TESIS PM-147501

ANALISIS FAKTOR-FAKTOR YANG MEMPENGARUHI CHURN RATE PADA PERUSAHAAN TELEKOMUNIKASI MENGGUNAKAN METODE *SUPPORT VECTOR MACHINES*

(Studi Kasus: PT Telekomunikasi XYZ)

SAMSUL ARIFIN
09211650053011

DOSEN PEMBIMBING
Dr.Eng. Febriliyan Samopa, S.Kom., M.Kom.

DEPARTEMEN MANAJEMEN TEKNOLOGI
BIDANG KEAHLIAN MANAJEMEN TEKNOLOGI INFORMASI
FAKULTAS BISNIS DAN MANAJEMEN TEKNOLOGI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

LEMBAR PENGESAHAN

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Manajemen Teknologi (M.MT)
di
Institut Teknologi Sepuluh Nopember
Oleh :

SAMSUL ARIFIN
NRP. 09211650053011

Tanggal Ujian : 6 Juli 2018
Periode Wisuda : September 2018

Disetujui oleh :

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

2. Erma Suryani, S.T., M.T., Ph.D (Penguji)
   NIP : 19700427 200501 2 001

3. Faizal Mahananto, S.Kom., M.Eng., Ph.D (Penguji)
   NIPH : 5200201301010

Dekan Fakultas Bisnis dan Manajemen Teknologi,

Prof. Dr. Ir. Udisubakti Ciptomulyono, M.Eng.Sc.
NIP. 19590318 198701 1 001

(Halaman ini sengaja dikosongkan)

Analisis Faktor-Faktor yang Mempengaruhi Churn Rate pada Perusahaan Telekomunikasi Menggunakan Metode Support Vector Machines (Studi Kasus: PT Telekomunikasi XYZ)

Nama : Samsul Arifin
NRP : 09211650053011
Pembimbing : Dr.Eng. Febriliyan Samopa, S.Kom., M.Kom.

ABSTRAK

Dalam dunia telekomunikasi, pelanggan adalah aset yang paling berharga bagi perusahaan untuk keberlangsungan proses bisnis. Perpindahan pelanggan tentunya tidak diharapkan yang menuntut perusahaan perlu melakukan analisis dengan melakukan segmentasi pelanggan untuk mempermudah dilakukannya segmentasi bisnis dalam mendukung pengambilan keputusan yang tepat dan terarah, keputusan yang tepat diharapkan menghasilkan *value* yang dapat meningkatkan kinerja dan profit perusahaan. Dalam melakukan analisis data pelanggan, data masa lalu adalah mutlak dibutuhkan dengan mengambil variabel yang dapat menggambarkan keadaan pelanggan di masa yang akan datang, dengan melakukan analisa potensi churn dan non churn.. Dalam hal ini dilakukan kalsifikasi data mining menggunakan metode *Support Vector Machine*. Metode SVM dipilih karena akurasi yang cukup tinggi dalam melakukan klasifikasi. Dari pengujian ini akan terbentuk *performance* yang menggambarkan hasil klasifikasi berupa *churn* dan *non churn*.

Penelitian ini mengambil 7 variabel sebagai uji coba di mana hasil yang telah dlakukan, terdapat 4 variabel dependen yang mempengaruhi churn rate secara signifikan dengan nilai X_2 lebih besar dari X_{tabel} atau nilai P_{value} lebih kecil dari 0.05 yaitu variabel *Handset*, *Usage Data in kb, Voice in Minutes* dan *Reload*. Model yang dihasilkan kemudian diuji dengan SVM, dengan performance lebih baik dibandingkan dengan data awal. Akurasi terbaik ada pada 78.2% dengan proporsi 70 data training dan 30 data testing. Sedangkan pada data sebelum model, akurasi terbaik 73.9%, sehingga model yang dihasilkan lebih baik jika dibandingkan data awal.

Kata kunci: *Data mining; SVM;* Pelanggan *Churn;* Telekomunikasi

(Halaman ini sengaja dikosongkan)

Analysis of Churn Rate Factors in Telecommunication Industry using Support Vector Machines Method (Case Study: PT Telecommunication XYZ)

Student Name : Samsul Arifin
Student Identity Number : 09211650053011
Supervisor : Dr.Eng. Febriliyan Samopa, S.Kom., M.Kom.

ABSTRACT

In the telecommunications industry, the customer is the most valuable asset for the company to sustain the business process. The moving out of customers is certainly not expected so that companies need to analyze the customers profile to facilitate the conduct of business segmentation in support of decision making, the right decision is expected to generate value that can improve the performance and profit of the company. In analyzing customer data, past data is absolutely necessary with variables that can describe the customer value in the future, whether the customer is potentially churn or non churn. In this case, using Support Vector Machine method for classification. The SVM method was chosen because of its high accuracy in classification prediction. From this test will formed a performance that describes the results of the classification of churn and non churn.

In this research, there are 7 attributs have been used for testing. there are 3 attributes that significantly influence the churn rate of total 7 attribute which have been tested. Usage Data in kb and Voice in Minutes and Reload where these attributes have a performance value smaller than 5% of the total performance of the overall attribute. While the attribute is approaching the significant in SMS that is equal to 9%. From these results, the telecommunications companies XYZ should maintain Data and Voice services in minimizing churn rate.

This study took 7 variables as a trial where the results have been used for testing, there are 4 dependent variables that affect or influence churn rate significantly with X_2 value greater than X_{tabel} or value of P_{value} smaller than 0.05. Those variables are Handset, Usage Data in kb, Voice in Minutes and Reload. The resulting model have been tested with SVM, with better performance than the initial data. The best accuracy is 78.2% with the proportion of 70 training data and 30 data testing. While in the data before the model, the best accuracy is 73.9%, so the resulting model is better when compared to the initial data.

Keywords: Data mining; SVM; Customer Churn; Telecommunications

(Halaman ini sengaja dikosongkan)

KATA PENGANTAR

Puji dan syukur penulis panjatkan kehadirat Allah SWT atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan Proposal Tesis yang berjudul "*Analisis Faktor-Faktor yang Mempengaruhi Churn Rate pada Perusahaan Telekomunikasi Menggunakan Metode Regresi Logistik dan Support Vector Machines (Studi Kasus: Pt Telekomunikasi XYZ)*". Tesis ini diajukan untuk memenuhi prasyarat untuk menyelesaikan studi Magister di Program Studi Magister Manajemen Teknologi, Konsentrasi Manajemen Teknologi Informasi, Institut Teknologi Sepuluh Nopember Surabaya.

Dalam penyelesaian Tesis ini, penulis telah mendapatkan banyak dukungan moral maupun material dari banyak pihak. Atas bantuan yang telah diberikan penulis ingin menyampaikan penghargaan dan ucapan terima kasih yang sebesar-besarnya kepada:

1. Bapak Prof. Dr. Ir. Udisubakti Ciptomulyono, M. EngSc selaku Dekan Fakultas Bisnis dan Manajemen Teknologi (FBMT).
2. Bapak Dr. Tech, Ir. R. V. Hari Ginardi, M.Sc. selaku Kepala Program Studi Departemen Manajemen Teknologi sekaligus Dosen Wali MTI yang telah memberikan pengarahan, dan ilmu pengetahuan.
3. Bapak Dr.Eng. Febriliyan Samopa, S.Kom., M.Kom selaku Pembimbing Tesis yang telah meluangkan waktu, tenaga dan pikiran dalam memberikan bimbingan, masukan, pengarahan, dan ilmu pengetahuan.
4. Seluruh dosen pengajar yang telah memberikan pengajaran dan ilmu yang begitu banyak. Serta seluruh karyawan MMT-ITS yang telah banyak membantu dalam proses administrasi, perkuliahan dan berbagai hal yang menunjang terselesainya pengajuan tesis ini.

5. Bapak Oki Kurniawan selaku Manajer Marketing di PT telekomunikasi XYZ yang telah banyak membantu dan memberikan banyak informasi yang dibutuhkan oleh penulis dan serta meluangkan waktunya untuk berdiskusi tentang banyak hal berkaitan dengan data penunjang dalam penelitian ini.
6. Kedua orang tua serta saudara yang selalu memberikan dukungan baik melalui doa ataupun material untuk kesuksesan dan kelancaran penelitian ini.
7. Saudari Ely Windasari yang selalu memberikan support berupa dukungan referensi, metode yang diberikan kepada penulis dalam menunjang penyelesaian penyusunan Tesis ini.
8. Teman-teman MTI angkatan 2016 yang selalu memotivasi, mengingatkan, memberi masukan, dan selalu memberi suntikan semangat kepada penulis dalam penyusunan Tesis ini.
9. Semua pihak yang tidak dapat disebutkan satu persatu, yang telah banyak memberikan berbagai macam bantuan dalam penyusunan Tesis ini.

Akhir kata, penulis berharap Tesis ini dapat memberikan manfaat kepada pembaca mengenai analisis faktor-faktor *churn rate* terutama pada *billing customer* di perusahaan telekomunikasi. Penulis menyadari bahwa tesis ini masih jauh dari kesempurnaan dan memiliki banyak kekurangan. Oleh karena itu, dengan kerendahan hati penulis mengharapkan masukan dan saran yang membangun dari pembaca untuk perbaikan ke depan.

Surabaya, 6 Juni 2018

Samsul Arifin

DAFTAR ISI

DAFTAR GAMBAR

(Halaman ini sengaja dikosongkan)

DAFTAR TABEL

(Halaman ini sengaja dikosongkan)

DAFTAR DIAGRAM

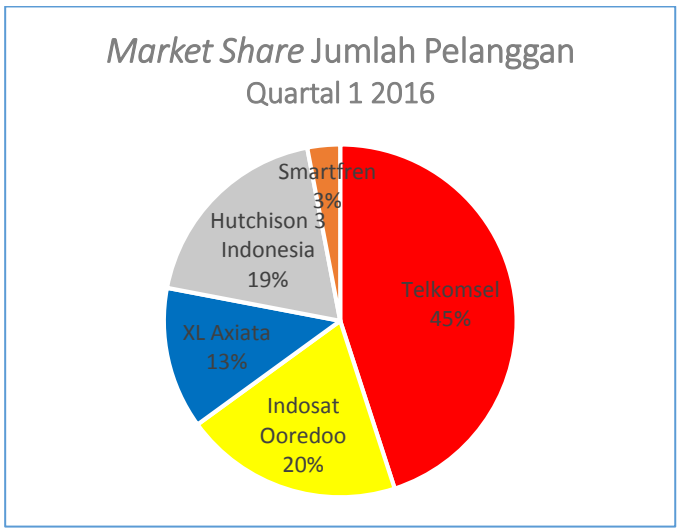
(Halaman ini sengaja dikosongkan)

BAB I
PENDAHULUAN

Pada bab ini akan dijelaskan tentang latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan dengan judul penelitian "*Prediksi dan Analisis Faktor-faktor yang Mempengaruhi Customer Churn pada Perusahaan Telekomunikasi Menggunakan Metode Support Vector Machine*".

1.1 Latar Belakang

Peran *customer* atau *subscriber* merupakan poin utama dalam kelangsungan operasional perusahan tidak terkecuali perusahaan telekomunikasi. Merupakan tantangan tersendiri bagi perusahaan operator telekomunikasi seluler dalam menurunkan jumlah pelanggan yang berhenti menggunakan layanan perusahaan dan pindah ke perusahaan kompetitor. Perilaku pelanggan yang meninggalkan layanan yang diberikan perusahaan pada industri operator telekomunikasi seluler disebut *churn*. Dalam hal ini, peran produk dan inovasi yang dikeluarkan oleh perusahaan menjadi peran penting dalam berkompetisi di dunia telekomunikasi.



Sumber: Internal Company

Gambar 1.1 Market Share Industri di Indonesia 2016

Pada gambar 1.1 terlihat bahwa secara *market share*, presentase jumlah pelanggan (*subscriber*) terbesar pengguna layanan telekomunikasi ada pada operator telkomsel yang menguasai hampir 50% market. Selanjutnya diikuti oleh tiga operator besar lainnya yaitu Indosat Ooredoo, XL Axiata dan Hutcison 3. Seiring dengan pertumbuhan industri telekomunikasi yang cukup dinamis, maka dimungkinkan adanya pergeseran jumlah pelanggan dari masing-masing operator dengan beberapa faktor yang mempengaruhi *customer* menjadi *churn* sehingga dapat berpindah ke operator lain.

Salah satu sumber internal perusahaan telekomunikasi pada gambar 2 menggambarkan bahwa dari kuartal 1, kuartal 2, kuartal 3 dan kuartal 4 mengalami peningkatan jumlah pelanggan *churn* cukup signifikan di mana masing-masing regional mengalami peningkatan *churn*. Jika dilihat lebih detail secara keseluruhan selama 2017, hanya ada satu region yang mengalami penurunan *churn rate*, yaitu regional *East* pada kuartal kedua. Secara nasional peningkatan *churn rate* pada masing-masing kuartal tumbuh mencapai 20%. Hal ini menjadi pekerjaan rumah perusahaan yang mengharuskan untuk melakukan evalusasi serta melakukan beberapa analisis tentang faktor-faktor yang berpengaruh terhadap potensi *churn rate*.

REGION	Q1-Q2	Q2-Q3	Q3-Q4	Q4-Q1
Central	15%	7%	16%	17%
East	39%	-19%	1%	23%
Jabo	2%	7%	9%	6%
North	19%	11%	32%	30%
West	3%	5%	22%	36%
National	**15%**	**0%**	**13%**	**20%**

Sumber: Internal Company

Gambar 1.2 *Churn* data Perusahaan Telekomunikasi XYZ Q1– Q4 2017

Beberapa metode klasifikasi belakangan ini sudah berkembang sangat pesat, terlebih dalam proses pengolahan data. Metode yang popular digunakan oleh para peneliti adalah berbasis *data mining* baik yang menggunakan klasifikasi atau *clustering*. Metode-metode tersebut seperti *Decision Tree* dan *K-Mean*. Namun kebanyakan paper hanya membahas tentang prediksi *churn* tanpa

melakukan analisis pada *historical* dari *customer billing*. Dalam penelitian ini, selain dilakukan prediksi *churn*, juga menganalisis *historical billing* yang meliputi penggunaan *voice* baik dalam *event* atau *seconds*, sms dalam *event*, serta penggunaan data dalam *kb* atau *event* dalam 3 bulan terakhir pada tahun 2017 menggunakan metode *Support Vector Machine* dengan menjadikan layanan *voice, sms* dan *data* tersebut sebagai variabel utama dalam pengujian, sehingga dari hasil uji coba dapat dijadikan acuan dalam menganalisis faktor-faktor yang mempengaruhi *customer churn*.

Dengan adanya prediksi *customer churn* dan analisis perilaku *customer* ini diharapkan dapat membantu divisi *marketing* dan CRM dalam melakukan segmentasi pelanggan yang dapat mempermudah melakukan *product campaign* sehingga dapat meminimalisir adanya *customer churn* dan meningkatkan profit perusahaaan. Dari analisa yang dilakukan menghasilkan variabel-variabel apa yang berpotensi besar menjadikan *customer* menjadi *churn*, sehingga dari variabel tersebut dapat dijadikan patokan dalam menentukan sebuah produk fokus tertentu sesuai dengan *profile* dari *customer*.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, maka masalah dalam penelitian ini adalah belum diketahui faktor-faktor yang mempengaruhi *churn rate* secara signifikan. Layanan yang ada di perusahaan telekomunikasi belum mencerminkan efektivitas masing-masing dalam mempengaruhi tingginya angka *churn rate*. Untuk itu, diperlukan upaya dalam menganalisis perilaku *billing* konsumen dalam menentukan potensi *churn rate*. Setelah didapati variabel terpilih sebagai faktor berpengaruh, belum diketahui apakah variabel terpilih akan menghasilkan klasifikasi terbaik dibandingkan dengan variabel total. Dari masalah ini, maka rumusan masalah pada penelitian ini adalah "faktor-faktor yang mempengaruhi *customer churn* secara signifikan pada perusahaan telekomunikasi dengan studi kasus PT. Telekomunikasi XYZ".

1.3 Tujuan Penelitian

Dari rumusan masalah di atas, maka dapat dibuat tujuan dari penelitian ini adalah untuk mengetahui faktor-faktor apa saja yang berpengaruh signifikan terhadap terjadinya *churn rate* terlebih dalam penggunaan layanan *billing customer* dengan memperhatikan korelasi antara variabel independen dengan variabel dependen. Selain itu, penelitian ini bertujuan untuk melakukan klasifikasi dan prediksi *churn rate* di masa yang akan datang berdasarkan pola pada data masa lalu.

1.4 Manfaat Penelitian

Dengan adanya penelitian ini, diharapkan mempunyai nilai manfaat sebagai berikut:

1. Meningkatkan kualitas dan strategi pemasaran yang lebih kompetitif di divisi *marketing* dengan menganalisa layanan yang kurang efektif yang nantinya akan sangat berpengaruh pada strategi dalam mengurangi *churn rate*.
2. Sebagai bahan acuan dalam menyelesaikan problematika industri pada perusahaan telekomunikasi.
3. Sebagai bahan referensi baik dalam penggunaan informasi data atau pengembangan metode dalam dunia akademisi.

1.5 Batasan Masalah

Dalam penelitian ini, ditentukan batasan masalah agar topik permasalahan menjadi terfokus dan dibahas secara rinci. Batasan-batasan tersebut adalah sebagai berikut:

1. Data yang digunakan adalah data *billing* dari penggunaan layanan pada perusahaan telekomunikasi yang meliputi *voice*, sms dan layanan data.
2. Pengumpulan data terbatas pada pelanggan salah satu *provider* telekomunikasi di area Jatim dan Bali Nusa.
3. Data yang menjadi objek penelitian adalah data pelanggan yang menggunakan layanan tetap dan aktif minimal 3 bulan selama tahun 2018.

4. Penelitian ini didasarkan pada metode yang menggunakan data *supervised learning*, di mana ada data training sebagai data acuan dalam menentukan klasifikasi *churn* dan *non churn*.

1.6 Sistematika Penulisan

Berikut adalah sistematika penulisan yang digunakan pada peneltian ini:

Bab I Pendahuluan

Bab ini menyajikan mengenai latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, metodologi dan sistematika penulisan

Bab II Tinjauan Pustaka

Bab ini menyajikan tentang studi literatur mengenai teori dan penelitian sebelumnya yang telah dilakukan yang mendasari penelitian.

Bab III Metode Penelitian

Bab ini menyajikan metode dan langkah-langkah yang dilakukan oleh peneliti dalam melakukan penelitian ini.

(Halaman ini sengaja dikosongkan)

BAB II
KAJIAN PUSTAKA

Bab ini menjelaskan tentang studi pustaka berupa teori-teori penunjang yang digunakan sebagai acuan dalam pembuatan laporan yang meliputi *customer churn, data mining*, klasifikasi, SVM dan regresi.

2.1 Customer Churn

Customer churn merupakan kondisi dimana pelanggan tidak melanjutkan berlangganan pada perusahaan yang dipilih dan berpindah pada perusahaan pesaing (Richeldi dan Perrucci, 2002). *Customer churn* merupakan istilah yang digunakan untuk mendefinisikan bahwa seorang *customer* (pelanggan) menghentikan hubungan bisnis dengan perusahaan (Liaou, 2007). *Customer churn* merupakan istilah untuk mendefinisikan perputaran pelanggan, atau lebih khusus disebut sebagai *customer churn management*. Manajemen *customer churn* merupakan sebuah konsep untuk mengidentifikasi para pelanggan yang berkeinginan untuk berpindah dari perusahaan yang telah dipilih ke perusahaan pesaing, sehingga sekali pelanggan tersebut teridentifikasi *churn* maka pelanggan tersebut menjadi target pemasaran proaktif sebagai upaya *retention* (Hadden dkk, 2005).

Customer churn dapat disebabkan oleh banyak hal, mulai dari tarif yang kompetitif antar operator, fitur dan fasilitas yang kompetitif, sampai bagaimana *provider* melayani, berinteraksi, dan mengelola hubungannya dengan pelanggan-pelanggannya namun dalam hal ini penulis akan fokus pada layanan dan *profile customer*. Masalah *customer churn* ini menjadi krusial, karena biaya yang dikeluarkan untuk mendapatkan pelanggan baru, untuk iklan, *marketing,* komisi, dan lain-lain akan jauh lebih besar dibandingkan biaya yang harus dikeluarkan untuk menjaga pelanggan yang sudah ada. Ditambah lagi belum kebanyakan pelanggan baru cenderung tidak lebih menghasilkan keuntungan dibandingkan pelanggan yang sudah lama dan bertahan (Herawati, 2016). Sehingga mempertahankan pelanggan yang sudah ada merupakan prioritas utama, dari pada

mencoba memenangkan persaingan untuk mendapatkan pelanggan baru (Hadden, 2005).

Pelanggan *churn* dapat dibagi ke dalam dua kelompok, yaitu pelanggan *churn voluntary* (sukarela) dan *non voluntary*. Pelanggan *churn voluntary* lebih sulit dideteksi dikarenakan pelanggan tersebut memutuskan untuk *churn* karena keputusan mereka sendiri, sedangkan pelanggan *churn non voluntary* tidak menghentikan layanan secara sukarela, melainkan layanan dihentikan oleh provider secara sepihak dikarenakan beberapa hal, antara lain karena menyalahgunakan layanan atau tidak membayar tagihan.

Analisis *churn* sering digunakan pada semua sektor, salah satunya adalah sektor telekomunikasi yang mempunyai potensi *churn* yang sangat tinggi, sehingga biaya *churn* yang sangat besar menyebabkan potensi kerugian yang besar (Gursoy, 2010). Biaya *churn* yang sangat besar menjadikan manajemen *churn* sebagai senjata paling kompetitif, dan menjadi dasar berbagai strategi pemasaran yang berfokus pada pelanggan. Perusahaan dapat menentukan jenis pelanggan yang paling mungkin untuk *churn*, dan yang paling mungkin untuk tetap setia dengan menggunakan produk perusahaan. Bagian dari proses ini adalah menentukan nilai pelanggan, karena terkadang perusahaan melepaskan pelanggan yang tidak menguntungkan atau tidak terlalu menguntungkan. Ketika informasi telah dimiliki oleh perusahaan, maka manajer pemasaran dapat mengambil tindakan tepat dan strategis untuk meminimalkan *churn*, memenangkan kembali *churn,* dan mengefektifkan biaya yang sesuai untuk pelanggan di masa depan, termasuk pelanggan yang memiliki kemungkinan kecil untuk *churn* (Richeldi dan Perrucci, 2002). Dengan memperkirakan pelanggan yang mungkin akan beralih pada *provider* lain, maka organisasi dapat menentukan usaha yang bertujuan untuk meningkatkan loyalitas pelanggan dan mengembangkan strategi pemasaran agar retensi pelanggan meningkat.

2.2 Data Mining

Data mining adalah suatu proses ekstraksi atau penggalian data dan informasi dengan *volume* besar, yang belum diketahui sebelumnya, namun dapat dipahami dan berguna, seta didapatkan dari sebuah *database* berkapasitas besar

serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting (Conolly dkk, 2005). Menurut Berry dan Linoff (2011, 7) *data mining* adalah suatu pencarian dan analisa dari jumlah data yang sangat besar dan bertujuan untuk mencari arti dari pola dan aturan. Turban (2007) menyatakan bahwa *data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar (Turban dkk. 2007). Terdapat beberapa istilah lain yang memiliki makna sama dengan *data mining*, yaitu *Knowledge discovery in databases* (KDD), ekstraksi pengetahuan (*knowledge extraction*), Analisa data/pola (*data/pattern analysis*), kecerdasan bisnis (*business intelligence*) dan *data archaeology* dan *data dredging* (Larose, 2005).

Dari beberapa teori yang dijabarkan oleh para ahli diatas, bahwa *data mining* adalah suatu pencarian dan analisa pada suatu koleksi data (*database*) yang sangat besar sehingga ditemukan suatu pola yang menarik dengan tujuan mengekstrak informasi dan *knowledge* yang akurat dan berpotensial, serta dapat dipahami dan berguna dari *database* yang besar serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting (Yudhistira, 2017).

Menurut Hoffer, Ramesh & Topi (2012), tujuan dari adanya *data mining* adalah:

- *explanatory*, yaitu untuk menjelaskan beberapa kegiatan observasi atau suatu kondisi.
- *confirmatory*, yaitu untuk mengkonfirmasi suatu hipotesis yang telah ada.
- *exploratory*, yaitu untuk menganalisis data baru suatu relasi yang janggal.

Karakteristik *data mining* menurut Turban (2007, 230):

a. Seringnya data terpendam dalam *database* yang sangat besar dan kadang datanya sudah bertahun-tahun.
b. Lingkungan *data mining* biasanya berupa arsitektur *client-server* atau arsitektur sistem informasi berbasis web.

c. *Tool* baru yang canggih, termasuk *tool* visualisasi tambahan, membantu menghilangkan lapisan informasi yang terpendam dalam *file-file* yang berhubungan atau *record-record* arsip publik.
d. Pemilik biasanya seorang *end user*, didukung dengan *data drill* dan *tool* penguasaan *query* yang lain untuk menanyakan pertanyaan dan mendapatkan jawaban secepatnya, dengan sedikit atau tidak ada kemampuan pemrograman.
e. *Tool data mining* dengan kesediannya dikombinasikan dengan *spreadsheet* dan *tool software* pengembangan yang lainnya.
f. Karena besarnya jumlah data dan usaha pencarian yang besar-besaran, kadang-kadang diperlukan penggunaan proses *parallel* untuk *data mining*.

Kelebihan *data mining* sebagai alat analisis:

a. *Data mining* mampu menangani data dalam jumlah besar dan kompleks.
b. *Data mining* dapat menangani data dengan berbagai macam tipe atribut.
c. *Data mining* mampu mencari dan mengolah data secara semi otomatis.
d. Disebut semi otomatis karena dalam beberapa teknik *data mining*, diperlukan parameter yang harus di *input* oleh *user* secara manual.
e. Data mining dapat menggunakan pengalaman ataupun kesalahan terdahulu untuk meningkat.

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu (Larose, 2005):

1. *Description* (Deskripsi)

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

2. *Estimation* (Estimasi)

   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah *numeric* daripada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh yaitu estimasi nilai indeks prestasi kumulatif mahasiswa program pasca sarjana dengan melihat nilai indeks prestasi mahasiswa tersebut pada saat mengikuti program sarjana.

3. *Prediction* (Prediksi)

   Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada dimasa mendatang. Contoh prediksi dalam bisnis dan penelitian adalah:

   a. Prediksi harga beras dalam tiga bulan yang akan datang.
   b. Prediksi tingkat penganguran lima tahun akan datang.
   c. Predisksi persentase kanaikan kecelakaanlalu lintas tahun depan jika batas bawah kecepatan dinaikan. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. *Classification* (Klasifikasi)

   Dalam klasifikasi, terdapat target variabel kategori. sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Contoh lain klasifikasi dalam bisnis dan penelitian adalah:

   a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
   b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.
   c. Mendiagnosis penyakit seorang pasien untuk mendapatkan termasuk penyakit apa.

5. *Clustering* (Pengklusteran)

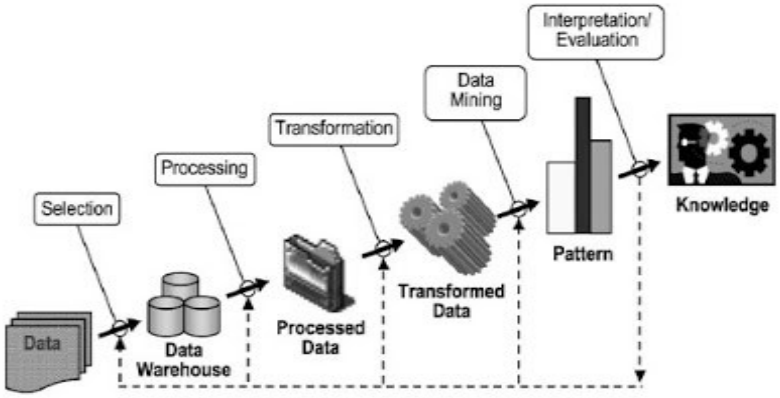
   Pengklusteran merupakan pengelompokan *record* pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidak miripan dengan *record-record* dalam kluster lain. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (*homogen*), yang mana kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal. Contoh pengklusteran dalam bisnis dan penelitian adalah:

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

6. *Association* (Asosiasi)

   Tugas asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja. Contoh asosiasi dalam bisnis dan penelitian adalah:

   a. Meneliti jumlah pelanggan dari perusahaan telekomunikasi seluler yang diharapkan untuk memberikan respon positif terhadap penawaran *upgrade* layanan yang diberikan.
   b. Menemukan barang dalam supermarket yang dibeli secara bersamaan dan barang yang tidak pernah dibeli secara bersamaan.



Gambar 2.1 Tahapan Data Mining

Pada gambar 2.1 tahapan yang dilakukan pada proses *data mining* diawali dari seleksi data dari data sumber ke data target, tahap *preprocessing* untuk memperbaiki kualitas data, transformasi, *data mining* serta tahap interpretasi dan evaluasi yang menghasilkan *output* berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik. Secara detail dijelaskan sebagai berikut (Fayyad, 1996):

a. *Data selection*

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing / cleaning*

   Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

3. *Transformation*

   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data mining*

   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation / evalution*

   Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.3 Algoritma Klasifikasi

Data Mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu Deskripsi, Estimasi, Prediksi, Klasifikasi, Pengklusteran, dan Asosiasi. Klasifikasi merupakan bagian dari algoritma *datamining*, klasifikasi ini adalah algoritma yang menggunakan data dengan target (*class/label*) yang berupa nilai kategorikal/nominal. Menurut Gorunescu (2011) proses klasifikasi didasarkan pada empat komponen mendasar, yaitu:

1. Kelas (*Class*)

   Variabel dependen dari model, merupakan variabel kategorikal yang merepresentasikan "label" pada objek setelah klasifikasinya. Contoh kelas semacam ini adalah: adanya kelas penyakit jantung, loyalitas pelanggan, kelas bintang (galaksi), kelas gempa bumi (badai), dll.

2. Prediktor (*Predictor*)

   Variabel independen dari model, direpresentasikan oleh karakteristik (atribut) dari data yang akan diklasifikasikan dan berdasarkan klasifikasi yang telah dibuat. Contoh prediktor tersebut adalah: merokok, konsumsi alkohol, tekanan darah, frekuensi pembelian, status perkawinan, karakteristik (satelit) gambar,

catatan geologi yang spesifik, kecepatan dan arah angin, musim , lokasi terjadinya fenomena , dll.

3. Pelatihan dataset (*Training dataset*)

   Kumpulan data yang berisi nilai-nilai dari kedua komponen sebelumnya dan digunakan untuk melatih model dalam mengenali kelas yang cocok/sesuai, berdasarkan prediktor yang tersedia. Contoh set tersebut adalah: kelompok pasien yang diuji pada serangan jantung, kelompok pelanggan supermarket (diselidiki oleh *intern* dengan jajak pendapat), *database* yang berisi gambar untuk monitoring teleskopik dan pelacakan objek astronomi, *database* badai, *database* penelitian gempa.

4. Dataset Pengujian (*Testing Dataset*)

   Berisi data baru yang akan diklasifikasikan oleh (*classifier*) model yang telah dibangun di atas sehingga akurasi klasifikasi (*model performance*) dapat dievaluasi.

2.4 Pengantar Support Vector Machine

Support Vectore Machine merupakan metode *data mining* yang digunakan untuk tujuan klasifikasi dan prediksi pada penelitian ini. Adapun pada pengantar ini akan dijelaskan tentang pengertian, kernel, kelebihan dan kekurangan SVM.

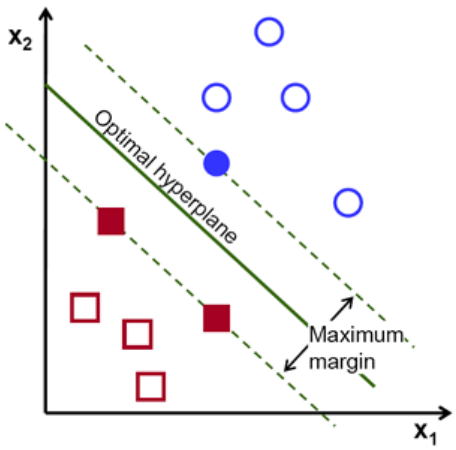
2.4.1 Pengertian Support Vector Machine

*Support Vector Machine (*SVM) dikembangkan oleh Boser, Guyon, Vapnik, dan pertama kali dipresentasikan pada tahun 1992 (Santosa, 2007) di *Annual Workshop on Computational Learning Theory*. Konsep dasar SVM sebenarnya merupakan kombinasi harmonis dari teori-teori komputasi yang telah ada puluhan tahun sebelumnya, seperti *marginhyperplane*. Menurut Y. Yin, Han, & Cai, (2011) *Support Vector Machine* (SVM) didefinisikan sebagai seperangkat metode pembelajaran terkait yang menganalisis data dan mengenali pola, yang kemudian digunakan untuk klasifikasi dan analisis regresi.

SVM mengambil satu set data input dan memprediksi untuk setiap masukan yang diberikan, yang berasal dari dua kelas yang kemudian di

klasifikasikan dengan mencari nilai *hyperplane* terbaik. Menurut Li, You, & Liu (2015) *Support Vector Machine* (SVM) merupakan pembelajaran yang mengarah ke pemrograman kuadratik dengan kendala *linear*. Berdasarkan minimalisasi risiko prinsip terstruktur, SVM berusaha untuk meminimalkan batas atas kesalahan generalisasi bukan kesalahan empiris, sehingga model prediksi baru efektif menghindari *over-pas* masalah. Selain itu, model SVM bekerja di ruang fitur berdimensi tinggi yang dibentuk oleh pemetaan *nonlinear* dari N-dimensi vektor *input* x ke dalam ruang fitur K-dimensi (K> N) melalui penggunaan fungsi φ *nonlinear* (x).

Hyperplane (batas keputusan) pemisah terbaik antara kedua kelas dapa ditemukan dengan mengukur *margin hyperplane* tersebut dan mencari titik maksimalnya. *Margin* adalah jarak antara *hyperplane* tersebut dengan data terdekat dari masing-masing kelas. Data yang paling dekat ini disebut *support vector*. Garissolid pada Gambar 2.2 menunjukkan *hyperplane* yang terbaik, yaitu yang terletak tepat pada tengah-tengah kedua kelas, sedangkan data lingkaran dan bujur sangkar yang dilewati garis batas *margin* (garis putus putus) adalah *support vector*. Usaha untuk mencari lokasi *hyperplane* ini merupakan inti dari peoses pelatihan *Support Vector Machine* (SVM) (Prasetyo, 2012).



Gambar 2.2 Algoritma SVM Sumber : Prasetyo (2012)

Menurut Santosa (2007) *hyperplane* klasifikasi linier SVM dinotasikan:

$$f(x) = \boldsymbol{w}^T \boldsymbol{x} + b$$

sehingga menurut Vapnik dan Cortes (1995) diperoleh persamaan:

$$[(\boldsymbol{w}^{T}.\boldsymbol{x}_i) + b] \geq 1 \; untuk \; y_i = +1$$

$$[(\boldsymbol{w}^{T}.\boldsymbol{x}_i) + b] \leq -1 \; untuk \; y_i = -1$$

dengan,$\boldsymbol{x}_i$ = himpunan *data training*, *i* = 1,2,...*n* dan *yi* = label kelas dari x_iUntuk mendapatkan *hyperplane* terbaik adalah dengan mencari *hyperplane* yang terletak di tengah-tengah antara dua bidang pembatas kelas dan untuk mendapatkan *hyperplane* terbaik itu, sama dengan memaksimalkan *margin* atau jarak antara dua set objek dari kelas yang berbeda (Santosa, 2007). *Margin* dapat dihitung dengan $\frac{2}{\|\boldsymbol{w}\|}$

Mencari *hyperplane* terbaik dapat digunakan metode *Quadratic Progamming* (QP) *Problem* yaitu meminimalkan

$$\frac{1}{2}\boldsymbol{w}^{T}w$$

dengan syarat $y_i(\boldsymbol{w}^{T}.\boldsymbol{x}_i + b) \geq 1, i = 1,2,3, \ldots . n$

Solusi untuk mengoptimasi oleh Vapnik (1995) diselesaikan dengan menggunakan fungsi L*agrange* sebagai berikut:

$$L(\boldsymbol{w}, b, a) = \tfrac{1}{2}\boldsymbol{w}^{T}\boldsymbol{w} - \textstyle\sum_{i=1}^{n} \alpha_i\{y_i[(\boldsymbol{w}^{T}.\boldsymbol{x}_i) + b] - 1\}$$

(1)

dengan α*i* = pengganda fungsi *Lagrange* dan = 1,2,..., *n*

Nilai optimal dapat dihitung dengan memaksimalkan L terhadap α*i*, dan meminimalkan *L* terhadap *w* dan *b.* Hal ini seperti kasus dual *problem*

$max_{\alpha}W(a) = max_{\alpha}\big(min_{w.b}L(w, b, \alpha)\big)$ (Gunn, 1998).

Nilai minimum dari fungsi *lagrange* tersebut diberikan oleh

$$\frac{\partial L}{\partial b} = 0 \rightarrow \textstyle\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{w}} = 0 \rightarrow \mathbf{w} = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\mathbf{x}_{\mathrm{i}}\mathrm{y}_{\mathrm{i}}$$

(2)

Untuk menyederhanakannya persamaan (1) harus ditransformasikan ke dalam fungsi *Lagrange Multiplier* itu sendiri, sehingga menurut Santosa (2007) persamaan (1) menjadi

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i(\boldsymbol{w}^T.\boldsymbol{x}_i) - b\sum_{i=t}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i$$

(3)

Berdasarkan persamaan (2), maka persamaan (3) oleh Hastie (2001) menjadi sebagai berikut:

$$L_d = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j y_i y_j \boldsymbol{x}_i^T\, \boldsymbol{x}_j$$

dan diperoleh *dual problem*

$$max_\alpha L_d = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\, \alpha_j y_i y_j \boldsymbol{x}_i^T x_j$$

dengan batasan, $\alpha_i \geq 0, i = 1,2, \ldots, n$ dan $\sum_{i=1}^{n} \alpha_i y_i = 0$

Data training dengan $\alpha_i > 0$ terletak pada *hyperplane* disebut *support vector*. *Data training* yang tidak terletak pada *hyperplane* tersebut mempunyai $\alpha_i = 0$. Setelah solusi permasalahan *quadratic progamming* ditemukan (nilai αi), maka kelas dari data yang akan diprediksi atau *data testing* dapat ditentukan berdasarkan nilai fungsi berikut.

$$f(x_t) = \sum_{s=1}^{ns} \alpha_s y_s \boldsymbol{x}_s.\boldsymbol{x}_t + b$$

dengan

$\boldsymbol{x}_t$ = data yang akan diprediksi kelasnya (*data testing*)

$\boldsymbol{x}_s$ = data *support vector, s* = 1,2, ...*ns*

ns = banyak data *support vector*

2.4.2 Kernel Trick SVM

Kernel trick merupakan suatu *trick* yang digunakan untuk memecahkan kasus non linier (Santosa, 2007). Dalam dunia nyata, berbagai kasus klasifikasi memperlihatkan ketidaklinieran sehingga membutuhkan *trick* untuk memecahkan masalah non-linier tersebut dengan kernel trick. Fungsi Kernel yang sering digunakan dalam literatur SVM (Karatzouglou, 2006) antara lain sebagai berikut:

a. Kernel *linear* adalah kernel yang paling sederhana dari semua fungsi kernel. Kernel ini biasa digunakan dalam kasus klasifikasi teks.

$K(x_i, x_j) = x_i{}^T x_j$

b. Kernel *Polynominal* adalah kernel yang sering digunakan untuk klasifikasi gambar.

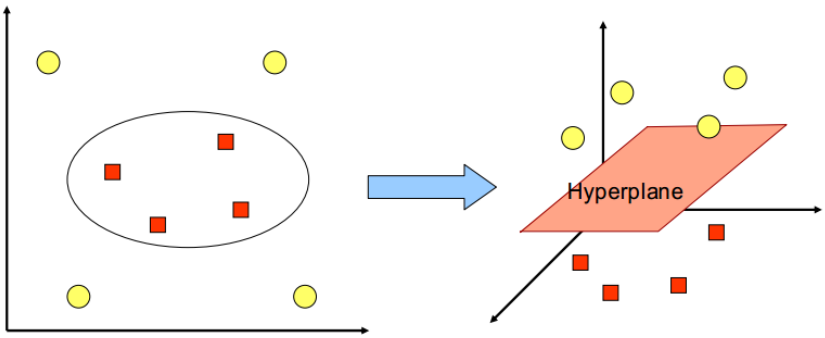
$K(x_i, x_j) = (\gamma x_i{}^T x_j + r)^p, \gamma > 0$

c. Kernel *Gaussian radial basis function (RBF)* adalah kernel yang umum digunakan untuk data yang sudah valid (*available*) dan merupakan *default* dalam *tools SVM*.

$K(x_i, x_j) = exp\left(-\gamma \|x_i - x_j\|^2\right), \gamma > 0$

d. Kernel *Tangent Hyperbolic* adalah kernel yang sering digunakan untuk *neural networks*.

$K(x_i, x_j) = \tanh\left(\gamma x_i{}^T x_j + r\right)$



Input Space X High-dimensional Feature space

Gambar 2.3 Pemetaan input *space* berdimensi dua dengan pemetaan ke dimensi tinggi

Ilustrasi dari konsep ini dapat dilihat pada gambar. Pada gambar di atas sisi kiri diperlihatkan data pada *class* kuning dan data pada *class* merah yang berada pada *input space* berdimensi dua tidak dapat dipisahkan secara *linear*. Selanjutnya gambar menunjukkan bahwa fungsi Φ memetakan tiap data pada *input space* tersebut ke ruang vektor baru yang berdimensi lebih tinggi (dimensi 3), dimana kedua *class* dapat dipisahkan secara *linear* oleh sebuah *hyperplane*.

2.4.3 Kelebihan dan Kekurangan SVM

Dalam memilih solusi untuk menyelesaikan suatu masalah, kelebihan dan kelemahan masing-masing metode harus diperhatikan. Selanjutnya metode yang tepat dipilih dengan memperhatikan karakteristik data yang diolah. Dalam hal SVM, walaupun berbagai studi telah menunjukkan kelebihan metode SVM dibandingkan metode konvensional lain, SVM juga memiliki berbagai kelemahan. Kelebihan SVM antara lain (Nugroho, 2003):

1. Generalisasi

   Generalisasi didefinisikan sebagai kemampuan suatu metode (SVM, *neural network*, dsb.) untuk mengklasifikasikan suatu *pattern*, yang tidak termasuk data yang dipakai dalam *fase* pembelajaran metode itu. Vapnik menjelaskan bahwa *generalization error* dipengaruhi oleh dua faktor: *error* terhadap *training set*, dan satu faktor lagi yang dipengaruhi oleh dimensi VC (*Vapnik-Chervokinensis*). Strategi pembelajaran pada *neural network* dan umumnya metode *learning machine* difokuskan pada usaha untuk meminimimalkan *error* pada *training-set*. Strategi ini disebut *Empirical Risk Minimization* (ERM). Adapun SVM selain meminimalkan *error* pada *training-set*, juga meminimalkan faktor kedua. Strategi ini disebut *Structural Risk Minimization* (SRM), dan dalam SVM diwujudkan dengan memilih *hyperplane* dengan *margin* terbesar. Berbagai studi empiris menunjukkan bahwa pendekatan SRM pada SVM memberikan *error* generalisasi yang lebih kecil dari pada yang diperoleh dari strategi ERM pada *neural network* maupun metode yang lain.

2. *Curse of dimensionality*

   Curse of dimensionality didefinisikan sebagai masalah yang dihadapi suatu metode *pattern recognition* dalam mengestimasikan parameter (misalnya jumlah *hidden neuron* pada *neural network*, *stopping criteria* dalam proses pembelajaran dsb.) dikarenakan jumlah sampel data yang relatif sedikit dibandingkan dimensional ruang vektor data tersebut. Semakin tinggi dimensi dari ruang vektor informasi yang diolah, membawa konsekuensi dibutuhkannya jumlah data dalam proses pembelajaran. Pada kenyataannya

seringkali terjadi, data yang diolah berjumlah terbatas, dan untuk mengumpulkan data yang lebih banyak tidak mungkin dilakukan karena kendala biaya dan kesulitan teknis. Dalam kondisi tersebut, jika metode itu "terpaksa" harus bekerja pada data yang berjumlah relatif sedikit dibandingkan dimensinya, akan membuat proses estimasi parameter metode menjadi sangat sulit.

Curse of dimensionality sering dialami dalam aplikasi di bidang *biomedical engineering*, karena biasanya data biologi yang tersedia sangat terbatas, dan penyediaannya memerlukan biaya tinggi. Vapnik membuktikan bahwa tingkat generalisasi yang diperoleh oleh SVM tidak dipengaruhi oleh dimensi dari input *vector*. Hal ini merupakan alasan mengapa SVM merupakan salah satu metode yang tepat dipakai untuk memecahkan masalah berdimensi tinggi, dalam keterbatasan sampel data yang ada.

3. Landasan teori

   Sebagai metode yang berbasis statistik, SVM memiliki landasan teori yang dapat dianalisa dengan jelas, dan tidak bersifat *black box*.

4. *Feasibility*

   SVM dapat diimplementasikan *relative* mudah, karena proses penentuan *support vector* dapat dirumuskan dalam *QP problem*. Dengan demikian jika kita memiliki *library* untuk menyelesaikan *QP problem*, dengan sendirinya SVM dapat diimplementasikan dengan mudah. Selain itu dapat diselesaikan dengan metode sekuensial sebagaimana penjelasan sebelumnya.

Disamping kelebihannya, SVM memiliki kelemahan atau keterbatasan, antara lain (Nugroho, 2003) :

1. Sulit dipakai dalam *problem* berskala besar. Skala besar dalam hal ini dimaksudkan dengan jumlah sampel yang diolah.
2. SVM secara teoritik dikembangkan untuk *problem* klasifikasi dengan dua *class*. Dewasa ini SVM telah dimodifikasi agar dapat menyelesaikan masalah dengan *class* lebih dari dua, antara lain strategi *One versus rest* dan strategi *Tree Structure*. Namun demikian, masing-masing strategi ini memiliki kelemahan, sehingga dapat dikatakan penelitian dan pengembangan SVM

pada *multiclass-problem* masih merupakan tema penelitian yang masih terbuka.

2.4.4 Multiclass Support Vector Machine (SVM)

Ada dua pilihan untuk mengimplementasikan *multiclass* SVM yaitu dengan menggabungkan beberapa SVM biner atau menggabungkan semua data yang terdiri dari beberapa kelas ke dalam sebuah bentuk permasalahn optimal. Namun pada pendekatan yang kedua permasalahn optimasi yang harus diselesaikan jauh lebih rumit. Berikut ini adalah metode yang umum digunakan untuk mengimplementasikan *multiclass SVM* dengan pendekatan yang pertama:

1. *Metode one-against-all* (satu lawan semua)

   Dengan menggunakan metode ini, dibagun k buah model SVM biner (k adalah jumlah kelas)
2. Metode *one-against-one* (satu lawan satu)

   Dengan menggunakan metode ini, dibagun k(k-1)/2 buah model klasifikasi biner (k adalah jumlah kelas). Terdapat beberapa metode untuk melakukan pengujian setelah keseluruhan k(k-1)/2 model klasifikasi selesai dibagun. Salah satunya adalah metode *voting* (Santosa, 2007).

2.5 Analisis Regresi *Logistic*

Menurut Hosmer dan Lemeshow (2000), regresi logistik adalah suatu metode yang dapat digunakan untuk mencari hubungan antara variabel respon yang bersifat *dichotomus* (skala nominal/ordinal dengan dua kategori) dengan satu atau lebih variabel prediktor berskala kategori atau kontinu. Model regresi logistik terdiri dari regresi logistik dengan respon biner, ordinal, dan multinomial.

Regresi logistik biner adalah suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (y) yang bersifat biner (*dichotomus*) dengan variabel prediktor (x) yang bersifat kategorik atau kontinu. Model regresi logistik biner digunakan untuk menganalisis hubungan antara satu variable respon dan beberapa variabel prediktor, dengan variabel responnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan

sebuah karakteristik dan bernilai 0 untuk menyatakan ketidakberadaan sebuah karakteristik (Tampil dkk, 2017).

Asumsi yang harus dipenuhi dalam Regresi Logistik antara lain:

1. Regresi logistik tidak membutuhkan hubungan linier antara variabel independen dengan variabel dependen.
2. Variabel independen tidak memerlukan asumsi *multivariate normality.*
3. Asumsi homokedastisitas tidak diperlukan.
4. Variabel bebas tidak perlu diubah ke dalam bentuk metrik (interval atau skala ratio).
5. Variabel dependen harus bersifat dikotomi (2 kategori, misal: tinggi dan rendah atau baik dan buruk.
6. Variabel independen tidak harus memiliki keragaman yang sama antar kelompok variable.
7. Kategori dalam variabel independen harus terpisah satu sama lain atau bersifat eksklusif.
8. Sampel yang diperlukan dalam jumlah relatif besar, minimum dibutuhkan hingga 50 sampel data untuk sebuah variabel prediktor (independen).
9. Regresi logistik dapat menyeleksi hubungan karena menggunakan pendekatan non linier log transformasi untuk memprediksi odds ratio. Odd dalam regresi logistik sering dinyatakan sebagai probabilitas.

Sebagaimana metode regresi biasa, Regresi Logistik dapat dibedakan menjadi 2, yaitu:

1. *Binary Logistic Regression* (Regresi Logistik Biner).
2. Regresi Logistik biner digunakan ketika hanya ada 2 kemungkinan variabel respon (Y), misal membeli dan tidak membeli. *Multinomial Logistic Regression* (Regresi Logistik Multinomial).
3. Regresi Logistik Multinomial digunakan ketika pada variabel respon (Y) terdapat lebih dari 2 kategorisasi.

Pada regresi logistik dapat disusun model yang terdiri dari banyak variabel prediktor, dikenal sebagai model multivariabel. Rata-rata bersyarat dari y jika diberikan nilai x adalah $\pi(x) = E(y \mid x)$. Model regresi logistik multivariabel dengan p variabel prediktor adalah sebagai berikut.

$$\pi(x)\frac{e^{\beta_0+\beta_1 x_{1+\cdots+\beta_p x_p}}}{1+e^{\beta_0+\beta_{1x_1}+\cdots+\beta_p x_p}}$$

Dimana, *p* = Banyaknya variabel predictor

Dengan menggunakan transformasi logit dari π (*x*) untuk mempermudah pendugaan parameter regresi yang dirumuskan sebagai berikut

$$\{\pi(x)\}\{1+e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}\}=e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}$$

$$\{\pi(x)\}\{\pi(x)e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}\}=e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}$$

$$\pi(x)=e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}-\pi(x)e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}$$

$$\pi(x)=\{1-\pi(x)\}e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}$$

$$\frac{\pi(x)}{1-\pi(x)}=e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}$$

$$ln\left(\frac{\pi(x)}{1+\pi(x)}\right)=\ln e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}$$

$$ln\left(\frac{\pi(x)}{1+\pi(x)}\right)=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p$$

Sehingga diperoleh persamaan sebagai berikut:

$$g(x)=\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p$$

g(*x*) disebut dengan fungsi logit model regresi logistik biner dengan *p* variabel prediktor. Model regresi logistik pada persamaan diatas dapat dituliskan dalam bentuk :

$$\pi(x)=\frac{\exp(g(x))}{1+\exp(g(x))}$$

2.5.1 Pendugaan Parameter Regresi Logistik

Pendugaan parameter dalam model regresi logistik dilakukan dengan menggunakan metode kemungkinan maksimum atau *Maximum Likelihood Estimation (MLE)* yaitu diperoleh dengan menurunkan fungsi kepekatan peluang bersama (Hosmer and Lemeshow 1989).

Fungsi *likelihood* memberikan kemungkinan mengamati data sebagai fungsi dari parameter yang tidak diketahui. MLE dipillih untuk memaksimalkan nilai fungsi tersebut. Estimasi maksimum likelihood merupakan pendekatan dari

estimasi Weighted Least Square, dimana matrik pembobotnya berubah setiap iterasi. Proses menghitung estimasi maksimum likelihood ini disebut juga sebagai Iteratif Reweighted Least Square.

Cara yang sesuai untuk kontribusi fungsi likelihood untuk setiap pengamatan (x_{1i} y_i) adalah sebagai berikut.

$$f(Y = yi) = \pi(xi)\, yi\, [(1 - \pi(xi))]1 - yi\,,\ yi = 0,1$$

Fungsi likelihood yang diperoleh dengan pengamatan yang diasumsikan independen adalah sebagai berikut.

$$I(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i]^{1-y_i}$$

dimana:

y_i = Pengamatan pada variable ke i

$\pi(x_i)$ = Peluang untuk variable predictor ke-i

Untuk memudahkan perhitungan maka dilakukan pendekatan log likelihood, didefinisikan sebagai:

$$L(\beta) = \sum_{i=1}^{n} \{y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]\}$$

Untuk mendapatkan nilai penafsiran koefisien regresi logistik $(\widehat{\beta})$ dilakukan dengan membuat turunan pertama $L(\beta)$ terhadap β dan disamakan dengan 0.

2.5.2 Pengujian Parameter Model Regresi Logistik Biner

Pengujian estimasi parameter merupakan pengujian yang digunakan untik menguji signifikansi koefisien β dari model. Pengujian ini dapat menggunakan uji secara serentak maupun parsial.

1. Uji Serentak

Pengujian serentak dilakukan untuk memeriksa signifikansi koefisien β secara keseluruhan (Hosmer & Lemeshow, 2000) dengan hipotesis sebagai berikut.

$H_0 : \beta_j = \beta_{2=\cdots=}\beta_P = 0$

H_1 : minimal ada satu $\beta_j \neq 0$; $j = 1,2,3, \ldots \ldots p$

$$G = -2LN\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{0}}{\prod_{i=1}^{n}\widehat{\pi}_i^{\widehat{y_i}}\,(1-\widehat{\pi}_i)^{(1-y_i)}}\right]$$

Daerah Penolakan : Tolak H0 jika G >χ2(p,α)

Keterangan :

n_0 = jumlah pengamatan dengan kategori $y = 0$

n_1 = jumlah pengamatan dengan kategori $y = 1$

n = jumlah pengamatan

P = banyaknya parameter

Jika terdapat *k* kategori pada suatu variable predictor, maka kontribusi untuk derajat bebas pada uji Likelihood adalah sebesar *k-1* (Homsmer & Lemeshor, 2000).

2. Uji Parsial

Pengujian secara parsial dilakukan untuk mengetahui signifikansi setiap parameter terhadap variabel respon. Pengujian signifikansi parameter menggunakan uji Wald (Hosmer & Lemeshw, 2000) dengan hipotesis sebagai berikut;

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0 \;; j = 1,2,3, \ldots \ldots p$

$$W = \frac{\hat{\beta}_J}{SE(\hat{\beta}_J)}$$

Daerah Penolakan : Tolak H0 jika W >Zα/ 2

Statistik uji W tersebut juga disebut sebagai statistika uji Wald dengan SE($\hat{\beta}$ j) adalah taksiran standart error parameter.

3. Uji Kesesuaian Model

Pengujian ini dilakukan untuk menguji apakah model yang dihasilkan berdasarkan regresi logistik multivariat/serentak sudah layak. Pengujian ini menggunakan statistik uji Hosmer dan Lemeshow (Hosmer & Lemeshow, 2000) dengan hipotesis yang digunakan sebagai berikut.

H0 : Model sesuai (tidak terdapat perbedaan yang signifikan antara hasil pengamatan dengan kemungkinan hasil prediksi model)

H1 : Model tidak sesuai (terdapat perbedaan yang signifikan antara hasil pengamatan dengan kemungkinan hasil prediksi model).

Statistik Uji

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k \overline{\pi}_k)^2}{n_k \overline{\pi}_k (1 - \overline{\pi_k})}$$

Daerah Penolakan : Tolak H0 jika Cˆ >χ2(g–2,α)

Keterangan :

Ok : observasi pada grup ke-k

$\overline{\pi}_k$: rata-rata taksiran peluang $\sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n_k}$

g : jumlah grup (kombinasi kategori dalam model serentak)

n_k' : banyaknya observasi pada grup ke-k

g : banyaknya kategori semua variabel predictor

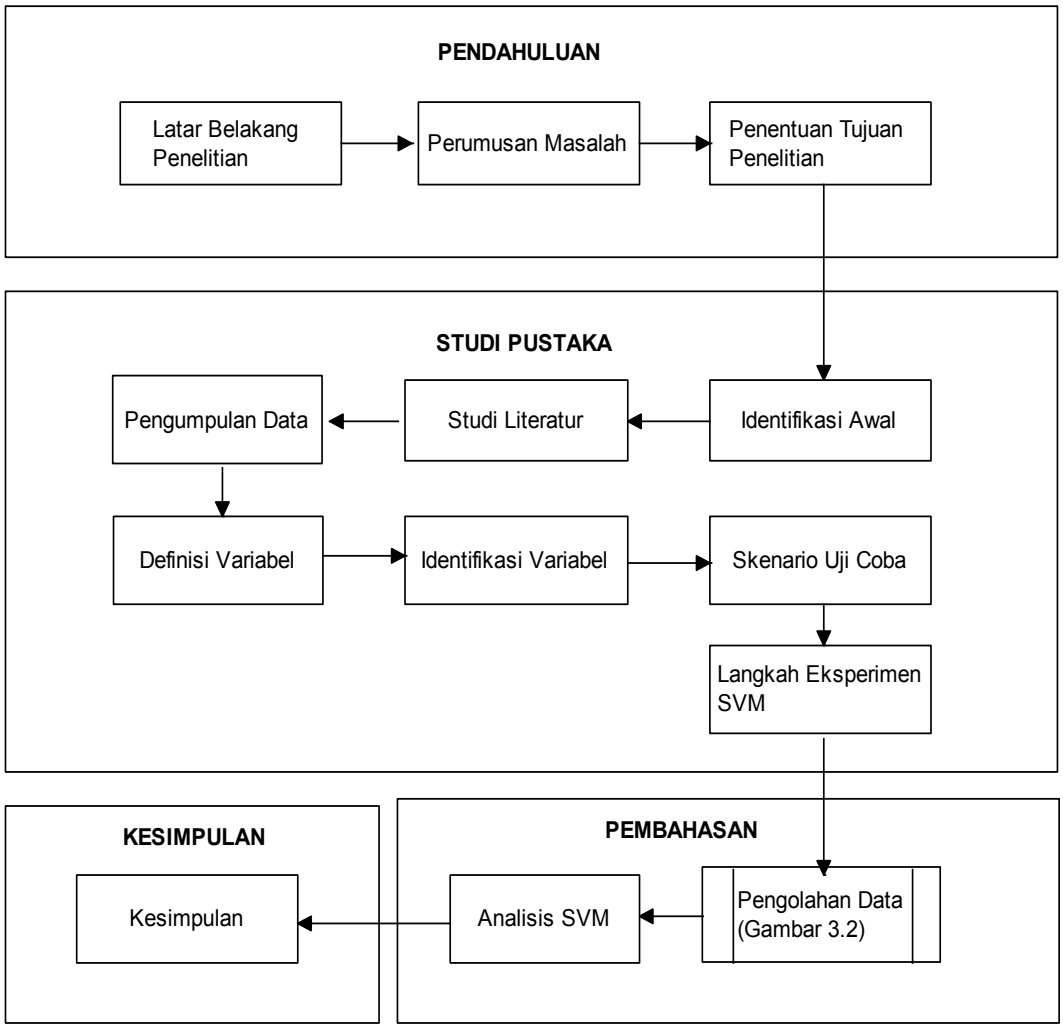
(Halaman ini sengaja dikosongkan)

BAB III
METODOLOGI PENELITIAN

Bab ini menjelaskan mengenai alur/tahapan pelaksanaan penelitian dan penulisan laporan penelitian, diagram alur, jenis sumber data, metode pengumpulan data, metode pengolahan data, dan jadwal rencana kegiatan penelitian.

3.1 Diagram Alir Penelitian

Untuk memperjelas alur dalam penelitian ini, maka dibuat sebuah rancangan penelitian yang akan dijadikan pedoman pengerjaan sampai tercapainya hasil yang diharapkan. Diagram alir tersebut dapat dilihat pada gambar 3.1 berikut:



Gambar 3.1 Diagram Alir Penelitian.

Pada gambar 3.1 terlihat bahwa diagram alir pada penelitian ini dapat dijelaskan secara detail sebagai berikut:

3.1.1 Pendahuluan

Pada pendahuluan ini akan dibahas hal-hal yang dilakukan dalam penelitian ini sebelum melakukan percobaan yang terdiri dari dari latar belakang, rumusan masalah dan penentuan tujuan penelitian yang telah dirumuskan.

3.1.1.1 Latar Belakang

Pada latar belakang ini dibahas hal-hal yang menjadi dasar atau titik tolak yang melatar belakangi dalam penelitian ini untuk disampaikan kepada pembaca mengenai kondisi *real* adanya *churn rate* pada perusahaan telekomunikasi dengan berdasarkan data. Segmentasi pelanggan yang selalu bergerak secara dinamis dan berpindah dari satu operator ke operator lain juga melatar belakangi adanya penelitian ini.

3.1.1.2 Rumusan Masalah

Pada rumusan masalah dibahas tentang masalah yang diangkat pada penelitian ini, rumusan masalah didasarkan dari latar belakang yang telah dijelaskan dan selanjutnya digunakan untuk menjelaskan masalah atau isu yang dibahas untuk disampaikan kepada para pembaca.

3.1.1.3 Penentuan Tujuan Penelitian

Dalam penentuan tujuan penelitian ini menjelaskan adanya sesuatu yang ingin dicapai dari adanya penelitian ini, meliputi kontribusi nyata bagi PT Telekomunikasi XYZ secara khusus dan perusahaan telekomunikasi pada umumnya dalam memaintain *subscriber* dengan meminimalisir adanya *churn rate*.

3.1.2 Studi Pustaka

Studi pustaka merupakan pembelajaran awal penelitian sampai dilakukannya uji coba penelitian melalui berbagai sumber yang terdiri dari *marketing discussion*, *studi literature* secara khusus yang meliputi penelitian terkait berupa jurnal-jurnal *data mining*, teknik SVM, *churn prediction* dan analisis regresi, pengumpulan data, definisi variabel, identifikasi variabel baik input maupun output.

3.1.2.1 Identifikasi Awal

Identifikasi awal ini dilakukan untuk mengetahui variabel-variabel yang berpengaruh pada *churn rate* serta penentuan variabel yang akan diambil untuk dijadikan variabel input pada percobaan nantinya sesuai dengan data yang *available* di perusahaan.

3.1.2.2 Studi Literatur

Studi literatur dimaksudkan mencari referensi secara khusus yang berkaitan dengan teori yang relevan dengan kasus atau permasalahan yang ditemukan dalam hal ini *churn rate,* SVM dan Regresi. Teknik ini dilakukan dengan tujuan untuk mengungkapkan berbagai teori-teori yang relevan dengan permasalahan yang sedang dihadapi/diteliti sebagai bahan rujukan dalam pembahasan hasil penelitian.

3.1.2.3 Pengumpulan Data

Pengumpulan data dilakukan untuk mendapatkan data uji coba yang dapat menjawab hipotesis yang dibuat. Pada data ini akan dijelaskan penentuan populasi data dan sampel, jenis-jenis sumber data, metode pengumpulan data serta identifikasi variabel-variabel data.

a. Penentuan populasi data dan Sampel

Populasi adalah wilayah generalisasi yang terdiri atas: obyek/subyek yang mempunyai kualitas dan karakteristik tertentu yang ditetapkan oleh peneliti untuk dipelajari dan kemudian ditarik kesimpulannya Sugiyono (2010: 117). Penyelidikan yang menggunakan hipotesis nol, akan berhadapan dengan masalah populasi dan sampel, karena penelitian selalu berhubungan dengan sekelompok subyek, gejala, nilai tes benda-benda ataupun peristiwa (Surakhmad, 2004).

Arikunto (2010: 174) menyatakan bahwa “sampel adalah sebagian atau wakil populasi yang diteliti”. Selanjutnya, untuk menentukan banyaknya sampel yang akan digunakan Arikunto (2010: 120) menjelaskan bahwa, “Apabila subjeknya kurang dari 100 lebih baik diambil semua hingga penelitiannya merupakan penelitian populasi. Selanjutnya jika jumlah subjeknya besar dapat diambil antara 10-15 %, atau 20-25% atau lebih”.

Populasi dalam penelitian ini adalah pengguna layanan jasa telekomunikasi di salah satu perusahaan telekomunikasi di Indonesia dengan area meliputi Jawa Timur, Bali Nusa yang tercatat sebagai pengguna aktif minimal 3 bulan terakhir. Teknik pengambilan sampel yang digunakan dalam penelitian ini adalah metode *purposive sampling,* yaitu pemilihan sampel yang dilakukan secara acak berdasarkan karakteristik tertentu yang telah ditetapkan sebelumnya. (Sugiyono, 2010).

Uma Sekaran (2006) memberikan acuan umum untuk menentukan ukuran sampel :

1. Ukuran sampel lebih dari 30 dan kurang dari 500 adalah tepat untuk kebanyakan penelitian.
2. Jika sampel dipecah ke dalam sub sampel (pria/wanita, junior/senior, dan sebagainya), ukuran sampel minimum 30 untuk tiap kategori.
3. Dalam penelitian mutivariate (termasuk analisis regresi berganda), ukuran sampel sebaiknya 10x lebih besar dari jumlah variabel dalam penelitian.
4. Untuk penelitian eksperimental sederhana dengan kontrol eskperimen yang ketat, penelitian yang sukses adalah mungkin dengan ukuran sampel kecil antara 10 sampai dengan 20.

Adapun sampel yang akan dilakukan uji coba sebanyak 1000 data dengan 600 data training dan 400 data testing di mana masing-masing terdapat data *churn* dan *non churn*. Adapun pertimbangan jumlah data training lebih banyak dibandingkan dengan data testing adalah agar lebih memaksimalkan fungsi algoritma, karena semakin banyak data testing, maka semakin akurat data yang dihasilkan.

b. Sumber Data

Sumber data penelitian yaitu sumber subjek dari tempat mana data bisa didapatkan. Jika penelitian merupakan studi kasus, maka rata-rata sumber data berasal dari internal instansi studi kasus. Namun dalam penelitian ini, terdapat dua jenis data yang digunakan yaitu data primer dan data sekunder sebagai berikut:

1) Data primer pada penelitian ini adalah data yang didapatkan secara langsung dari sumber perusahaan tanpa adanya pihak ketiga. Data primer disebut juga sebagai data asli atau data baru yang memiliki sifat *up to date.* Dalam hal ini data diperoleh dari *source* database internal perusahaan.
2) Data Sekunder, adalah data yang mengacu pada informasi yang dikumpulkan dari sumber yang telah ada. Sumber data sekunder adalah catatan atau dokumentasi perusahaan, publikasi pemerintah, analisis industri oleh media, situs Web, internet dan seterusnya (Sekaran, 2011). Pada penelitian ini, data sekunder adalah data yang didapatkan dari hasil kajian pustaka (peneliti sebagai tangan kedua) meliputi jurnal, buku, internet, dan literature yang relevan dengan penelitian ini.

c. Waktu Pengumpulan Data

Dalam pengumpulan data, pada penelitian ini dilakukan dengan menggunakan *Time Series*. *Time Series*, merupakan pengumpulan data yang secara berkala dari waktu ke waktu untuk menggambarkan perkembangan suatu keadaan atau kecenderungan sebuah data. data yang diambil adalah berbentuk data *time series* di mana terdapat *history* penggunaan layanan voice, sms dan data selama 3 bulan terkahir pada tahun 2018 dalam jumlah total (akumulasi per bulan).

3.1.2.4 Definisi Variabel

Defnisi variabel merupakan variabel yang digunakan untuk *pre-defined* data dalam mempermudah peneliti untuk mengumpulkan data. Variabel in akan dijadikan acuan dalam melakukan analisis faktor-faktor layanan dan demografi yang akan berpengaruh pada *churn rate* pelanggan. Variabel ini diambil berdasarkan studi *literature* (Shaaban, 2012) serta diskusi dengan divisi *marketing* terkait variabel yang tersedia pada *source* data yang dapat berpengaruh pada *churn rate*. Variabel dan definisi operasional variabel pada penelitian ini dituangkan pada Tabel 3.1.

Tabel 3.1 Variabel dan Definisi Operasional Variabel

No	Variabel	Definisi Operasional	Indikator
1	MSISDN	*Mobile Subscriber Integrated Services Digital Network Number* yang digunakan oleh *customer*	*No Handphone* pengguna layanan
2	POC	Merupakan nama tempat di mana MSISDN digunakan	Nama tempat MSISDN digunakan
3	*Handset*	*Device barier* yang digunakan pada perangkat pengguna	3G/4
4	*Voice Minutes*	V*oice call* untuk melakukan panggilan telepon pelanggan baik sesama operator atau operator lain	*Outcoming call* dalam jumlah menit
5	*SMS Event*	*Short message service* untuk mengirim pesan antar pengguna layanan *provider*	Penggunaan layanan SMS dalam jumlah *event*
6	Data *kb*	Merupakan layanan yang digunakan untuk akses inter *connection-networking*	Penggunaan layanan data dalam jumlah *kb*
7	Packet Name	Nama Paket yang digunakan oleh customer khususnya di layanan mobile data	Nama Paket yang digunakan user
8	Reload	Jumlah pulsa yang ada pada MSISDN customer untuk dilakukan penggunaan layanan	Jumlah Pulsa di MSISDN

3.1.2.5 Identifikasi Variabel

Dalam melakukan identifikasi variabel dimaksudkan untuk mengetahui variabel-variabel yang cukup berpengaruh pada faktor *churn rate* yang

menjadi topik dalam penelitian ini. Adapun variabel yang dimaksud terdiri dari variabel *input* dan variabel *output*.

a. Variabel *Input*

Variabel *input* merupakan variabel yang keberadaannya dapat mempengaruhi variabel *output* dan dapat mengeluarkan suatu respon. Variabel *input* keberadaanya tidak bergantung pada variabel lain, atau biasa disebut dengan variabel independen. Dalam penelitian ini, variabel *input* yang akan diuji terdiri dari variabel *numeric* dan *character* (text). Jika variabel *input* terdiri dari *text*, maka untuk dapat dilakukan uji dengan *Support Vectore Machine* (SVM) harus dilakukan konversi menjadi *numeric*. Pada table 3.1 merupakan variabel-variabel yang penggunaanya perlu dilakukan konversi.

Tabel 3.2 Variabel tipe *numeric*

No	Nama Variabel
1	MSISDN
2	*Voice Minutes*
3	*SMS Event*
4	Data kb

Sedangkan variabel yang memerlukan konversi menjadi *numeric* terdapat pada tabel 3.3 di bawah dengan jumlah empat variabel. Dari keempat ini membutuhkan kategorikal *value* sebelum dilakukan konversi ke *numeric*.

Tabel 3.3 Variabel tipe *text* untuk konversi *numeric*

No	Nama Variabel	***Range* Konversi *(Numeric)***
1	POC	1-22
2	*Handset*	1-2
3	*Nama Paket*	1-2

Pada tabel 3.3 terlihat bahwa terdapat empat variabel yang membutuhkan konversi menjadi *numeric*. Konversi *numeric* ini dilakukan dengan perhitungan angka mulai dari 1 sampai dengan jumlah kategori teks yang merupakan *value* dari variabel. Pada POC, terdapat *value* variabel kemungkinan sebanyak 22 POC yang berada di area sampel sehingga *range* konversi 1 sampai 22. Sedangkan untuk *value* pada *handset* terdapat dua kategori nilai *handset,* yaitu 3G dan 4G, konversi *numeric* yang dilakukan adalah angka 2 dan 2. Adapun untuk variabel nama paket mempunyai nilai nama paket yang digunakan user sehingga rangenya antara 1 sampai jumlah paket yang ada.

b. Variabel *Output*

Variabel *output* merupakan keluaran yang diharapkan dari adanya variabel *input*. Variabel *output* dihasilkan melalui proses pemodelan SVM yang telah dibahas sebelumnya. Variabel *output* ini juga bisa disebut variabel dependen yang berarti bahwa adanya variabel *output* bergantung pada adanya variabel *input*. Keluaran yang dihasilkan sangat tergantung pada pemilihan antribut yang akan dilakukan *testing.* Pada penelitian ini, *output* yang diharapkan adalah tingkat level dari variabel yang diuji dalam mempengaruhi *churn rate*.

Tabel 3.4 Variabel *output* analisis faktor *churn*

Level	Nama Variabel	Hasil Performance SVM (sampel)
Level 1	POC	90%
Level 2	*Voice Minutes*	85%
Level 3	Data kb	80%
Level 4	*SMS Event*	79%
Level 5	*Handset*	75%
Level 6	Nama Paket	70%
Level 7	Reload	65%

Pada tabel 3.4 merupakan variabel keluaran yang diharapkan setelah dilakukan uji coba. *Output* ini sangat bergantung pada variabel *input*.

Semakin kecil besar nilai *performance* yang dihasilkan, semakin bagus klasifikasi yang dihasilkan. Artinya pada penelitian ini faktor-faktor yang mempengaruhi *churn rate* ditentukan oleh seberapa besar nilai *performance* pada masing-masing variabel *output.* Pada tabel 3.4 terlihat POC merupakan faktor yang mempengaruhi *churn rate* teratas dengan nilai performance 90%. Disusul dengan penggunaan *voice* dalam *minutes* dengan nilai performance 85%, sedangkan faktor ketiga yang paling berpengaruh adalah penggunaan jumlah data dalam kb yang mencapai 80%.

3.1.2.6 Skenario Uji Coba

Pada bagian skenario uji coba ini dijelaskan bagaimana skenario yang dilakukan dalam penelitian ini untuk mengetahui kemampuan algoritma dalam memperoleh klasifikasi yang baik dan dapat menjelaskan variabel apa yang paling berpengaruh terhadap *churn rate* pelanggan.

Dataset yang digunakan sebanyak 1000 dataset di mana data dibagi menjadi dua bagian, yaitu 600 *data training* dan 400 untuk *data testing*. Data akan diuji dengan tiga skenario berikut:

1. Uji performa metode SVM dengan mengubah tipe kernel yang digunakan sebagai perbandingan. Beberpa kernel tersebut antara lain *Polynomial kernel* dan *Gaussian radial basis function* (RBF) kernel. Pemilihan kernel ini dilakukan untuk mengetahui tipe kernel yang paling sesuai dengan dataset yang diuji.
2. Uji perubahan atribut dataset, setiap atribut mempunyai nilai yang berbeda. Dengan pemilihan atribut, maka akan menghasilkan akurasi yang berbeda, sehingga masing-masing atribut mempunyai nilai dalam mempengaruhi prediksi *churn rate*. Tujuan dari skenario ini adalah untuk mengetahui akurasi pengaruh masing-masing variabel terhadap prediksi *churn rate*.
3. Uji regresi pada masing-masing variabel untuk mengetahui seberapa besar pengaruh variabel bebas terhadap variabel tetap. Masing-masing variabel mempunyai nilai tersendiri di mana nilai akhir yang akan dijadikan sentimen apakah variabel bebas berpengaruh secara signifikan

terhadap variabel tetap dengan *threshold* atau tingkat kepercayaan sebesar 0.05 (Fraticasari, 2018). Tujuan dari uji regresi ini adalah untuk mengetahui variabel apa saja yang secara signifikan berpengaruh dan tidak berpengaruh terhadap terjadinya *churn rate*.

3.1.2.7 Langkah Eksperimen SVM

Langkah eksperimen dengan SVM ini dilakukan untuk memahami cara kerja SVM secara mendalam. Eksperimen dilakukan dengan menggunakan perhitungan pada Microsoft excel dengan data dummy sebagai ujicoba data. Detail eksperimen sebagai berikut:

1. Data Uji

   Data uji ini terdiri dari dua atribut yaitu x1 dan x2 dan terdapat dua kelas 1 dan -1 sebagai berikut:

   Tabel 3.5 Data uji Eksperimen SVM

X1	X2	y
2.947814	6.626878	1
2.530388	7.785050	1
3.566991	5.651046	1
3.156983	5.467077	1
2.582346	4.457777	-1
2.155826	6.222343	-1
3.273418	3.520687	-1

2. Constraint

   Setelah diketahui data uji, maka langkah selanjutnya adalah menentukan constraint untuk kedua atribut dengan menggunakan rumus:

   $$y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right).$$

   Di mana nilai b awal adalah 0, sehingga didapat hasil constraint sebagai berikut:

Tabel 3.6 Mencari nilai constraint

w1.x1	w2.x2	constraint
8.560183	7.960593	1.785812
7.348017	9.351858	1.96491
10.35822	6.788367	2.41162
9.167592	6.567373	1
7.498897	5.354943	1.881125
6.260322	7.474642	1
9.50571	4.229255	1

Nilai constraint ini yang nanti akan digunakan untuk mencari nilai w1, w2 dan b.

3. Cari nilai w1,w2 dan b

   Setelah didapati nilai constraint, langkah selanjutnya adalah dengan cara mencari nilai w1, w2 dan b dengan menggunakan rumus berikut:

$$\min_{w,b} \frac{1}{2} \| w \|^2$$

subject to

$$y_i (wx_i + b) \geq 1, \qquad i = 1,...m$$

   Pada rumus di atas, untuk menghitung nilai quadratic programming digunakan add-ins solver pada excel, sehingga penulis tidak lagi menghitung secara manual. Sehingga didapati nilai w1, w2 dan b sebagai berikut;

Tabel 3.7 Nilai w1, w2 dan b

w1	w2	b
2.903909	1.201258	-14.735

4. Scoring

   Setelah didapati hasil w1, w2 dan b, maka langkah selanjutnya adalah menentukan score pada masing-masing atribut dengan menggunakan rumus scoring berikut:

$$\Delta_i = \mathbf{w}^T \mathbf{x}_i + b$$

   Di mana b adalah nilai dari b yang baru. Sehingga didapati hasil scoring sebagamana tabel 3.8.

Tabel 3.8 Nilai score pada masing-masing data uji

Score
1.785812
1.96491
2.41162
1
-1.88113
-1
-1

Setelah didapati nilai score, maka langkah selanjutnya adalah dengan melakukan klasifikasi dengan rumus berikut:

$$sign(x) = \begin{cases} +1 & if\ x \geq 0 \\ -1 & if\ x < 0 \end{cases}$$

Artinya jika setelah dilakukan sign score hasil lebih besar sama dengan 0, maka termasuk klasifikasi nilai +1, sedangkan jika lebih kecil dari 0, maka masuk pada klasifikasi -1 sebagai berikut:

Tabel 3.9 Hasil klasifikasi

Score	Classification
1.785812	1
1.96491	1
2.41162	1
1	1
-1.88113	-1
-1	-1
-1	-1

5. Confusion Table

Nilai klasifikasi sebagaimana tabel 3.9. dicocokkan dengan data asli kelas awal untuk menentukan seberapa akurasi metode SVM ini dalam menentukan klasifikasi sebagai berikut:

Tabel 3.10 Confusion Table

Prediction			
True Class	1	-1	**sum**
1	4	0	4
-1	0	3	3
Sum	**4**	**3**	**7**

Tabel 3.11 Prediction Table

True Class	1	2	Prediction
1	100.00%	0.00%	100%
2	0.00%	100.00%	100%
Percent correctly predicted			**100.00%**

Pada tabel 3.11 terlihat bahwa hasil klasifikasi 100% akurasi, di mana terdapat 4 untuk kelas +1 dan 3 untuk kelas -1.

6. Hyperplane

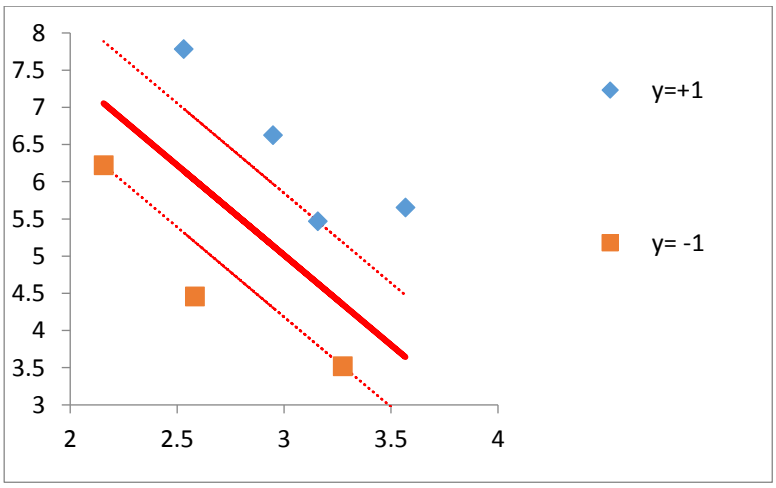
Langkah selanjutnya setelah dilakukan scoring adalah dengan mencari nilai hyperplane dengan menggunakan rumus berikut:

$w_1 x_1 + w_2 x_2 + b = 0$.

Tabel 3.12 Hyperplane

Decision	Margin-	Margin+
x2	x2	x2
5.14026	4.3078	5.972720605
6.149341	5.316881	6.981801186
3.643468	2.811008	4.475928658
4.634617	3.802157	5.467077143
6.023739	5.191279	6.856199375
7.054803	6.222343	7.887263608
4.353147	3.520687	5.185607597

Pada tabel 3.12 setelah dilakukan plot dua kelas akan terbentuk grafik 3.1 berikut:



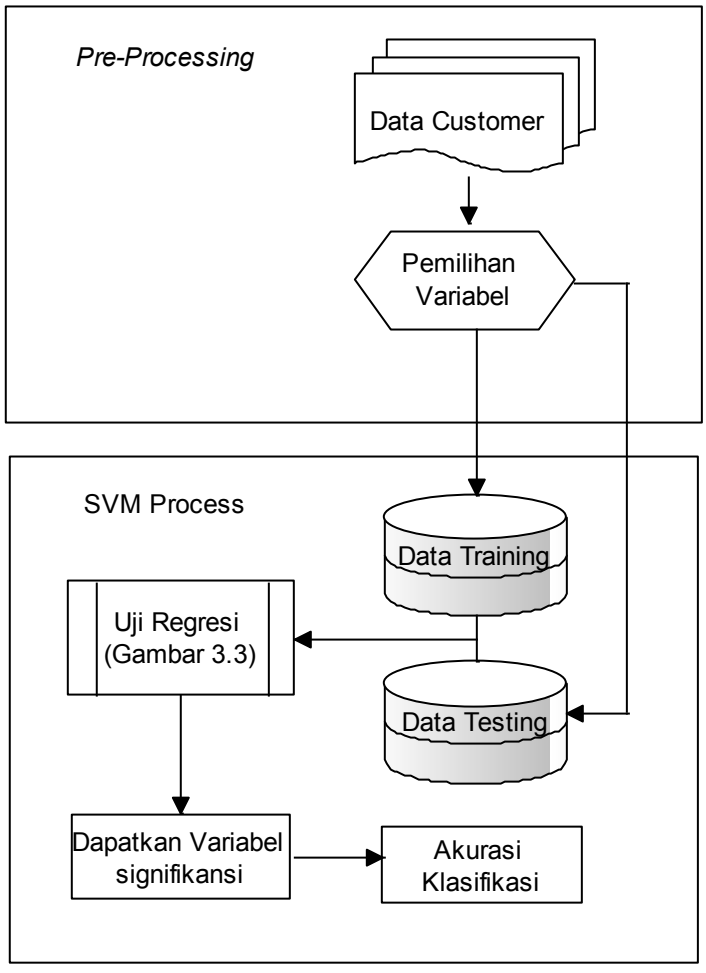
Grafik 3.1 Hyperplane

3.1.3 Pembahasan

Pada pembahasan ini nantinya akan dijelaskan secara detail penelitian ini meliputi pengolahan data, uji coba penelitian, prediksi *churn rate* dengan SVM, serta analisa variabel-variabel input yang mempunyai secara signifikan terhadap *churn rate* dengan SVM.

3.1.3.1 Pengolahan Data

. Pengolahan data dilakukan dengan SVM (gambar 3.2) untuk mengetahui hasil prediksi, dan variabel-variabel yang berpengaruh signifikan terhadap *churn rate* (gambar 3.3).



Gambar 3.2 Pengolahan data Regresi

Sebagaimana gambar 3.2 terlihat bahwa data yang telah dikumpulkan, akan dilakukan pemilihan variabel yang nantinya akan dilakukan ujicoba. Pada awal percobaan, semua variabel akan dimasukkan untuk mengetahui akurasi awal di mana MSISDN menjadi x1 dan variabel input lainnya

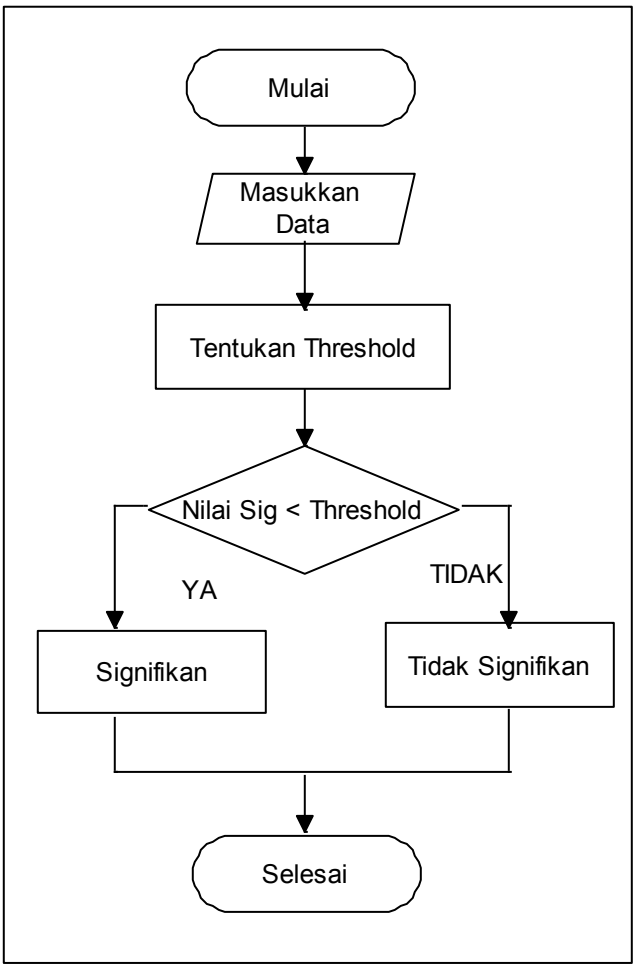
menjadi x2 dan seterusnya. Sedangkan pada variabel *churn* merupakan label untuk nilai b. Setelah diketahui akurasi awal, maka langkah selanjutnya adalah dengan menguji tiap-tiap variabel. Hal ini dilakukan untuk mengetahui seberapa besar pengaruh masing-masing variabel terhadap *churn rate.* Dalam pengujian ini akan terdapat dua variabel uji dan 1 vriabel sebagai label. Variabel uji tersebut adalah MSISDN sebagai x1 dan variabel input sebagai x2 sedangkan untuk variabel churn sebagai label nilai b.

Data ujicoba baik untuk keseluruhan variabel atau masing-masing variabel dengan dibagi dua kategori data yaitu *training* dan *testing*. Untuk data *training* dan *testing* berupa data *history* pengguna layanan yang sudah menjadi *churn* dan *non churn*, untuk *non churn* data diambil dari pengguna layanan yang aktif minimal 3 bulan dengan variabel yang sama dengan data *training*.

Kemudian kedua data *training* dan *testing* akan dilakukan ujicoba dengan pemodelan Regresi Logistik (gambar 3.3) menggunakan SPSS dengan teknik *supervised learning*, yakni teknik pendekatan *data mining* dengan membuat model yang mengacu pada *dataset history* sebelumnya untuk dilihat pola data sehingga bisa diimplementasikan pada data uji untuk mendapatkan nilai signifikansi. Variabel ini yang nantinya akan dijadikan analisis dalam menentukan model yang dihasilkan.

Pada gambar 3.3 merupakan cara kerja pemodelan regresi dimulai dari input data yang akan diuji, dari data tersebut kemudian dilakukan pemilihan atribut yang diinginkan untuk dilakukan uji coba. Pada awalnya akan dilakukan secara simultan untuk menentukan adanya pengaruh keseluruhan variabel terhadap churn, diikuti oleh uji secara parsial.

Pada saat uji parsial, perlu menentukan nilai cutoff agar klasifikasi yang dihasilkan mencerminkan pada prediksi yang sesungguhnya. Dalam hal ini penentuan nilai cutoff klasifikasi sebesar 0.5.



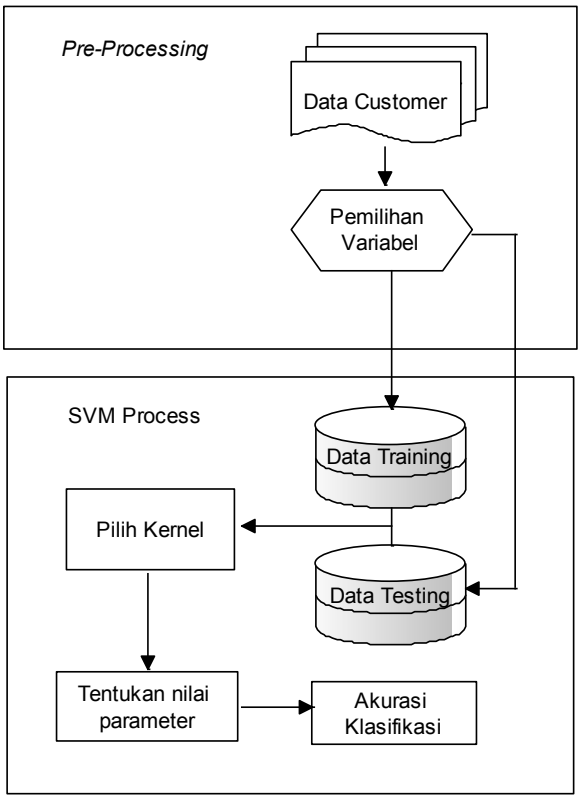
Gambar 3.3 Pemodelan SVM.

Tingkat signifikansi merupakan ambang batas yang digunakan untuk menentukan signifikansi. Jika nilai p kurang dari atau sama dengan tingkat signifikansi (threshold), data tersebut dianggap signifikan secara statistik. Sebagai aturan umum, tingkat signifikansi (alfa) ditetapkan sebesar 0,05 [1], berarti probabilitas kedua kelompok data tersebut 5%. Jika hasil ujicoba lebih kecil atau sama dengan *threshold,* maka variabel akan dikategorikan sebagai variabel signifkan, namun jika lebih besar, maka variabel dapat dikategorikan sebagai variabel yang tidak signifikan mempengaruhi variabel dependen.

3.1.3.2 Analisis SVM

Sebagaimaan pada gambar 3.4 bahwa analisis SVM dilakukan sebagaimana regresi logistik, namun terdapat perbedaan pada pemilihan kernel yang biasa digunakan adalah polyinomial dan radial basis function. Serta penetuan parameter untuk mendapatkan hasil uji klasifikasi terbaik.

Setelah masing-masing fungsi kernel diujikan, maka akan didapati hasil terbaik yang akan dijadikan acuan klasifikasi.



Gambar 3.4 Pengolahan data SVM

3.1.4 Kesimpulan

Pada kesimpulan, akan dijelaskan hasil akhir yang didapatkan setelah dilakukan pengujian. Pada tahap ini merupakan proses untuk menarik kesimpulan dan saran atas apa yang dilakukan selama pengerjaan Penelitian. Dasar pengambilan kesimpulan dan saran diantaranya adalah hasil analisa dari pembahasan.

(Halaman ini sengaja dikosongkan)

BAB IV
HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan tentang hasil ujicoba dan pembahasan, meliputi distribusi data masing-masing variabel, uji hipotesis, dan pembahasan hipotesis.

4.1 Distribusi Data

Dalam distribusi data ini, tujuan utama adalah untuk memberikan gambaran detail tentang penyebaran data yang telah dilakukan uji coba pada penelitian ini. Pada setiap variabel, dijelaskan masing-masing sebaran data, baik berupa kategorikal atau numerik.

4.1.1 Distribusi data POC

POC merupakan *point of control* berupa daerah yang menjadi penyebaran pengguna layanan pada PT Telekomunikasi XYZ. Variabel POC ini terdiri dari 22 POC yang ada di *East Region* meliputi Jawa Timur dan Bali Nusa.

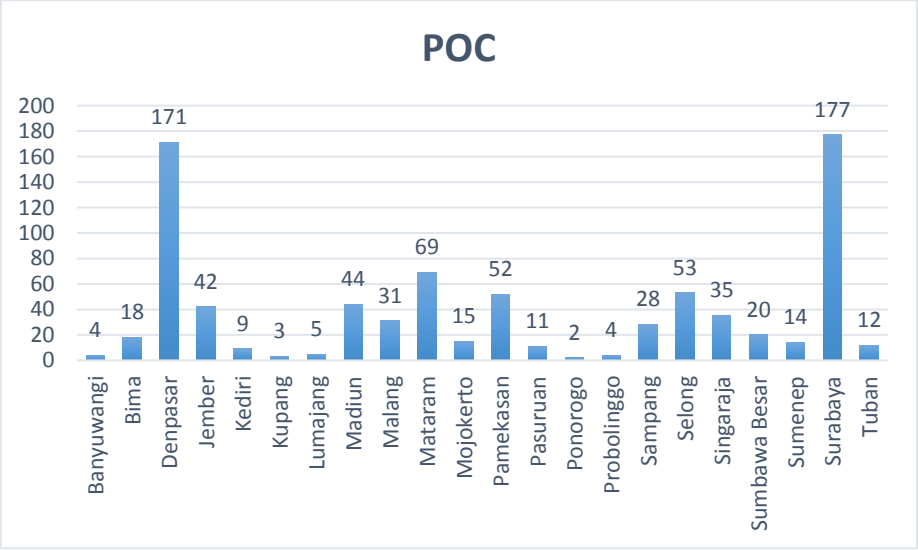


Diagram 4.1 Distbusi data POC

Sebagaimana digambarkan pada grafik 4.1 bahwa pengujian data POC didominasi oleh dua kota besar, yaitu Surabaya dan Denpasar di mana jumlah sampel untuk POC Surabaya sebesar 177 MSISDN dan POC Denpasar sebesar 171 MSISDN dari total sampel 819 MSISDN. Untuk total POC terendah terletak di Ponorogo untuk daerah Jawa Timur yaitu sebesar 2 MSISDN diikuti dengan

POC Kupang sebanyak 3 MSISDN. Distribusi Data POC ini juga bisa dijadikan gambaran terkait penyebaran pengguna layanan secara umum pada PT Telekomunikasi XYZ.

4.1.2 Distribusi Data *Handset*

Handset merupakan *device barier* yang digunakan pada perangkat pengguna layanan pada PT Telekomunikasi XYZ. Variabel data *handset* ini terdiri dari 7 jenis jaringan yaitu 2G, 2.5G, 2.75G, 3G, 3.5G, 3.75G dan 4G di mana Untuk jaringan 2G merupakan *second-generation* yang digunakan untuk tujuan *Global System for Moble Communication* (GSM) yang dapat melakukan komunikasi dan layanan pesan teks. Adapun 2.5G merupakan pengembangan dari jaringan 2G untuk tujuan *General Packet Radio System* (GPRS) yang dapat melakukan transfer dan menerima data. Sedangkan untuk 2.75G merupakan pengembangan dari 2.5G dengan tujuan *Enhanced Data Rates for GSM Evoultion* (EDGE) di mana memiliki kecepatan transfer teoritis maksimum 500 kbit/ detik. Adapun jaringan 3G merupakan singkatan dari istilah dalam bahasa Inggris yaitu third-generation technology yaitu sebuah standar yang ditetapkan oleh International Telecommunication Union (ITU) yang diadopsi dari IMT-2000 untuk diaplikasikan pada jaringan telepon seluler yang memiliki kemampuan transmisi berkisar antara 384 Kbps – 2Mbps.

Teknologi 3G terus berkembang yang menghasilkan teknologi 3.5G atau biasa disebut Turbo 3G dengan teknologi High Speed Packet Access (HSPA) dengan kecepatan hingga 14 Mbps untuk download dan 5.76 untuk upload. Teknologi 3.5G kemudian dikembangkan menjadi 3.75G yang ditingkatkan dari HSPA menjadi HSPA+ atau bisa disebut juga *Evolved HSPA* dengan kecepatan transmisi hingga 42 Mbps. Sedangkan 4G adalah singkatan dari istilah dalam bahasa Inggris yaitu *fourth-generation technology* atau yang biasa disebut dengan *Fourth Generation Long Term Evolution* (LTE). Istilah ini umumnya digunakan mengacu kepada standar generasi keempat dari teknologi telepon seluler. Jaringan 4G merupakan pengembangan dari teknologi 3G dengan kecepatan transmisi berkisar antara 100 Mbps – 1Gbps.

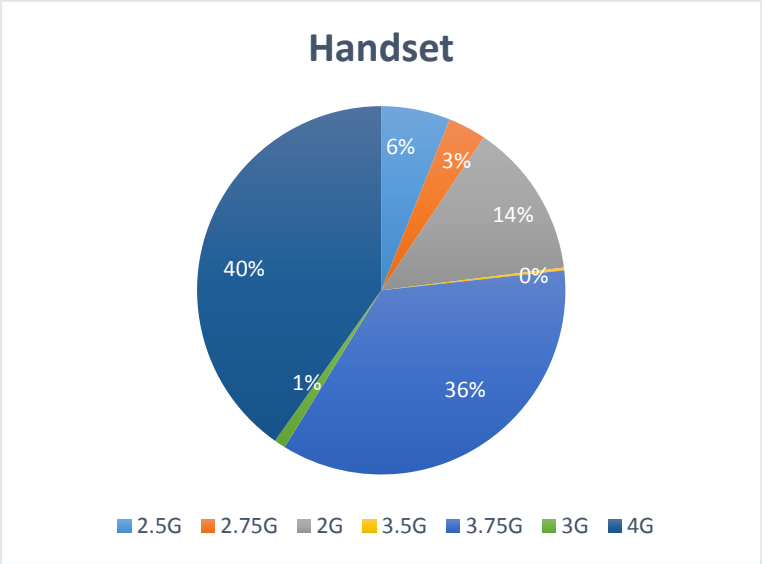


Diagram 4.2 Distribusi data *handset*

Sebagaimana digambarkan pada diagram 4.2 bahwa pengujian data *handset* didominasi oleh penggunaan jaringan 4G yaitu sebesar 40% diikuti oleh 3.75G atau H+ dengan presentase 36%. Sedangkan penggunaan jaringan terendah yaitu 3,5G dengan presentase sebesar 0% di mana pada dasarnya data 3.5G sudah dialih fungsikan pada *barrier* 3.75G dan 4G. Data tersebut juga bisa dijadikan gambaran secara umum terkait barrier yang ada pada device pengguna pengguna di jaringan PT Telekomunikasi XYZ.

4.1.3 Distribusi Data Penggunaan *Voice*

V*oice call* merupakan panggilan telepon yang dilakukan oleh pelanggan baik sesama operator ataupun berbeda operator. Variabel penggunaan *voice call* ini yaitu menilai jumlah telepon masuk dan keluar dalam satuan *minutes* dari pengguna layanan PT Telekomunikasi XYZ. Layanan *voice call* ini sudah termasuk *night call* dan *day call* periode Januari hingga Maret 2018 dengan data distribusi pada Diagram 4.3.

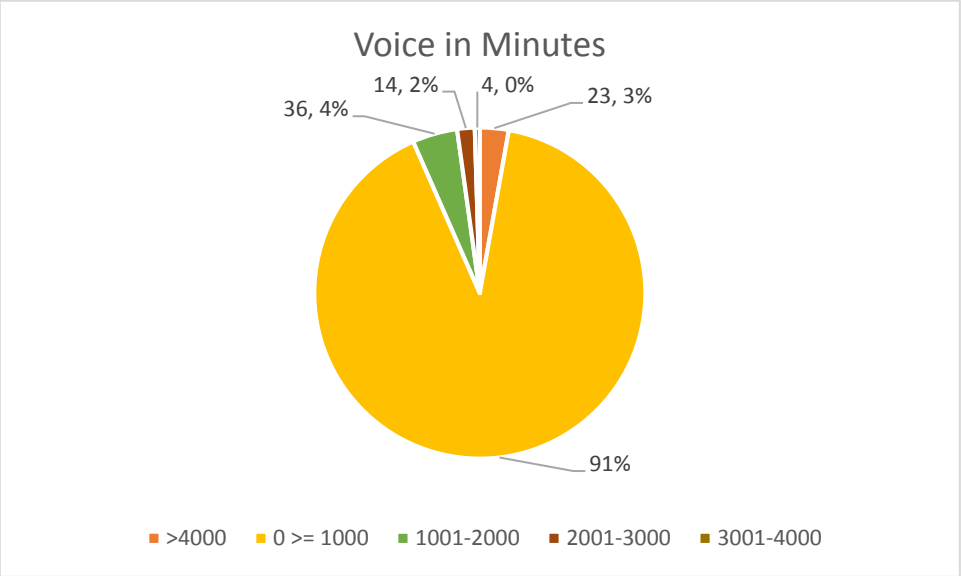


Diagram 4.3 Distribusi data *voice*

Pada diagram 4.2 terlihat bahwa penggunaan *voice call* tertinggi yaitu 91% dengan jumlah waktu panggilan masuk dan keluar sebesar 0 hingga 1000 *minutes* selama periode Januari hingga Maret 2018. Sedangkan penggunaan *voice call* terendah yaitu 4.0% dengan jumlah waktu panggilan masuk dan keluar sebesar 3001 hingga 4000 *minutes* diikuti dengan panggilan masuk dan keluar sebesar 2001 hingga 3000 *minutes* dengan presentase 2% selama periode Januari hingga Maret 2018. Dari data ini, bisa digambarkan secara umum bahwa pengguna *Voice Call* di PT Telekomunikasi XYZ berkisar di angka maksimum 333 menit per bulan (1000 Minutes /3 bulan) per MSISDN.

4.1.4 Distribusi Data Penggunaan SMS

SMS merupakan *Short Message Service* yaitu mengirim pesan antar pengguna layanan *provider* baik sesama operator maupun berbeda operator. Variabel penggunaan SMS ini yaitu menilai jumlah SMS masuk dan keluar dalam satuan *event. Event* dalam dunia telekomunikasi dimaksudkan dengan jumlah, artinya jumlah dari pengguna layanan SMS pada PT Telekomunikasi XYZ periode Januari hingga Maret 2018.

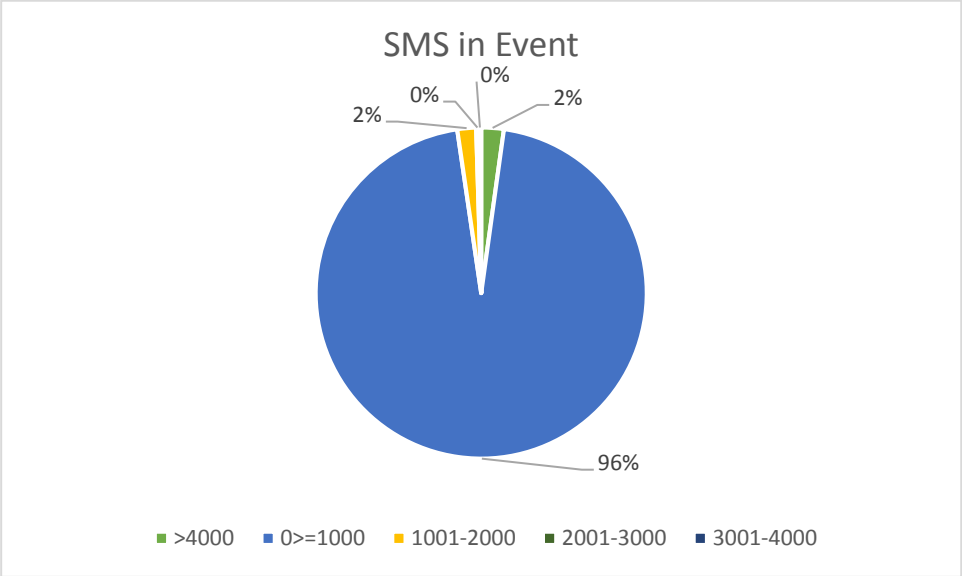


Diagram 4.4 Distribusi data SMS

Sebagaimana digambarkan pada diagram 4.4 bahwa penggunaan SMS tertinggi yaitu 96% dengan jumlah SMS masuk dan SMS keluar sebesar 0 hingga 1.000 *event* selama periode Januari hingga Maret 2018. Sedangkan penggunaan SMS terendah yaitu 0% dengan jumlah SMS masuk dan keluar sebesar 2001 hingga 3000 *event,* 3.001 hingga 4000 *event,* diikuti dengan jumlah SMS sebanyak 1001 hingga 2000 serta >4000 *event* dengan presentase 2% selama periode Januari hingga Maret 2018. Artinya penggunaan layanan SMS rata-rata per bulan pada PT Telekomunikasi XYZ sebesar 333 *event* yaitu maksimum 1000 SMS dibagi dengan 3 bulan.

4.1.5 Distibusi Penggunaan Layanan Data

Penggunaan paket data merupakan layanan yang digunakan untuk akses inter *connection-networking* atau internet dalam satuan *kb* dari pengguna layanan data PT Telekomunikasi XYZ selama periode Januari hingga Maret 2018.

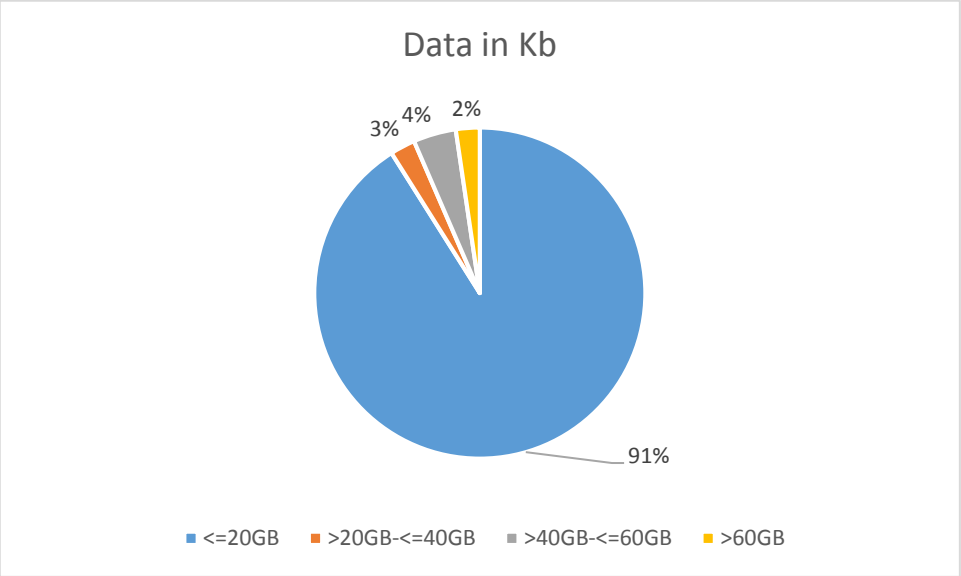


Diagram 4.5 Distribusi penggunaan Data in Kb

Sebagaimana digambarkan pada diagram 4.5 bahwa penggunaan paket data tertinggi yaitu 91% dengan jumlah penggunaan paket data sebesar 0 hingga 20 GB selama periode Januari hingga Maret 2018. Sedangkan penggunaan paket data terendah yaitu 2% dengan jumlah penggunaan paket data sebesar lebih besar dari 60GB selama periode Januari hingga Maret 2018. Dari penggunaan layanan data *in kb* ini, bisa disimpulkan bahwa rata-rata penggunaan data per bulan pada PT Telekomunikasi XYZ berkisar maskimum 6.6GB dengan membagi angka 20GB dengan 3 bulan yang menggambarkan pola konsumsi pelanggan data secara umum.

4.1.6 Distribusi Data Paket Name

Packet Name ini merupakan nama paket yang digunakan oleh pelanggan dalam menggunakan layanan PT Telekomunikas XYZ. Nama paket biasanya *inline* dengan layanan yang diberikan oleh provider. Misal untuk nama paket Super Ngobrol, ini akan memberikan layanan berupa voice.

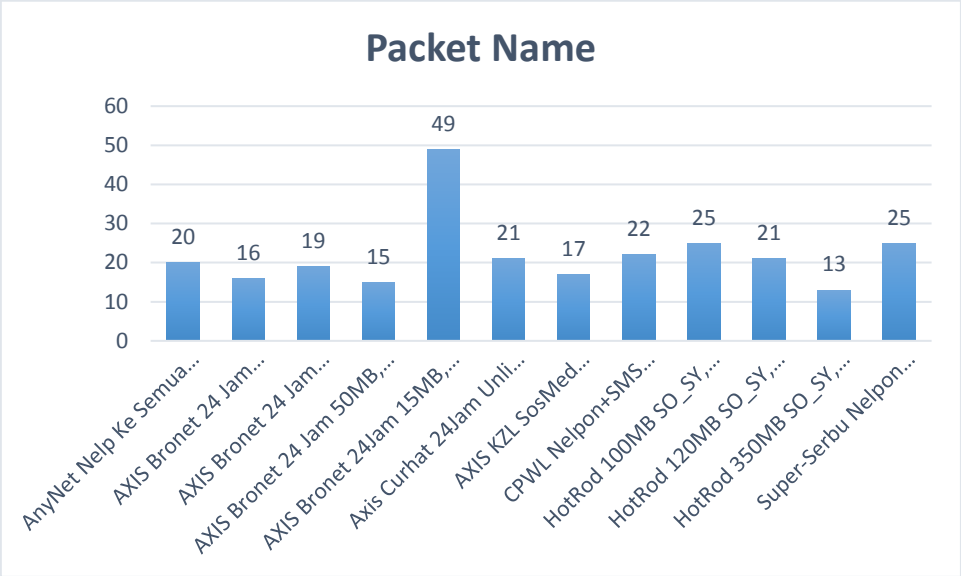
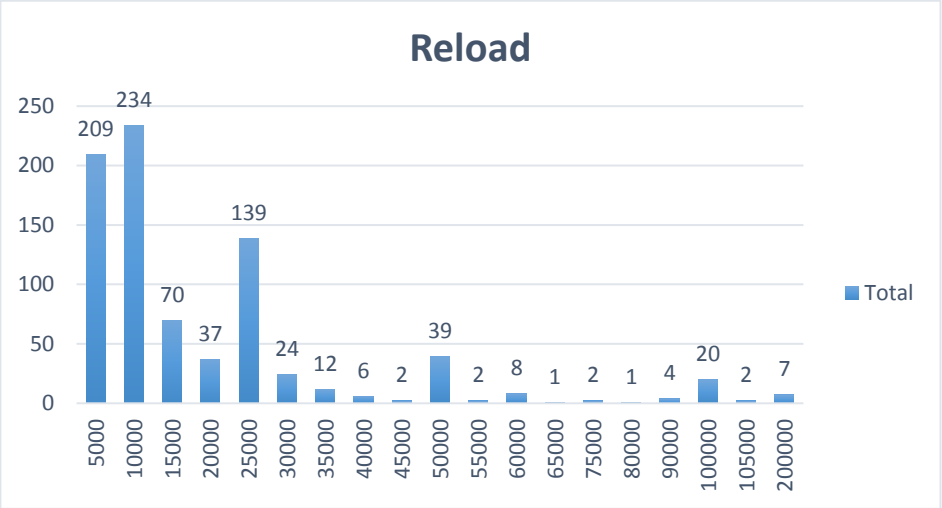


Diagram 4.6 Distribusi Packet Name

Sebagaimana pada diagram 4.6, bahwa distribusi nama paket yang paling banyak digunakan adalah Axis Bronet 24 Jam dengan total MSISDN sebesar 49 pada bulan Januari hingga Maret, 2018. Sedangkan nama paket yang paling sedikit digunakan adalah Paket HotRod 350MB dengan total MSISDN sebesar 13 MSISDN. Dari distribusi ini, dapat disimpulkan bahwa Mayoritas pelanggan lebih tertarik dengan nama paket Axis Bronet 24 Jam jika dibandingkan dengan paket yang lain, hal ini inline dengan harga yang ditawarkan paket tersebut cukup murah dengan kualitas jaringan yang baik.

4.1.7 Distribusi Data Reload

Data *reload* merupakan pengisian pulsa kembali yang dilakukan oleh pengguna layanan guna bisa tetap menggunakan layanan yang disediakan PT Telekomunikasi XYZ berupa *voice call*, SMS maupun paket data.



Grafik 4.7 Distribusi *Reload*

Sebagaimana digambarkan pada grafik 4.7 bahwa *reload* yang dilakukan pelanggan didominasi oleh pulsa 10.000 dan pulsa 5.000,di mana jumlah sampel untuk reload pulsa 10.000 sebesar 234 MSISDN dan reload pulsa 5.000 sebesar 209 MSISDN dari total sampel 819 MSISDN. Untuk total reload terendah terletak pada reload pulsa 65.000 dan 80.000 yaitu sebesar masing-masing 1 MSISDN. Data tersebut juga bisa dijadikan gambaran terkait reload pulsa pelanggan secara umum pada PT Telekomunikasi XYZ.

4.2 Karakteristik Data Uji

Karakteristik Data uji merupakan karakteristik data uji yang dilakukan pengelompokkan untuk memudahkan pada saat melakukan testing menggunakan regresi logistik. Data berupa kategorikal untuk masing-masing variabel independen dan variabel dependen. Pada variabel dependen, data hanya terdiri dari dua kategori data misal iya dan tidak, atau 1 dan 0, dalam hal ini adalah churn dan non churn. Sedangkan pada variabel independen bisa terdiri dari dua kategori atau lebih.

Tabel 4.1 Karakteristik data POC

POC	Churn	Non Churn	Grand Total
Jatim	176	274	450
Bali Nusa	143	226	369
Grand Total	**319**	**500**	**819**

Pada tabel 4.1, dijelaskan bahwa data uji untuk variabel POC terdiri dari dua kategori, yaitu POC yang ada di Jawa Timur dan POC yang ada di Bali Nusa. Masing-masing POC terdiri dari churn dan non churn di mana angka churn di Jatim sebesar 176 dan 274 non churn di mana sedikit lebih banyak jika dibandingkan dengan data churn dan non churn yang ada di Bali Nusa sebesar 143 dan 226.

Tabel 4.2 Karakteristik data Handset

Handset	Churn	Non Churn	Grand Total
Unknown	66	120	186
2G	2	9	11
3G	141	156	297
4G	110	215	325
Grand Total	**319**	**500**	**819**

Pada tabel 4.2 dapat dilihat bahwa untuk jumlah angka non churn lebih besar pada masing-masing handset jika dibandingkan dengan angka churn yaitu sebesar 500 untuk churn dan 319 untuk non churn. Pada masing-masing handset, angka terbesar ada pada handset 4G sebesar 325 dengan 215 angka churn dan 110 angka non churn. Untuk 2G merupakan gabungan data dari 2.5G dan 2.75G, begitu juga dengan 3G merupakan gabungan dari 3.5G dan 3.75G. Hal ini dilakukan untuk memudahkan kategorikal pada saat uji regresi logistik. Adapun *unknown* adalah device yang tidak support di database perusahaan telekomunikasi XYZ sehingga tidak bisa tercapture jenis handsetnya.

Tabel 4.3 Karakteristik data Voice in Minutes

Voice in Minutes	Churn	Non Churn	Grand Total
0	78	69	147
1<=500	181	316	497
>500	60	115	175
Grand Total	**319**	**500**	**819**

Pada tabel 4.3 merupakan karakteristik data dari penggunaan voice dalam jumlah minutes. Dapat digambarkan bahwa jumlah keseluruhan, data non churn sebesar 500 sedangkan untuk data churn sebesar 319. Adapun jumlah penggunaan

voice dengan kategori lebih kecil dari 500 Minutes selama 3 bulan mendominasi sebesar 497 jika dibandingkan data penggunaan di atas 500 minutes yang hanya berjumlah 175. Sedangkan angka 0 (tidak ada penggunaan voice), jumlah churn lebih banyak jika dibandingkan dengan non churn. Hal ini bisa diasumsikan bahwa terjadi churn dikarenakan tidak memanfaatkan layanan voice.

Tabel 4.4 Karakteristik data SMS in Event

SMS in Event	Churn	Non Churn	Grand Total
0	21	25	46
1<=500	280	431	711
>500	18	44	62
Grand Total	**319**	**500**	**819**

Seperti dijelaskan pada tabel 4.3, karakteristik SMS terdiri dari 319 data Churn dan 500 data non churn. Data terbesar ada pada penggunaan SMS di bawah 500 selama 3 bulan, yaitu churn sebesar 280 dan non churn sebesar 431. Sedangkan penggunaan SMS lebih besari dari 500 hanya sebesar 62, hanya sedikit lebih besar jika dibandingkan dengan angka 0 SMS yaitu sebesar 46.

Tabel 4.5 Karakteristik data Data in kb

Layanan Data	Churn	Non Churn	Grand Total
0	44	62	106
1<=10GB	256	327	583
>10GB	19	111	130
Grand Total	**319**	**500**	**819**

Tabel 4.5 merupakan jumlah data penggunaan layanan data dalam jumlah kb, namun untuk memudahkan analisis data, digunakan kategori dengan menjadikan data kb sebagai GB. Dari tabel 4.5 dapat dilihat bahwa data penggunaan layanan data dibawah 10GB sebesar 583, artinya lebih banyak jika dibandingkan dengan penggunaan layanan data di atas 10 GB yang hanya sebesar 130. Dari data 583 tersebut, jumlah data non churn sebesar 327, lebih banyak jika dibandingkan dengan data churn yang sebesar 256. Secara total angka churn sebesar 319 dan non churn sebanyak 500.

Tabel 4.6 Karakteristik data Layanan Paket

Packet	Churn	Non Churn	Grand Total
A	107	193	300
B	149	179	328
Unknown	63	128	191
Grand Total	**319**	**500**	**819**

Pada tabel 4.6 merupakan karakteristik data jenis layanan paket. Pada awalnya paket ini adalah layanan berbagai nama paket yang digunakan oleh customer. Namun dengan jumlah paket yang cukup banyak, maka dibuat suatu kategori khusus dengan Paket dari layanan A dan Paket dari layanan B dengan jumlah total sebesar 819 yang terdiri dari 300 data untuk paket A, 328 untuk paket B dan unknown sebesar 191. Dari ketiga kategori tersebut, data non churn sebesar 500, lebih banyak jika dibandingkan dengan data churn, sebesar 319. Adapun untuk data Unknown di sini adalah data yang tercapture sebagai layanan suatu paket, namun tidak ada di database yang menggambarkan nama paket tersebut, namun tetap tercapture secara usage baik voice, sms ataupun data.

Tabel 4.7 Karakteristik data layanan Reload

Reload	Churn	Non Churn	Grand Total
<=25000	293	396	689
>25000	26	104	130
Grand Total	**319**	**500**	**819**

Karakteristik data terkahir adalah karakteristik data pada layanan reload. Sebagaimana dijelaskan pada tabel 4.7, bahwa data berjumlah 819 dengan data churn sebesar 319 dan non churn sebear 500. Data reload lebih kecil dari 25 ribu, merupakan akumulasi reload selama sebulan dengan total pembelian reload lebih kecil dari 25000 dengan jumlah data churn sebesar 293 dan non churn sebesar 394. Adapun reload di atas 25000 hanya sebesar 130 dengan data churn sebanyak 26 dan non churn sebesar 104. Artinya customer dengan reload di atas 25000 75% tidak terjadi churn.

Tabel 4.8 Variabel dan Kategorikal

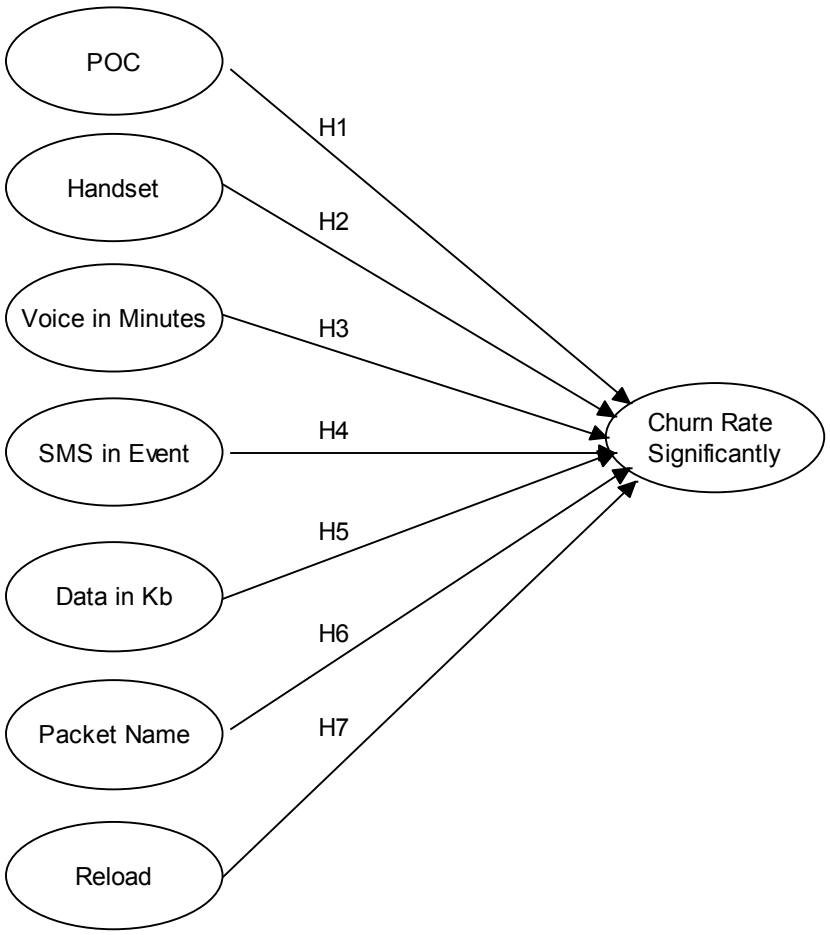
Variabel	**Kategori**
POC $(X1)_{(0)}$	Jatim
POC $(X1)_{(1)}$	Bali Outer
Handset$(X2)_{(0)}$	Unknown
Handset $(X2)_{(1)}$	2G
Handset $(X2)_{(2)}$	3G
Handset $(X2)_{(3)}$	4G
Voice in Minutes $(X3)_{(0)}$	0
Voice in Minutes $(X3)_{(1)}$	1<=500
Voice in Minutes $(X3)_{(2)}$	>500
SMS in Event $(X4)_{(0)}$	0
SMS in Event $(X4)_{(1)}$	1<=500
SMS in Event $(X4)_{(2)}$	>500
Data in Kb $(X5)_{(0)}$	0
Data in Kb $(X5)_{(1)}$	1<=10GB
Data in Kb $(X5)_{(2)}$	>10GB
Packet $(X6)_{(0)}$	A
Packet $(X6)_{(1)}$	B
Packet $(X6)_{(2)}$	Unknown
Reload $(X6)_{(0)}$	<=25000
Reload $(X6)_{(1)}$	>25000

Tabel 4.8 merupakan jumlah variabel beserta kategori masing-masing variabel sesuai dengan karakteristik data uji . Variabel didefinisikan sesuai dengan urutan, misal handset sebagai urutan ke 1 dengan variabel pengganti sebagai X1. Data ini nantinya bisa dijadikan acuan dalam mendefiniskan masing-masing variabel terpilih dan rekomendasi hasil penelitian akhir.

4.3 Hipotesa awal

Hipotesa awal 0 (H0) merupakan hiptesa bahwa semua variabel independen tidak berpengaruh secara signifikan terhadap *churn rate*, artinya hipotesis pembandingnya (Ha) adalah bahwa semua variabel berpengaruh signifikan terhadap terjadinya *churn rate*. Dalam hal ini terdapat tujuh variabel independen,

yaitu *POC, Handset, Voice in Minutes, SMS in Event, data in Kb*, packet name dan reload yang secara bersama menyatakan bahwa tidak ada pengaruh signifikan terhadap churn sebagai variabel dependen.



Gambar 4.1. Hipotesis Penelitian

Pada gambar 4.1 bahwa terdapat tujuh hipotesis awal dengan POC sebagai hipotesis 1 selanjutnya disebut H1, *Handset* sebagai H2, Voice in Minutes sebagai H3, SMS in Event sebagai H4, Data in Kb sebagai H5, Packet Name sebagai H6 dan Reload sebagai H7 dengan masing-masing H0 sebagai hipotesis pembanding. Hipotesis awal (H0) menyatakan bahwa semua variabel independen tidak berpengaruh secara signifikan terhadap *churn rate* (dependen). untuk membuktikan hipotesis tersebut, dilakukan suatu uji Statisik dengan uji independensi.

4.4 Analisis Uji Independensi

Uji Independensi digunakan untuk mengetahui ada atau tidaknya hubungan antara variabel independen dengan variabel dependen. Dalam hal ini adanya hubungan antara customer churn dengan faktor-faktor yang mempengaruhinya. Hipotesis yang digunakan adalah sebagai berikut:

H_0 : Tidak terdapat hubungan antara customer churn dengan variabel yang diduga mempengaruhinya.

H_1 : Terdapat hubungan yang signifikan antara customer churn dengan variabel yang diduga mempengaruhinya.

Taraf Siginifikan : $\alpha = 0{,}05$

Daerah kritis : Tolak H_0 jika $X^2 > X^2_{(df,\alpha)}$ atau $P_{value} < \alpha$

Dari pengujian serentak, kemudian didapati variabel independen yang kemudian secara tegas menyatakan signifikan terhadap variabel dependen yang digambarkan pada tabel 4.2

Tabel 4.9 Hasil uji Independensi

Variabel	**X^2**	**Df**	**$X^2_{7;0.05}$**	**P value**	**Keputusan**
POC (X_1)	.004	1	3.841	.950	H_0 diterima
Handset (X_2)	3.900	1	3.841	.048	H_0 ditolak
Voice in Miniutes (X_3)	13.595	1	3.841	.000	H_0 ditolak
SMS Event (X_4)	1.480	1	3.841	.224	H_0 diterima
Usage Data (X_5)	20.481	1	3.841	.000	H_0 ditolak
Packet (X_6)	1.002	1	3.841	.317	H_0 diterima
Reload (X_7)	14.933	1	3.841	.000	H_0 ditolak

Dari hasil uji independensi pada tabel 4.9 bahwa terdapat empat variabel independen yang mempengaruhi variabel dependen dengan nilai X^2 lebih besar dari X_{tabel} atau nilai P_{value} lebih kecil dari 0.05 yaitu variabel Hanset (X2), variabel variabel Voice in Minutes (X3), variabel Usage Data (X5), dan variabel Reload (X7). Artinya keempat variabel tersebut ada hubungan signifikan terhadap churn rate yang diduga mempengaruhinya, sehingga perlu dilakukan uji parsial. Sedangkan ketiga variabel lainnya tidak mempengaruhi secara signifikan.

4.5 Analisis Regresi Logistik

Regresi logistik biner merupakan suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel respon (y) yang bersifat biner atau dikotomus dengan variabel prediktor (x) yang bersifat politokomus.

4.5.1 Uji Signifikansi secara simultan

Uji signifikansi secara simultan dilakukan untuk mengetahui apakah variabel-variabel bebas dapat mempengaruhi secara signifikan terhadap variabel terikat dengan hipotesa sebagai berikut:

H_0 : $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$ bahwa semua variabel bebas tidak berpengaruh secara signifikan terhadap variabel terikat.

H_1 : $\beta_i \neq 0$ di mana i=1,2,3,4,5,6,7,8 minimal terdapat satu variabel bebas berpengaruh signifikan terhadap variabel terikat.

Taraf Siginifikan : $\alpha = 0{,}05$

Daerah kritis : Tolak H_0 jika $X^2 > X^2_{(df,\alpha)}$ atau $P_{value} < \alpha$

Tabel 4.10 Hasil uji signifikansi variabel secara simultan

	X^2	Df	$X^2_{7;0.05}$	P value
Model	90.645	13	22.362	0.000

Dari tabel 4.10 dapat dilihat bahwa nilai X^2 sebesar 90.645 lebih besar dari X_{tabel} yaitu sebesar 22.362 atau nilai P_{value} sebesar 0.000 lebih kecil dari nilai α 0.05, sehingga dapat dibuat keputusan bahwa H^0 ditolak, artinya terdapat minimal satu variabel independen yang berpengaruh signifikan terhadap variabel dependen.

4.5.2 Uji Signifikansi secara parsial

Uji signifikansi parameter secara parsial dilakukan untuk mengetahui apakah variabel yang signifikan dari hasil uji secara simultan yang telah dilakukan memberikan pengaruh yang signifikan terhadap model yang telah didapatkan. Berikut hasil uji pengujian signifikansi parameter secara parsial.

1. H_0: $\beta_2 = 0$ bahwa reload yang dilakukan oleh customer tidak berpengaruh secara signifikan terhadap churn dan non churn.

   H_1: $\beta_2 \neq 0$ bahwa reload yang dilakukan oleh customer berpengaruh secara signifikan terhadap churn dan non churn.
2. H_0: $\beta_4 = 0$ bahwa handset yang ada pada device customer tidak mempengaruhi secara signifikan terhadap churn dan non churn.

   H_1: $\beta_4 \neq 0$ bahwa handset yang ada pada device customer berpengaruh secara signifikan terhadap churn dan non churn.
3. H_0: $\beta_5 = 0$ bahwa penggunaan layanan data tidak mempengaruhi secara signifikan terhadap churn dan non churn.

   H_1: $\beta_5 \neq 0$ bahwa penggunaan layanan data berpengaruh secara signifikan terhadap churn dan non churn.
4. H_0: $\beta_7 = 0$ bahwa penggunaan layanan voice tidak mempengaruhi secara signifikan terhadap churn dan non churn.
5. H_1: $\beta_7 \neq 0$ bahwa penggunaan layanan voice berpengaruh secara signifikan terhadap churn dan non churn.

Taraf Siginifikan : $\alpha = 0{,}05$

Daerah kritis : Tolak H_0 jika $|W^2| > Z_{\alpha/2}$ atau $P_{value} < \alpha$

Tabel 4.11 Hasil uji signifikansi variabel secara parsial

	B	**Wald**	**df**	**Sig.**	**Exp(B)**
Handset $(X_2)_{(0)}$	-.194	11.665	3	.009	0.823
Voice $(X_3)_{(0)}$	.469	18.024	2	.000	1.598
Voice$(X_3)_{(1)}$	-.999	16.496	1	.000	0.369
Data$(X_5)_{(1)}$	-1.545	18.228	1	.000	0.213
Data$_{(}(X_5)_{2)}$	-1.358	23.341	1	.000	0.257
Reload$(X_7)_{(0)}$	-.786	10.251	1	.001	0.455
Constant	2.710	57.194	1	.000	15.036

Pada tabel 4.10, menghasilkan sebuah model logit sebagai berikut:

$$g(x) = 2.710 - 0.194X_{2(0)} + 0.469X_{3(0)} - 0.999X_{3(1)} - 1.545X_{5(1)} - 1.358X_{5(2)} - 0.786X_{7(1)}$$

Model logit tersebut dapat digunakan padarumus logit di bawah untuk mendapatkan besar peluang.

$$\pi_1(x) = \frac{\exp(g(x))}{1 + \exp(g(x))}$$

$$= \frac{e^{2.710 - 0.194X_{2(0)} + 0.469X_{3(0)} - 0.999X_{3(1)} - 1.545X_{5(1)} - 1.358X_{5(2)} - 0.786X_{7(1)}}}{1 + e^{2.710 - 0.194X_{2(0)} + 0.469X_{3(0)} - 0.999X_{3(1)} - 1.545X_{5(1)} - 1.358X_{5(2)} - 0.786X_{7(1)}}}$$
$$= 0.937$$

Tabel 4.12 Keterangan variabel terpilih

Variabel	**Keterangan Data**
Handset $(X_2)_{(0)}$	Unknown
Voice $(X_3)_{(0)}$	0 minutes
Voice$(X_3)_{(1)}$	1<=500 minutes
Data$(X_5)_{(1)}$	1<=10GB
Data$_{(}(X_5)_{2)}$	>10GB
Reload$(X_7)_{(0)}$	<=25000
Constant	2.710

Dari tabel 4.12, dijelaskan masing-masing variabel terpilih beserta keterangan kategori data. Pada Handset, variabel terpilih adalah Handset $(X_2)_{(0)}$ dengan data Unknown. Sedangka pada voice, terpilih dua kategori, yaitu 0 (minutes) dan 1<=500 (minutes). Pada variabel data, juga terpilih dua kategori, yaitu 1<=10GB dan >10GB. Dan untuk variabel reload, kategori terpilih adalah >25000.

4.6 Analisis Klasifikasi dengan SVM

Analisis klasifikasi SVM ini dilakukan untuk membandingkan antara nilai akurasi total variabel setelah pemilihan variabel terpilih dengan akurasi variabel terpilih. Analisis ini menggunakan dua fungsi kernel, yaitu *Polynomial*. Pada fungsi polynomial menggunakan parameter sebanyak 3 jenis, yaitu p=1, p=2 dan p=3.

Pada pengujian SVM, dilakukan dengan menggunakan tools Weka 3.8.2 dengan memilih menu Explorer, kemudian pada saat muncul popup, klik tombol open untuk memilih file. Pilih jenis file type, dikarenakan penulis menggunakan csv, maka extensi yang dipilih adalah csv. Pada menu filter, penulis menggunakan filter->unsupervised->attribute->Numeric to Nominal untuk antisipasi jika terdapat attribute tipe nominal kemudian tekan apply. Pada tab classify, pilih classifier->function->SMO untuk memilih jenis klasifikasi dalam hal ini SVM. Pada opsi Test Option, penulis memilih percentage split dengan membagi data menjadi data training dan data testing. Kemudian pilih salah satu attribute yang akan dijadikan kelas, terakhir klik start untuk memulai proses klasifikasi.

4.6.1 Analisis perbandingan klasifikasi SVM

Pada analisis klasifikasi SVM ini digunakan nilai c=1. C adalah sebuah parameter yang digunakan SVM untuk menghindari kesalahan klasifikasi. Semakin besar nilai C, semakin kecil optimasi hyperplane yang digunakan oleh SVM dan sebaliknya, semakin kecil nilai C, maka optimasi hyperplane yang digunakan smakin besar. Fungsi yang digunakan dalam hal ini adalah fungsi kernel *Polynomial*. Masing-masing mempunyai 3 parameter, yaitu p=1,p=2 dan p=3.

Tabel 4.13 Perbandingan akurasi klasifikasi variabel terpilih dan total variabel

Proporsi	**Variabel Terpilih Polynomial**			**Semua Variabel Polynomial**		
	p=1	**p=2**	**p=3**	**p=1**	**p=2**	**p=3**
80:20	73.3	73.3	73.3	66.6	66.6	66.6
70:30	78.2	78.2	78.2	73.9	73.9	73.9
60:40	77.4	77.41	77.41	54.8	54.8	54.8

Berdasarkan tabel 4.13 dijelaskan bahwa nilai akurasi tertinggi terdapat pada proporsi data 70:30 dengan 70 sebagai data training dan 30 sebagai data testing, dengan parameter p=1,p=2 dan p=3 dengan variabel terpilih, sebesar 78.2%, sedangkan pada total variabel, akurasi tertinggi ada pada proporsi

70:30 dengan akurasi 73.9%. Sehingga dapat dibuat kesimpulan bahwa vairabel terpilih secara akurasi lebih bagus jika dibandingkan dengan variabel total dengan range sebesar 4.3%.

Tabel 4.14 Perbandingan akurasi klasifikasi variabel terpilih dan total data

Proporsi	**Variabel Terpilih Polynomial**			**Total Data dan variabel Polynomial**		
	p=1	**p=2**	**p=3**	**p=1**	**p=2**	**p=3**
80:20	73.3	73.3	73.3	58.5	68.2	68.9
70:30	78.2	78.2	78.2	71.5	68.6	71.5
60:40	77.4	77.41	77.41	60.6	70.1	70.4

Pada tabel 4.14 menjelaskan bahwa variabel terpilih pada saat dilakukan komparasi dengan uji total data (tanpa pemilihan kategorikal terpilih) sekaligus total variabel, masih lebih baik dengan akurasi tertinggi pada 78.2%, sedangkan pada uji total data dan variabel, akurasi terbaik adalah 71.5% pada parameter p=1. Hal ini menunjukkan bahwa variabel terpilih, masih lebih bagus jika dilakukan uji total data sekaligus total variabel dengan gap akurasi sebesar 6.7%.

Tabel 4.15 Perbandingan akurasi variabel terpilih dan total data variabel terpilih

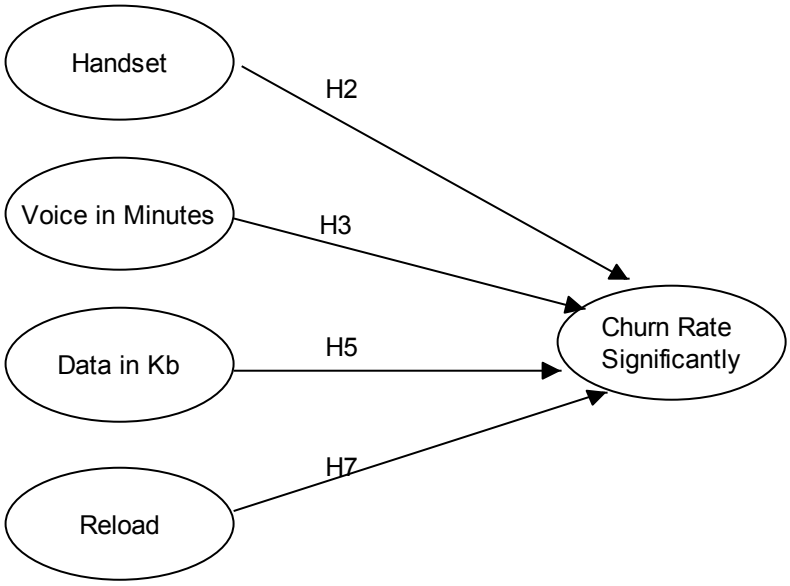
Proporsi	**Variabel Terpilih Polynomial**			**Total Data dan variabel terpilih Polynomial**		
	p=1	**p=2**	**p=3**	**p=1**	**p=2**	**p=3**
80:20	73.3	73.3	73.3	58.5	61.5	60.3
70:30	78.2	78.2	78.2	59.7	63.8	65.8
60:40	77.4	77.41	77.41	60.6	64.9	66.4

Pada tabel 4.15 merupakan pengujian variabel terpilih dengan total data dan variabel terpilih, di mana menggambarkan bahwa untuk nilai akurasi tertinggi pada variabel terpilih sebesar 78.2% berada pada proporsi 70:30, sedangkan pada total data

dan variabel terpilih, akurasi tertinggi 66.4% pada proporsi 60:40 dengan parameter p=3. Hal ini masih jauh lebih baik akurasi variabel terpilih dengan presentase gap sebesar 11.8%.

4.7 Model Akhir Penelitian

Setelah dilakukan pengujian, maka didapatkan suatu model akhir dari penelitian ini di mana terdapat 4 hasil akhir sebagaimana gambar 4.2 yang menyatakan bahwa variabel *Handset*, *Voice in Minutes*, *Data in Kb* serta variabel *Reload* secara bersama-sama berpengaruh secara signifikan terhadap terjadinya *churn rate*. Model yang diperoleh telah sesuai dengan uji kesesuaian model pada regresi logistik.



Gambar 4.2 Model akhir Penelitian

4.8 Analisis Odds Ratio dan Rekomendasi Penelitian

Setelah didapatkan model penelitian sebagaimana pada gambar 4.2, maka dilakukan analisis odds rasio untuk mengetahui seberapa besar peluang yang dihasilkan terhadap churn dan non churn serta diberikan rekomendasi-rekomendasi pada masing-masing variabel terpilih dengan tujuan sebagai masukan pada perusahaan telekomunikasi demi perkembangan dan perbaikan layanan ke depannya.

4.8.1 Variabel Handset

Pada tabel 4.10, nilai odds rasio pada variabel handset sebesar 0.823, artinya peluang handset dengan kategori unknown, cenderung memiliki kemungkinan churn sebesar 0.823 jika dibandingkan dengan handset 2G, 3G dan 4G, artinya perusahaan telekomunikasi xyz perlu melakukan perbaikan pada sisi hardware jaringan dalam mendeteksi handset dari device customer, dikarenakan potensi 1 MSISDN yang unknown, hampir menghasilkan 1 angka churn dengan angka exponensial (B) sebesar 0.82.

4.8.2 Analisis Voice in Minutes

Variabel Voice in Minutes, sebagaimana pada tabel 4.10 terdapat dua kategori yang mempengaruhi churn rate secara signifikan, yaitu voice dengan nilai 0 dengan nilai Odds Ratio 1.598 yang berarti bahwa satu MSISDN yang angka voicenya 0, memungkinkan terjadinya churn sebesar 1.59 sehingga pengaruhnya terhadap churn sangat besar. Adapun voice dengan nilai 1<=500 juga berpengaruh terhadap churn dengan nilai Odds Ratio sebesar 0.3, artinya 1 MSISDN dengan penggunaan voice 1 sampai 500 minutes probabilitas MSISDN menjadi churn sebesar 0.369, yang masih kecil jika dibandingkan dengan voice dengan nilai 0. Dari kedua angka ini bisa disimpulkan bahwa customer masih banyak menggunakan layanan voice yang mengakibatkan angka probabilitas signifikansinya cukup tinggi, sehingga perusahaan telekomunikasi xyz perlu melakukan promosi layanan voice lebih gencar lagi dengan tarif yang relatif murah dengan kualitasi yang sama, hal ini memungkinkan user untuk tetap setia dengan layanan voice yang diberikan.

4.8.3 Analisis Data in Kb

Adapun untuk variabel Data in Kb, pada tabel 4.10 dijelaskan bahwa terdapat dua kategori yang mempengaruhi churn secara signifikan, yaitu kategori data dengan usage 1<50GB dan usage >10 GB. Namun nilai yang dihasilkan sangat kecil. Pada kategori 1<50 GB, nilai Odds Ratio yang dihasilkan sebesar 0.213, artinya setiap 1 MSISDN dengan pemakaian usage data dari angka 1 sampai 50GB selama tiga bulan, probabilitas churn sebesar 0.213 yang bisa dibilang relative kecil. Sedangkan pada pemakaian data > 50 GB selama 3 bulan, probabilitas churnnya sebesar 0.257, sedikit lebih besar

jika dibandingkan dengan nilai 1<50 GB. Artinya kedua kategori sama-sama berpengaruh churn namun dengan angka yang relative kecil.

Rekomendasi yang diberikan pada perusahaan telekomunikasi XYZ adalah terus meningkatkan infrastruktur yang baik pada layanan data dengan terus membangun jaringan berkualitas di masing-masing daerah POC yang menjadi area fokus sehingga user tetap merasa nyaman dengan layanan yang diberikan. Selain itu, juga perlu dilakukannya riset dengan memanfaatkan *Big Data* analisis sehingga mendapatkan pelanggan yang potensial dengan memanfaatkan *historical* penggunaan data yang kemudian dilakukan analisis yang tepat dalam memberikan layanan sesuai kebutuhan pelanggan.

4.8.4 Analisis Reload

Pada tabel 4.10 bisa dilihat bahwa untuk Reload pulsa berpengaruh terhadap churn rate dengan nilai Odds Ratio sebesar 0.455. yang artinya 1 MSISDN yang melakukan transaksi reload lebih kecil dari 25000, probabilitas angka churn sebesar 0.455. secara angka belum besar, akan tetapi bisa dijadikan tolak ukur dalam melakukan maintain terhadap churn.Yang perlu dilakukan oleh perusahaan telekouminasi XYZ adalah bagaimana customer sesering melakukan reload, baik dengan angkan kecil maupun angka besar dengan cara menawarkan paralel bonus. Misalnya jika customer melakukan reload pulsa dengan angka minimal 10 ribu, maka paralel akan mendapatkan bonus layanan data sebesar 500 MB atapun 1GB, sehingga menarik interest customer dalam melakukan reload.

BAB V
PENUTUP

Pada bab ini akan dijelaskan tentang kesimpulan dari percobaan yang telah dilakukan mengenai analisis faktor-faktor yang mempengaruhi *churn rate* secara signifikan pada perusahaan telekomunikasi. Terdapat juga saran-saran yang mendukung upaya penyempurnaan penelitian ini untuk penelitian kedepannya.

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, maka dapat diambil suatu kesimpulan sebagai berikut;

1. Faktor yang mempengaruhi churn rate secara signifikan adalah *Handset*, penggunaan layanan *Voice in Minutes,* penggunaan layanan *Data in Kb* dan *Reload* dengan nilai signifikansi lebih kecil dari 0.05 (5%).
2. Jika dilihat pada nilai Odds Ratio pada masing-masing variabel terpilih, nilai voice dengan angka 0, merupakan faktor yang paling berpengaruh dengan 1 MSISDN probabilitas churn sebesar 1.59.
3. Nilai akurasi pada klasifikasi variabel terpilih meningkat sebesar 4.3% jika dibandingkan nilai akurasi total variabel yang mempunyai akurasi sebesar 73.9% untuk total variabel dan akurasi 78.2% untuk akurasi variabel terpilih.
4. Nilai klasifikasi SVM lebih baik dengan mengunakan prooporsi data 70:30, 70 untuk data training dan 30 data testing baik untuk variabel terpilih, ataupun variabel total.
5. Rekomendasi yang diberikan untuk variabel terpilih Hanset adalah perbaikan dari sisi hardware jaringan yang menyebabkan handset unknown. Sedangkan untuk *Voice in Minutes* adalah dengan tetap gencar melakukan promosi layanan voice dengan tarif yang relative murah. Adapun rekomendasi untuk variabel terpilih *Data in Kb* adalah terus meningkatkan infrastruktur yang baik pada layanan data dengan terus membangun jaringan berkualitas di masing-masing daerah POC yang menjadi area fokus. Sedangkan rekomendasi untuk variabel terpilih *reload* adalah dengan cara menawarkan paralel bonus, yaitu mendapatkan bonus layanan data dengan nominal reload tertentu.

5.2 Saran

Saran yang ingin disampaikan penulisan untuk kemanfaatan dan pengembangan lebih lanjut terkait penelitian ini sebagai berikut:

1. Bahwa perusahaan telekomunikasi XYZ, perlu melakukan maintain terhadap tiga layanan utama, yaitu layanan *Voice,* Data dan *Reload*, di mana ketiganya berpengaruh secara signifikan terhadap terjadinya *churn rate* terutama pada layanan data yang menjadi kebutuhan *primer* pelanggan.

2. Sebagai komparasi dengan Klasifikasi SVM, dapat dilakukan dengan metode *multivariant* regresi Logistik untuk mendapatkan hasil implementasi klasifikasi yang lebih maksimal.

3. Perlu dilakukan penelitian lebih lanjut tentang *customer oriented,* di mana MSISDN yang dipakai user bisa jadi hanya sebagai *second number* yang berarti bahwa MSISDN tersebut bukan *primery used* yang suatu waktu dapat dibuang.

DAFTAR PUSTAKA

Arikunto, S. (2010). *Prosedur Penelitian Suatu Pendekatan Praktik*. Jakarta: PT Rineka Cipta.

Berry, Michael J.A. & Gordon S. Linoff. (2011). *Data Mining Techniques for Marketing, Sales, Customer Relationship Management,* Second Edition. Wiley Publishing, Inc.

Conolly, Thomas, & Begg, Carolyn. (2005). *Database Systems a Practical Approach to Design, Implementation, and Management. Fourth Edition,* USA.

Fayyad, Usama dkk (1996). *From Data Mining to Knowledge Discovery in Databases*, AI Magazine Volume 17 Number 3.

Fraticasari S (2018). *Optimasi Pemodelan Regresi Linier Berganda Pada Prediksi Jumlah Kecelakaan Sepeda Motor Dengan Algoritme Genetika,* Vol. 2, No. 5 e-ISSN: 2548-964X.

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. New York: Springer-Verlag.

Gunn, S. R. (1998). *Support Vector Machines for Classification and Regression*. Southampton: University of South ampton.

Gursoy, U. (2010). *Customer Churn Analysis in Telecommunication sector, Istanbul University Journal of the School of Business Administration,* Vol 39, No 1, 35 – 49, ISSN: 1303-1732.

Hadden, J. T. (2005). *Computer Assisted Customer Churn Management: State of The Art and Future Trends,* Science Direct, Computer & Operations Research, Volume 34, 2902 – 2917.

Hastie, T., Tibshirani, R., dan Friedman, J. (2001). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. California: Springer.

Herawati, Meyrina. (2016). *Prediksi Customer Churn Menggunakan Algoritma Fuzzy Iterative Dichotomiser 3*. J. Math. and Its Appl, Vol. 13, No. 1.

Hoffer Jeffrey, Ramesh V & Topi Heiki (2012). *Modern Database Management*, Pearson Education Publishing.

Karatzoglou, Alexander (2006) *Support Vector Machines in R.* Volume 15, Issue 9.

Larose, Daniel T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* Jhon Willey & Sons, Inc.

Li, G., You, J., & Liu, X. (2015). *Support Vector Machine (SVM) based prestack AVO inversion and its applications*. Journal of Applied Geophysics, 120 60– 68.

Liao, T. d. (2007). *Recent Advances in Data Mining of Enterprise Data: Algorithms and Applications,* World Scientific, Singapore, pp. 1- 109.

Nugroho, Witarto & Handoko (2003) *Support Vector Machine Teori dan Aplikasinya dalam Bioinformatika*, ilmukomputer.com.

Permatasari, A. I., (2015). *Pemodelan Regresi Linear Dalam Konsumsi KWH Listrik Di Kota Batu Menggunakan Algoritma Genetika*, Malang: Universitas Brawijaya.

Prasetyo, E. (2012). *Data Mining: Konsep dan Aplikasi Menggunakan Matlab*. Indonesia: Andi Yogyakarta.

Richeldi Marco, Perucci Alessandro (2002). *Churn Analysis case study*. Deliverable D17.2, IST Project MiningMart, IST-11993.

Rodan, Faris, Alsakran & Al-Kadi, *A Support Vector Machine Approach for Churn Prediction in Telecom Industry*, King Abdulla II School for Information Technology the University of Jordan Amman, 11942, Jordan, 2014:6.

Santosa, Budi (2007). *Data minig: Teknik Pemanfaatan data untuk keperluan bisnis*. Yogyakarta: Graha Ilmu.

Sekaran, Uma. (2011). *Research Methods for business*. Jakarta: Salemba Empat.

Shaaban, E (2012). *A Proposed Churn Prediction Model*, MUST University, Vol. 2, Issue 4, ISSN: 2248-9622.

Sugiyono. (2010). *Metode Penelitian Pendidikan Pendekatan Kuantitatif, kualitatif, dan R&D.* Bandung: Alfabeta.

Surakhmad W. (2004). *Pengantar Penelitian Ilmiah, Dasar, Metode, dan Teknik*. Bandung: Tarsito

Turban, E., dkk. (2007). *Decision Support Systems and Intelligent Systems*, Yogyakarta: Andi Offset.

Vapnik, V dan Cortes, C. (1995). *Support Vector Networks. Machine Learning*, 20, 273-297.

Yin, Y., Han, D., & Cai, Z. (2011). *Explore Data Classification Algorithm Based on SVM and PSO for Education Decision*. Journal of Convergence Information Technology, 6 (10), 122–128.

Yudhistira, Ali (2017). *Analisa Customer Churn pada perusahaan Internet Service Provider XYZ menggunakan Backpropagation Neural Network*, eISSN: 2319-1163.

DAFTAR LAMPIRAN

Sampel Data Uji Coba

PACKET_DESC	TOTAL_RELOAD_AMOUNT	HOME_POC_NAME	Device	usage_data_kb	sms	voice	Status
Axis Curhat 24Jam Unli Nelpon+SMS UnliMnt+6000SMS, 30hr, Rp14900	20000	Denpasar	2G	0	2378	2782	1
AXIS Bronet 24Jam 15MB, 1hr, Rp888	25000	Surabaya	3.75G	1705090	303	935	1
INTERNET GAUL WEEKLY 200MB	30000	Madiun	4G	4181745	57	15	1
AXIS Bronet 24 Jam 50MB, 1hr, Rp2500	5000	Singaraja	4G	6098289	15	6	1
AXIS SMS+Internet 15MB+1000SMS, 1hr, Rp1500	5000	Denpasar	3.75G	828043	7682	1	1
AXIS KZL Combo Unlimited, 1hr, Rp1500	5000	Jember	3.75G	9607777	1	0	1
HotRod 120MB SO_SY, 7hr, Rp6000	5000	Surabaya	4G	1139235	2	178	1
AXIS SMS+Internet 15MB+1000SMS, 1hr, Rp1500	10000	Denpasar	3.75G	828043	7682	1	1
HotRod 16GB, 30hr, Rp220rb	200000	Denpasar	4G	62226973	349	1068	1
AXIS KZL Combo Unlimited, 30hr, Rp19900	25000	Bima	3.75G	2651484	1	0	1
Nelpon Kapan Aja 100Mnt, 5hr, Rp7000	10000	Mataram	2G	0	18	160	1
Axis Curhat 24Jam Unli Nelpon+SMS UnliMnt+200SMS, 1hr, Rp2500	5000	Denpasar	3.75G	1869506	59	730	1

BIOGRAFI PENULIS

Samsul Arifin. Lahir di Sampang pada 30 November 1992. Merupakan anak kedua dari 4 bersaudara. Penulis menempuh pendidikan formal dari tahun 1998-2004 di SD Negeri Tambak 2 Omben, 2005-2008 di SMP AL-MIFTAH TERPADU, Pamekasan, dan 2008-2011 di SMA AL MIFTAH Pamekasan. Tahun 2011 penulis melanjutkan jenjang pendidikan D3 di jurusan Teknik Informatika, Politeknik Elektronika Negeri Surabaya dan lulus pada 2014. Kemudian 2014-2016 melanjutkan pendidikan D4 di Kampus yang sama. Setelah itu penulis bekerja di PT XL Axiata, Tbk sebagai Software Engineer dan Channel Operation. Kemudian pada tahun 2016 penulis melanjutkan studi S2 di Program Manajemen Teknologi Informasi yang berada dalam Fakultas Bisnis dan Manajemen Teknologi, Institut Teknologi Sepuluh Nopember. Email : Samsul.arifin.me@gmail.com