vol. 9, no. 4* , 642-645.

Liu, H., & Yu, L. (2005). Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 491 - 502.

Liu, H., Sun, J., Liu, L., & Zhang, H. (2009). Feature selection with dynamic mutual information. *Pattern Recognition 42*, 1330 -- 1339.

Maldonado, S., Flores, A., Verbraken, T., Baesens, B., & Weber, R. (2015). Profit-based feature selection using support vector machines General framework and an application for customer retention. *Applied Soft Computing 35*, 740-748.

Mandasari, V., & Tama, B. A. (2011). Analisis Kepuasan Konsumen Terhadap Restoran Cepat Saji Melalui Pendekatan Data Mining: Studi Kasus XYZ. *Jurnal Generic Vol. 6 No. 1*, 25-28.

Mastrogiannisa, N., Boutsinas, B., & Giannikos, I. (2009). A Method for Improving the Accuracy of Data Mining Classification Algorithms. *Computers & Operations Research*, 2829 -- 2839.

Michie, D. (1998). Learning concepts from data. *Expert Systems with Applications*, 193–204.

Mo, D., & Huang, S. H. (2012). Fractal-Based Intrinsic Dimension Estimation and Its Application in Dimensionality Reduction. *IEEE Transactions on Knowledge and Data Engineering*, 59-71.

Motiwalla, L. F., & Thompson, J. (2012). *Enterprise Systems for Management Second Edition.* New Jersey: Pearson Education, Inc.

Musanto, T. (2004). Faktor-Faktor Kepuasan Pelanggan dan Loyalitas Pelanggan: Studi Kasus pada CV. Sarana Media Advertising Surabaya. *Jurnal Manajemen & Kewirausahaan*, 123 - 136.

Musriadi. (2014, Oktober 16). *Riset Indonesia Pasar Potensial Produk FMCG*. Retrieved Oktober 31, 2015, from Antara Bengkulu: http://www.antarabengkulu.com

Ng, K., & Liu, H. (2000). Customer Retention via Data Mining. *Artificial Intelligence Review14*, 569–590.

Ngai, E., Xiu, L., & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 2592–2602.

Novita, M. (2008). *Regresi Linier Sederhana.* Retrieved November 2, 2015, from Academia: http://www.academia.edu/4378028/regresi_linier_sederhana

Nugraha, A. (2014). Pengaruh Ekuitas Merek Terhadap Keputusan Pembelian Produk Mie Instan (Studi Pada Mie Sedaap). *Universitas Negeri Yogyakarta*.

Oliver, R. (1999). Whence consumer loyalty? 33–44.

Olson, D., & Shi, Y. (2008). *Pengantar Ilmu Penggalian Data Bisnis.* Jakarta: Salemba Empat.

Patel, B. R., & Rana, K. K. (2014). A Survey on Decision Tree Algorithm for Classification. *International Journal of Engineering Development and Research*.

Pawlak, Z., & Skowron, A. (2007). Rough sets: Some extensions. *Information Sciences 177*, 28-40.

Peng, H., Long, F., & Ding, C. (2005). Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8*, 1226-1238.

Pranatha, A. A. (2012). Analisis Perbandingan Lima Metode Klasifikasi Pada Dataset Sensus Penduduk. *Jurnal Sistem Informasi, Volume 4 Nomor 2, Institut Teknologi Sepuluh Nopember*.

Prasetyo, E. (2014). *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab.* Yogyakarta: Andi Offset.

Purbasari, I. Y., & Nugroho, B. (2013). Bencmarking Algoritma Pemilihan Atribut Pada Klasifikasi Data Mining. *SNASTIA*, 47-54.

Putra, R., Suprayogi, A., & Kahar, S. (2013). Aplikasi SIG Untuk Penentuan Daerah Quick Count Pemilihan Kepala Daerah (Studi Kasus : Pemilihan Walikota Cirebon 2013, Jawa Barat). *Jurnal Geodesi Undip*.

Qian, W., & Shu, W. (2015). Mutual Information Criterion for Feature Selection from Incomplete Data. *Neurocomputing Journal*.

Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning.* California: Morgan Kaufmann.

Quinlan, J. R. (1994). C4.5: Programs for Machine Learning. *Machine Learning* , 235-240.

Rabinovich, E., & Bailey, J. P. (2004). Physical distribution service quality in Internet retailing: service pricing, transaction attributes, and firm attributes. *Journal of Operations Management* , 651–672.

Rastogi, R., & Shim, K. (2000). PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning. *Data Mining and Knowledge Discovery* .

Ravald, A., & Gronroos, C. (1996). The Value Concept and Relationship Marketing. 19–30.

Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review 68 (5)*, 105–111.

Rohman, I. F. (2015). Penerapan Algoritma C4.5 Pada Kepuasan Pelanggan Perum DAMRI. *Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang*, 1-14.

Rygielski, C., Wang, J.-C., & Yen, D. C. (2002). Data mining techniques for customer relationship management. *Technology in Society*, 483–502.

Santoso, T. B. (2012). Analisa dan Penerapan Metode C4.5 untuk Prediksi Loyalitas Pelanggan. *Jurnal Ilmiah Fakultas Teknik LIMIT'S Vol. 10 No.1*.

Sebban, M., & Nock, R. (2002). A hybrid filter/wrapper approach of feature selection using information theory. *Pattern Recognition 35* , 835–846.

Slowinski, R., & Vannderpooten, D. (2000). A Generalized Definition of Rough Approximation Based on Similarity. *IEEE Transactions on Knowledge and Data Engineering*, 331-336.

Stone, M., Woodcock, N., & Wilson, M. (1996). Managing the Change from Marketing Planning to Customer Relationship Management . *Long Range Planning, Vol. 29, No. 5, pp. 675 to 683*, 675-683.

Sundari, C. (2014, Juni 12). *Mengenal Fast Moving Consumer Goods*. Retrieved Oktober 31, 2015, from Kompasiana: http://www.kompasiana.com.

Tang, W., & Mao, K. (2007). Feature selection algorithm for mixed data with both nominal and continuous features. *Pattern Recognition Letters 28*, 563–571.

Tremblay, M. C., Berndt, D. J., & Studnicki, J. (2006). Feature Selection for Predicting Surgical Outcomes. *Proceedings of the 39th Hawaii International Conference on System Sciences*, 1-9.

Tsai, C.-F., & Chen, M.-Y. (2010). Variable selection by association rules for customer churn prediction of multimedia on demand. *Expert Systems with Applications*, 2006–2015.

Tu, C. M., Shin, D., & Shin, D. (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. *Autonomic and Secure Computing*, 183-187.

Udo, G. J., Bagchi, K. K., & Kirs, P. J. (2010). An assessment of customers' e-service quality perception, satisfaction and intention. *International Journal of Information Management*, 481–492.

Vergara, J. R., & Este´vez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Comput & Applic*, 175–186.

Wei, M., Chow, T. W., & Chan, R. H. (2015a). Clustering Heterogeneous Data with k-Means by Mutual Information-Based Unsupervised Feature Transformation. *Entropy*.

Wei, M., Chow, T. W., & Chan, R. H. (2015b). Heterogeneous Feature Subset Selection usng Mutual Information-Based Feature Transformation. *Neurocomputing*.

Wibowo, S. (2014). Neural Network dengan Algoritma Genetika sebagai Pemilihan Fitur pada Prediksi Loyalitas Pelanggan. *Majalah Ilmiah Pawiyatan, Program Studi Diploma III Teknik Elektro, Universitas PGRI Semarang*, 78-91.

Winarso, K. (2010). Kepuasan dan Loyalitas Pelanggan Pada Produk Susu Bayi Menggunakan Service Quality dan Path Analysis. *Jurnal Manajemen Teori dan Terapan, Tahun 3 No. 1*, 82-104.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining Practical Machine Learning Tools and Techniques*. USA: Morgan Kaufmann.

www.innovationpei.com. (n.d.). How You Can Profit from E-Business. *Customer Relationship Management* . Ontario: Ministry of Economic Development and Innovation and the Ontario Queen's Printer.

Yu, L., & Liu, H. (2006). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proceedings of the Twentieth International Conference on Machine Learning*.

Zeithaml, V. A., & Berry, L. L. (1996). The Behavioral Consequences of Service Quality. *Article in Journal of Marketing*, 31-46.

Zhang, K., Li, Y., Scarf, P., & Ball, A. (2011). Feature selection for high-dimensional machinery fault diagnosis data using multiple models and Radial Basis Function networks. *Neurocomputing*, 2941–2952.

Zhang, L., Chen, Y., Liang, Y., & Li, N. (2008). Application of Data MiningClassification Algorithms in Customer Membership Card Classification Model. *Innovation Management and Industrial Engineering* , 211 - 215.

# BIOGRAFI PENULIS

Heni Sulistiani. Lahir di Pagelaran, 12 Oktober 1986, anak pertama dari dua bersaudara. Penulis menempuh pendidikan formal mulai dari tahun 1992-1998 di SD N 2 Pringsewu Lampung, 1998-2001 di SMP N 1 Pringsewu Lampung, 2001-2004 di SMA N 1 Pringsewu Lampung. Selanjutnya pada tahun 2007 penulis menyelesaikan pendidikan Diploma Tiga di Akademi Manajemen Informatika dan Komputer (AMIK) Teknokrat Lampung jurusan Komputerisasi Akuntansi. Tahun 2012 penulis menyelesaikan pendidikan Strata Satu di Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Teknokrat Lampung jurusan Teknik Informatika. Pada tahun 2012 penulis mulai bekerja di Perguruan Tinggi Teknokrat Lampung sebagai Pengajar di bidang komputer dan Sekretaris Bidang Kemahasiswaan. Tahun 2014 penulis diterima sebagai mahasiswi Program Pascasarjana Institut Teknologi Sepuluh Nopember Surabaya, Fakultas Teknologi Informasi, Jurusan Sistem Informasi dengan NRP. 5214201009. *E-mail*: henie.tekno@gmail.com.