ANALISIS CHURN PREDICTION MENGGUNAKAN METODE LOGISTIC REGRESSION DAN SMOTE (Synthetic Minority Over-sampling Technique) PADA PERUSAHAAN TELEKOMUNIKASI

CHURN PREDICTION ANALYSIS USING LOGISTIC REGRESSION AND SMOTE (Synthetic Minority Over-sampling Technique) METHOD IN TELECOMMUNICATIONS COMPANY

[1]Muhammad Faruq Mujaddid, [2]Adiwijaya, [3]Said Al-Faraby

[1,2,3] Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom

[1]faruqmujaddid@gmail.com, [2]adiwijaya@telkomuniversity.ac.id, [3]saidalfaraby@telkomuniversity.ac.id

Abstract

The competition between mobile telecominication companies at this time is to certain customers. Customer becomes one of the main factors in the success achieved in mobile telecomunication company. Customers can choose according to their wants and need, this is the main factor triggering the churn. Churn prediction is the method used to predict the likely customer churn and the customer that persist in a particular company. Churn prediction should be done to find out the possibility of customers switching service. In most cases the churn customer data has a lower number than the non-churn data, this fact raises the problem at the time of classification that is imbalanced data. In dealing with churn problems and imbalanced data used several methods of data mining. Problems with imbalanced data problems, the authors apply the SMOTE technique for data handling. Then to classify churn and non-churn classes using logistic regression method. The logistic regression method is a prediction model used to derive possibilities between two churn values. The data used is customer data from the WITEL PT. Telecommunications Regional 7. The research using logistic regession method and data imbalance handlin with SMOTE has a high performance result with an accuracy of 92,4% and f1-measure of 31,27%.

Keyword: SMOTE, churn, churn prediction, imbalanced data, logistic regression, classification.

Abstrak

Persaingan antara perusahaan telekomunikasi seluler pada masa ini adalah dengan mempertahankan pelanggan. Pelanggan menjadi salah satu faktor utama dalam kesuksesan yang dicapai dalam perusahaan telekomunikasi seluler. Pelanggan dapat memilih sesuai dengan keinginan dan kebutuhan, hal ini menjadi faktor utama pemicu terjadinya *churn. Churn prediction* adalah metode yang digunakan untuk memprediksi pelanggan yang kemungkinan *churn* dan pelanggan yang tetap bertahan pada suatu perusahaan tertentu. *Churn prediction* harus dilakukan untuk mengetahui kemungkinan pelanggan berpindah layanan. Dalam sebagian besar kasus data pelanggan *churn* memiliki jumlah lebih rendah dibanding dengan data non-*churn,* fakta ini memunculkan permasalahan pada saat klasifikasi yaitu *imbalanced data.* Dalam menangani permasalahan *churn* dan *imbalanced data* digunakan beberapa metode *data mining.* Permasalahan pada *imbalanced data,* penulis menerapkan teknik *SMOTE* untuk penanganan data. Kemudian untuk mengklasifikasikan kelas *churn* dan non-*churn* menggunakan metode *logistic regression.* Metode *logistic regression* merupakan model prediksi yang digunakan untuk mendapatkan kemungkinan diantara dua nilai *churn*. Data yang digunakan adalah data pelanggan dari WITEL PT.Telekomunikasi Regional 7. Penelitian yang dilakukan menggunakan metode *logistic regression* dan penanganan *imbalance data* dengan SMOTE memiliki hasil performansi dengan tingkat akurasi sebesar 92,4% dan f1-measure sebesar 31,27%

Kata kunci: ***SMOTE, churn, churn prediction, imbalanced data, logistic regression, klasifikasi***

1. Pendahuluan

Persaingan antara perusahaan telekomunikasi seluler pada masa ini adalah bagaimana cara mempertahankan pengguna layanan yang masih aktif dalam menggunakan layanan dan menarik pengguna baru

yang berpotensi menjadi pelanggan. Pelanggan menjadi salah satu faktor utama dalam kesuksesan yang dicapai dalam perusahaan telekomunikasi seluler [1]. Munculnya berbagai jenis penyedia jasa telekomunikasi dapat menjadi faktor perpindahan pelanggan menjadi masalah utama atau yang sering disebut *churn.*

Churn adalah salah satu aksi yang berkaitan dengan pelanggan, yaitu pelanggan beralih dari satu perusahaan ke perusahaan yang lain [2]. *Costumer churn* merupakan masalah yang sangat diperhitungkan oleh perusahaan, terdapat proses yang dimiliki oleh perusahaan untuk menangani masalah ini yaitu *costumer management*. Proses ini bertujuan untuk mempertahankan pelanggan yang menguntungkan. Perusahaan menyadari bahwa strategi bisnis yang berorientasi pelanggan akan sangat penting untuk mempertahankan keuntungan [1].

Salah satu strategi yang dibutuhkan dalam penanganan kasus *churn* adalah *Costumer Relationship Management* (CRM). Strategi ini mendapatkan informasi lansung dari masing-masing pelanggan, data yang dimasukan dalam proses registrasi. Strategi ini dapat dilakukan untuk mengidentifikasi pelanggan yang kemungkinan akan membatalkan kontrak mereka [3]. *Churn prediction* harus dilakukan untuk mengetahui kemungkinan pelanggan berpindah layanan. Bayak perusahaan yang menggunakan teknik data mining untuk analisis churn. Berbagai teknik data mining seperti *sequential patterns, genetic modelling, classification trees, neural networks, and SVM* telah dilakukan untuk mengeksplorasi *churn* [4].

Penerapan teknik *churn prediction* menggunakan metode data mining untuk melakukan klasifikasi, dan terdapat masalah yang timbul pada kasus ini yaitu *imbalance data.* Masalah ini disebabkan karena jumlah dari data *non-churn* dibandingkan data *churn* tidak seimbang [5]. Dengan perbandingan data yang tidak seimbang maka akan mempengaruhi metode klasifikasi dari data mining untuk memprediksi kelas *churn*, akurasi tinggi sering dicapai dalam klasifikasi tanpa penanganan *imbalance data* karena hanya berfokus pada data mayoritas dan untuk data minoritas dianggap sebagai data langka atau data tidak disengaja [6].

Pada penelitian Tugas Akhir ini menggunakan metode *Logistic Regression* (LR) sebagai metode klasifikasi. *Logistic regression* merupakan model regresi biner statistik standar [2]. Metode *logistic regression* mempelajari hubungan antara variabel dependen dan beberapa variabel independen. Nama regresi logistik digunakan apabila variabel dependen hanya memiliki dua nilai, contohnya 0 dan 1 [7]. Kegunaan regresi logistik yang dapat melakukan klasifikasi dengan nilai dependen variabel 0 dan 1 akan sangat berguna bagi *churn prediction* yang memiliki nilai variabel dependen atau kelas *churn* 0 dan 1.

Salah satu metode yang dapat digunakan untuk menangani *imbalance data* adalah metode *Synthetic Minority Over-sampling Technique* (SMOTE). SMOTE merupakan salah satu metode *oversampling* untuk menangani masalah *imbalance data* dengan melakukan peningkatan pada data minor dengan cara duplikasi. Dalam hal sampling sintetis merupakan metode yang ampuh, telah terbukti pada setiap penelitian [8].

2. Metodologi Penelitian

2.1 Dataset

Pada penelitian Tugas Akhir ini digunakan dataset yang berasal dari WITEL PT. Telekomunikasi Indonesia Regional 7. Dataset memiliki 53 atribut dengan 2 kelas label *churn* dan *non-churn* berjumlah 200,361 *data* (192,848 *data* data *non-churn* dan 7,513 *data* data *churn*), ratio terdapat kelas *churn* hanya antara 3% dari semua data, terdapat dataset yang bersifat diskrit dan kontinu, deskripsi data yang akan diolah terdapat pada tabel 1.

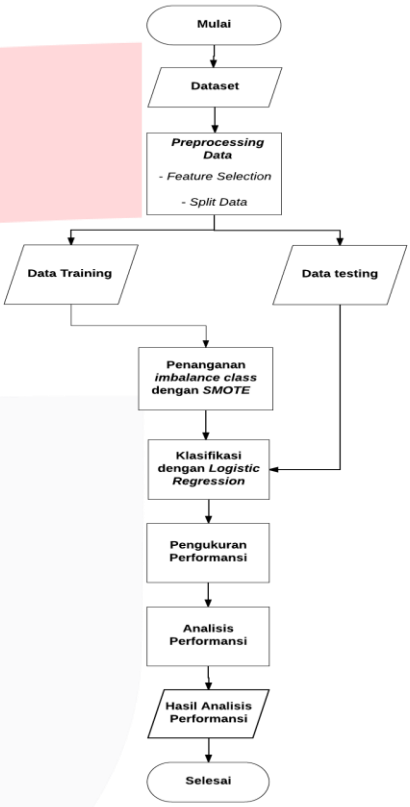
Tabel 1. Data Pelanggan

Atribut	Keterangan
SEGMEN_ID	Segmen pelanggan.
UMUR_PLG	Lama pelanggan menggunakan layanan.
PAKET_SPEEDY_ID	Paket internet yang digunakan pelanggan.
WITEL	Area regional 7.
TAG_N	Jumlah tagihan yang dibayarkan pada bulan tertentu selama 12 bulan. Terdapat 12 atribut TAG_N yaitu TAG_N sampai TAG_N11.
STATUS_BAYAR_N	Pembayaran pelanggan pada bulan tertentu selama 12 bulan. Terdapat 12 atribut STATUS_BAYAR_N yaitu

	STATUS_BAYAR_N sampai STATUS_BAYAR_N11.
GGN_N	Jumlah complain pelanggan pada bulan tertentu selama 12 bulan. Terdapat 12 atribut GGN_N yaitu GGN_N sampai GGN_N11.
USAGE_N	Total penggunaan layanan pelanggan pada bulan tertentu selama 12 bulan. Terdapat 12 atribut USAGE_N yaitu USAGE_N sampai USAGE_N11.
CHURN	Status pelanggan.

2.2 Gambaran Umum Sistem

Gambar 1 adalah flowchart gambaran umum sistem yang dibangun dalam Tugas Akhir ini.



Gambar 1 gambaran umum sistem

2.3 Deskripsi Tahapan Proses

Deskripsi dari masing masing tahapan dalam gambaran umum sistem akan dijelaskan dalam subbab dibawah ini.

2.3.1 Preprocessing Data

Terdapat dua tahap dalam preprocessing dalam Tugas Akhir ini, yaitu:

- **Seleksi Atribut**

Pada Tugas Akhir ini digunakan dataset dengan atribut sebanyak 53, pada tahap ini digunakan korelasi pearson untuk mengetahui keterkaitan antara atribut. 10^{-2} dan $5.10^{-2.}$ Batasan korelasi 10^{-2} menghasilkan 31 atribut, sedangkan batasan korelasi 5.10^{-2} menghasilkan 7 atribut. Berikut adalah penjabaran dari atribut dari hasil seleksi atribut:

- Data 7 atribut: paket speedy id, status bayar n, status bayar n-1, status bayar n-2, status bayar n-3, usage n, churn.
- Data 31 atribut: umur plg, paket speedy id, witel, tag n, tag n-2, tag n-3, tag n-4, tag n-5, tag n-6, tag n7, tag n-8, tag n-11, status bayar n, status bayar n-1, status bayar n-2, status bayar n-3, status bayar n-4, status bayar n-5, status bayar n-6, status bayar n-7, status bayar n-8, status bayar n-9, status bayar n-10, status bayar n-11, ggn n, usage n, usage n-1, usage n-2, usage n-3, usage n-4, churn.

- **Partisi Data**

Pada tahap ini dilakukan pemisahan data atau yang disebut partisi data dengan maksud membuat data *training* dan data *testing*. Partisi data dalam tugas akhir ini menggunakan komposisi data *training* dan *testing* dengan perbandingan 70:30, 80:20, 90:10. Deskripsi hasil dari komposisi data dapat dilihat pada tabel 2.

Tabel 2. Partisi Data

Komposisi Data	Jenis Data	Jumlah Data	Data nonchurn	Data churn
70:30	Data training	140,253 data	135,053 data	5,200 data
	Data testing	60,108 data	57,810 data	2,298 data
80:20	Data training	160,289 data	154,306 data	5,983 data
	Data testing	40,072 data	38,557 data	1,515 data
90:10	Data training	180,325 data	173,574 data	6,751 data
	Data testing	20,036 data	19,289 data	747 data

2.3.2 Menangani imbalance data menggunakan SMOTE

Pada tahap penanganan *imbalance data* yaitu dilakukan proses membuat data menjadi seimbang, dikarenakan pada kasus *churn prediction* pasti tingkat data *churn* akan lebih sedikit dibandingkan dengan data non-*churn*. Dengan menggunakan metode SMOTE akan dilakukan penyeimbangan data dengan cara pembuatan data *synthetic data* dari data minoritas dan untuk penentuan data mana yang menjadi kandidat untuk diduplikat akan digunakan *k nearest neughbors* dengan nilai k = 3,5, dan 7. Pemilihan data duplikasi dilakukan secara random. Jumlah data duplikasi sesuai dengan nilai N yang diinput, representasi nilai N berlaku untuk setiap data. Metode SMOTE dilakukan berulang-ulang sehingga kesenjangan antara data minoritas dengan data mayoritas tidak berbeda jauh. Gambar 2 merupakan flowchart proses SMOTE. Pada tugas akhir ini dilakukan penanganan *imbalance data* hanya pada data training, duplikasi data dilakukan sesuai dengan langkah-langkah sesuai pada gambar 2. Berikut deskripsi data hasil SMOTE pada data training ditampilkan pada Tabel 3.

gambar 2 SMOTE

Tabel 3. Duplikasi Data Training

Jenis data	Jumlah atribut	Duplikasi	K	Data *churn*	Data *churn*+SMOTE
Data training (70%)	7	50 kali	3,5,7	2509 data	127,959 data
	31	25 kali	3,5,7	5191 data	134,966 data
	53	25 kali	3,5,7	5200 data	135,200 data
Data training (80%)	7	51 kali	3,5,7	2,869 data	149,188 data
	31	25 kali	3,5,7	5,970 data	155,220 data
	53	25 kali	3,5,7	5,983 data	155,558 data
Data training (90%)	7	50 kali	3,5,7	3,195 data	162,945 data
	31	25 kali	3,5,7	6,736 data	175,136 data
	53	25 kali	3,5,7	6,751 data	175,526 data

2.3.3 Membangun Model Prediksi

Membangun model untuk prediksi data pelanggan *churn* atau non-*churn* pada perusahaan PT. Telekomunikasi dikhususkan pada metode *logistic regression*. Pada kasus data PT. Telekomunikasi tipe data dari variabel respon (y) merupakan data nominal yang hanya merepresentasikan 2 jenis data. Untuk variabel respon terdapat y=1 sebagai kelas *churn* dan y=0 sebagai kelas *non-churn*. Pada tahap ini digunakan *logistic regression* sebagai klasifikasi untuk mengetahui output prediksi. Dari hasil data penanganan *imbalance data* dengan SMOTE dari data 53 atribut, 31 atribut, 7 atribut akan dilakukan prediksi dengan *logistic regression*. Metode ini merupakan metode yang memiliki output variabel respon biner yaitu y=1 dan y=0. Hasil dari model ini merupakan sebuah persamaan $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_{53} X_{53}$ dimana $\beta_0, \beta_1, \beta_2, \ldots\ldots, \beta_{53}$ merupakan koefisien regresi untuk setiap atribut. Salah satu model yang dibentuk dan memiliki nilai terbaik adalah data dengan 31 atribut dan ketetangaan 5 adalah:

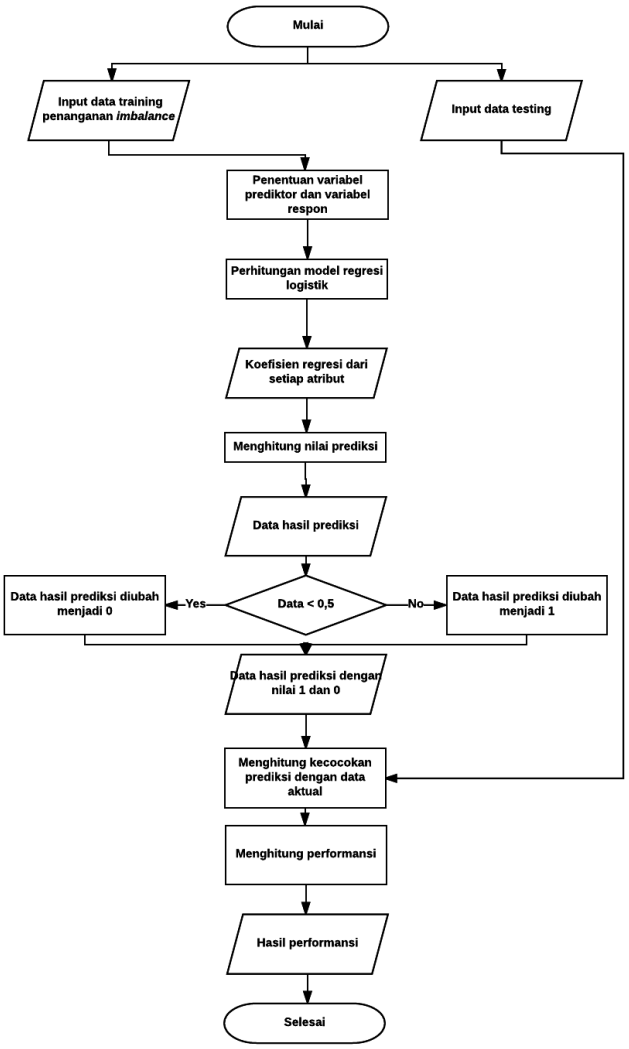
$$g(x) = \frac{1}{1+\exp[(\beta_0 + 3{,}865X_1 + -0{,}1085X_2 + -0{,}0221X_3 + \cdots . + 8{,}3918X_{31})]}$$

Persamaan diatas dapat dimasukan kedalam persamaan (2.4) dengan menggunakan persamaan logistik. Sehingga diperoleh persamaan sebagai berikut:

$$g(x) = \exp[(\beta_0 + 3{,}865X_1 + -0{,}1085X_2 + -0{,}0221X_3 + \cdots . + 8{,}3918X_{31} \quad (2.1)$$

Dalam perhitungan estimasi parameter menggunakan persamaan (2.6) didapat nilai β untuk setiap variabel independen yang berguna untuk memaksimalkan nilai model .Untuk mendapatkan nilai prediksi digunakan fungsi sigmoid yang memiliki rentang antara 0-1. Terdapat batas antara 1-0 yang menjadi batas nilai keluaran yaitu 0,5. Batas ini diambil karenan rentang yang dimiliki adalah 0-1 dan 0,5 merupakan batas tengah dari nilai rentang. Keluaran persamaan > 0,5 akan diubah menjadi 1 dan keluaran persamaan dengan < 0,5 akan diubah menjadi 0.

Nilai ini merupakan nilai prediksi yang akan dibandingkan dengan data aktual atau data testing. Gambar 3 merupakan flowchart proses *logistic regression*.



Gambar 3 logistic regression

2.3.4 Evaluasi Model Klasifikasi

Salah satu acuan dalam data mining adalah matriks evaluasi, beberapa perhitungan untuk mengevaluasi hasil dapat diketahui dengan matriks evaluasi yaitu, recall, precision, F1-Measure dan akurasi. Semua nilai dinyatakan dalam bentuk persentase(%).Berikut deskripsi matriks evaluasi ditampilkan pada Tabel 4.

Tabel 4. Matriks Evaluasi

Aktual/prediksi	Kelas minor	Kelas mayor
Kelas minor	TP(True Positive)	FN(False Negative)
Kelar mayor	FP(False Positive)	TN(True Negative)

Keterangan:

- True Positive (TP) : Jumlah *instance* kelas positif yang benar diprediksi sebagai kelas positif
- False Positive (FP) : Jumlah *instance* kelas negatif yang diprediksi sebagai kelas positif
- False Negative (FN) : Jumlah *instance* kelas positif yang diprediksi sebagai kelas negatif
- True Negative (TN) : Jumlah *instance* kelas negatif yang benar diprediksi sebagai kelas negatif

Terdapat pengukuran performansi adalah Recall, Precision, F1-measure, dan Akurasi

a). Recall

Recall dihitung untuk mengevaluasi seberapa besar *coverage* suatu model dalam memprediksi suatu kelas tertentu. *Recall* didapatkan dengan menghitung perbandingan antara jumlah data untuk satu kelas tertentu yang diprediksi dengan benar dibagi jumlah total kelas tersebut [4].

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (2.2)$$

b). Precision

Precision dihitung untuk mengevaluasi seberapa baik kepastian model dalam memprediksi suatu kelas dengan benar. *Precision* merupakan perhitungan untuk mendapat nilai antara perbandingan antara junlah data untuk satu kelas tertentu yang diprediksi dengan benar dibagi jumlah keseluruhan prediksi kelas. [9]. Rentang nilai precision adalah 0 sampai 1, jika nilai mendekati 0 maka ketetapan dalam memprediksi tidak baik dan mendekati 1 jika ketetapan dalam memprediksi baik. Untuk mendapatkan nilai dalam persentase maka nilai dikalikan dengan 100.

$$\textbf{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (2.3)$$

c). F1-Measure

F1-Measure adalah perhitungan kombinasi antara *recall* dan *precision.*Rentang nilai precision adalah 0 sampai 1, jika nilai mendekati 0 maka model prediksi tidak baik dan mendekati 1 jika model prediksi baik. Untuk mendapatkan nilai dalam persentase maka nilai dikalikan dengan 100.

$$\textbf{F1 – Measure} = \frac{\boldsymbol{2*Precision*Recall}}{\boldsymbol{Precision+Recall}} \quad (2.4)$$

d). Akurasi

Akurasi adalah nilai ketepatan model memprediksi data dengan perbandingan data aktualnya dan sebagai pengukur model seberapa akurat untuk melakukan prediksi. Terdapat kasus yang menunjukan nilai akurasi tinggi namun buruk terhadap prediksi *churn*, yaitu ketika jumlah data *non-churn* memiliki ketimpangan yang sangat tinggi. pengukuran yang lebih tepat dalam model prediksi tanpa *imbalance data* adalah *F1-Measure.*

$$\boldsymbol{Akurasi} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FN+FP}} \quad (2.5)$$

2.3.5 Skenario Pengujian

Skenario yang akan diuji pada sistem adalah:

1) Pengujian dilakukan dengan percobaan penanganan *imbalance data* menggunakan metode SMOTE hanya untuk dilakukan pada data training dan klasifikasi menggunakan metode *logistic regression*. Tujuan dari skenario ini untuk melihat pengaruh dari *logistic regression*
2) Pengujian dilakukan percobaan dengan nilai k = 3,5,7 pada SMOTE atau jumlah tetangga yang dimiliki pada data sebagai obyek dengan menggunakan penanganan *imbalance data* menggunakan SMOTE hanya pada data training dan klasifikasi menggunakan *logistic regression.*

Pada masing masing skenario akan dianalisis performansi model klasifikasi yang dibangun kemudian dinyatakan dengan akurasi, dan nilai *f1-measure* pada prediksi *churn.*

3. Hasil dan Pembahasan

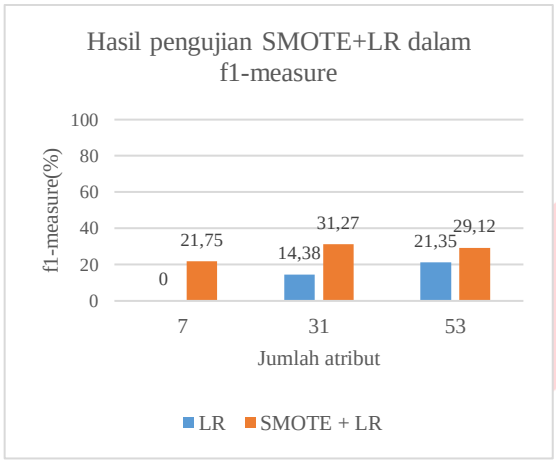
Pada subbab ini akan dijelaskan mengenai pengujian model sistem dan analisis model sistem yang telah dibuat. Pengujian ini dilakukan dengan beberapa skenario yang telah dirancang pada subbab sebelumnya.

a. Pengaruh klasifikasi Logistic Regression (LR) terhadap hasil performansi

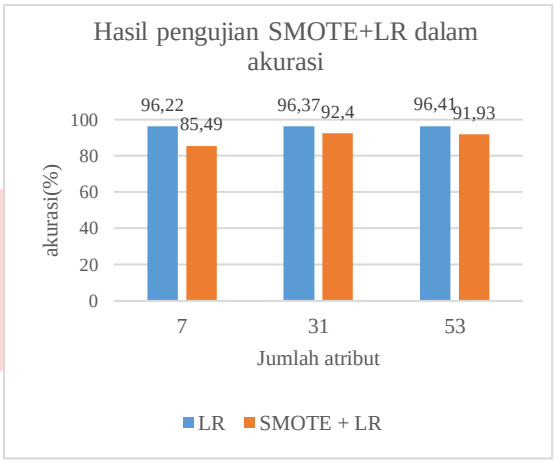
Pengujian ini dilakukan untuk melihat pengaruh klasifikasi *logistic regression* terhadap hasil performansi terhadap data yang telah dilakukan penanganan *imbalance data.* Penanganan *imbalance data* dalam pengujian ini menggunakan metode SMOTE untuk data trainingnya dengan nilai ketetangaan 3,5,7. partisi data yang dilakukan pengujian adalah 70:30, untuk aribut menggunakan 7 atribut, 31 atribut dan 53 atribut. Perbandingan yang dilakukan dengan nilai rata-rata *F1-Measure* dan akurasi pada masing-masing komponen pengujian. Tujuan dari pengujian ini adalah untuk mendapatkan nilai performansi dari

klasifikasi *logistic regression* dengan penanganan *imbalance class*. Hasil pengujian ini disajikan pada Gambar 4. Berdasarkan gambar 4 nilai *F1-Measure* tertinggi untuk klasifikasi *logistic regression* dengan penanganan *imbalance data* menggunakan SMOTE adalah pada atribut 31 yaitu sebesar 31,27% dan untuk yang terkecil terdapat pada atribut 7 dengan 21,75. *Logistic regression* menggunakan estimasi koefisien regresi yang didapatkan dari masing-masing atribut untuk kemudian didapatkan nilai(z) yang didapatkan dari fungsi sigmoid *f(z)*. Setelah didapat nilai prediksi maka di sortir dengan fungsi sigmoid yaitu 0.5 dan dihasilkan nilai Y atau output prediksi yang akan dicocokan dengan data aktual. Performansi klasifikasi dihasilkan dengan pengaruh penggunaan estimasi koefisien regresi dari *logistic regression* dari setiap atribut yang sangat berpengaruh.

.



gambar 4 Hasil Pengujian SMOTE+LR dalam f1-measure

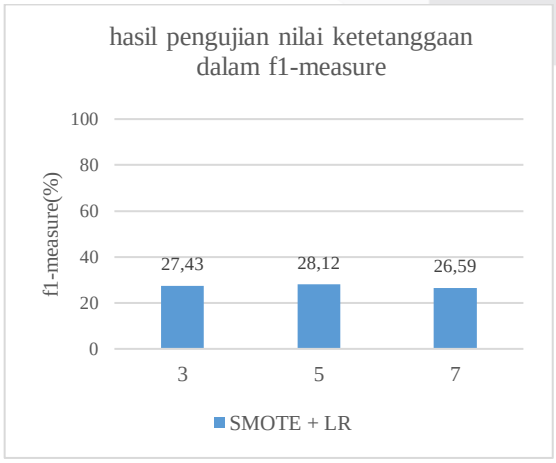


gambar 5 Hasil Pengujian SMOTE+LR dalam akurasi

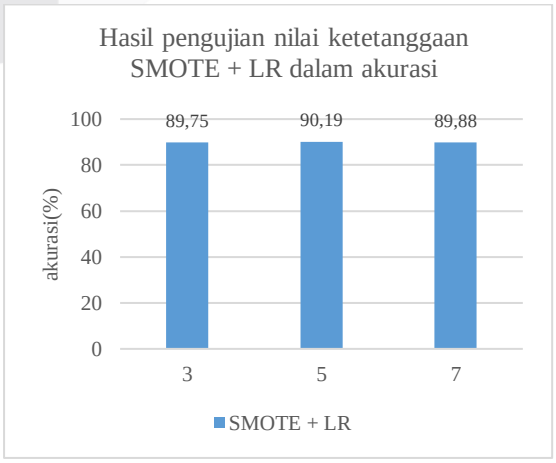
Berdasarkan gambar 5 nilai akurasi tertinggi untuk klasifikasi *logistic regression* dengan penanganan *imbalance data* menggunakan SMOTE adalah pada atribut 31 yaitu sebesar 92,4% dan untuk yang terkecil terdapat pada atribut 7 dengan 85,49%. Sesuai dengan pengujian sebelumnya bahwa atribut 31 memiliki performansi yang terbaik walaupun perbedaan dengan atribut 53 tidak terlalu jauh.

b. Pengaruh nilai ketetanggaan SMOTE terhadap hasil klasifikasi

Pengujian ini dilakukan untuk melihat pengaruh nilai ketetanggan terhadap model klasifikasi dengan menggunakan data yang memiliki komposisi data 70:30 dan data 7,31,53 atribut. Analisis pengujian menggunakan *Logistic Regression* dengan penanganan *imbalance data* menggunakan SMOTE dengan jumlah ketetanggaan yaitu k = 3,5,7, partisi data yang dilakukan pengujian adalah 70:30 dan data 7,31,53 atribut. Perbandingan yang dilakukan dengan nilai *F1-Measure* dan akurasi pada masing-masing jumlah atribut. Tujuan dari pengujian ini adalah untuk mendapatkan nilai performansi terbaik antara nilai ketetangaan k = 3,5,7. Hasil pengujian ini disajikan pada Gambar 7.



gambar 6 hasil pengujian nilai k dalam f1-measure



gambar 7 hasil pengujian nilai k dalam akurasi

Berdasarkan gambar 7 dapat dilihat bahwa untuk model SMOTE+LR yang menggunakan nilai ketetangaan k=5 memiliki nilai *F1-Measure* yang tinggi dibanding nilai ketetangaan lainnya. Untuk ketetanggaan bernilai 3 dan 7 cenderung mengalami penurunan. Pengujian ini menunjukan bahwa nilai ketetanggaan tidak terlalu berpengaruh terhadap nilai performansi karena data yang dimiliki bersifat random. Berdasarkan gambar 6 dapat dilihat bahwa hasil pengujian berdasarkan nilai ketetangaan yang memiliki akurasi tinggi adalah k = 5 dengan nilai 90,19%, namun perbedaan yang dimiliki dengan k = 7 tidak terlalu besar yaitu 89,88%. pengujian ini menunjukan tidak terdapat perbedaan yang sangat berarti dalam hal peningkatan data yang dilakukan terhadap obyek pengujian k = 5 dan k = 7. Nilai terkecil dalam aspek akurasi terdapat pada k = 3 yaitu 89,75%.

4. Kesimpulan dan Saran

Berikut adalah kesimpulan yang didapatkan dari hasil penelitian yang telah dilakukan pada Tugas Akhir ini:

1. Penanganan masalah *imbalance data* menggunakan SMOTE pada data training berpengaruh terhadap hasil klasifikasi jika dibandingkan tanpa penanganan *imbalance data* atau hanya menggunakan klasifikasi *logistic regression.*
2. Penanganan masalah *imbalance data* menggunakan SMOTE pada data training sangat berpengaruh terhadap hasil klasifikasi jika dibandingkan tanpa penanganan *imbalance data* atau hanya menggunakan klasifikasi *logistic regression*. pengujian ini menunjukan bahwa penggabungan antara SMOTE dengan *logistic regression* dapat menghasilkan performansi nilai rata-rata *F1-Measure* sebesar 31,27%. Sedangkan hasil pengujian klasifikasi *logistic regression* tanpa penangana *imbalance data* memiliki nilai rata-rata 14,38%.
3. Estimasi koefisien setiap atribut independen yang memberikan pengaruh terhadap model yang dibangun dan hasil performansi klasifikasi untuk prediksi churn.

Saran untuk pengembangan Tugas Akhir adalah:

1. Data yang digunakan memiliki waktu yang lebih lama dibandingkan data pada penelitian ini.
2. Melakukan pemisahan data yang bersifat katagorikal dan menganalisis pengaruhnya terhadap kelas *churn.*
3. Penanganan seleksi atribut untuk data kategorikal.
4. Nilai ketetangaan diperbesar kelipatannya.
5. Dilakukan korelasi spearman untuk data kategorikal.

5. References

[1] Chih-Fong Tsai, Yu-Hsin Lu , "Customer churn prediction by hybrid neural networks," *Expert Systems with Applications* , vol. 36, p. 12547–12553, 2009.

[2] Nicolas Glady, Bart Baesens, Christophe Crouxa , "Modeling Churn Using Customer Lifetime Value," *European Journal of Operational Research* , 2009.

[3] Vladislav Lazarov, Marius Capota, "Churn Prediction," *Business Analytics Course. TUM Computer Science,* 2007.

[4] Yaya Xie a, Xiu Li, E.W.T. Ngai , Weiyun Ying , "Customer churn prediction using improved balanced random forests," *Expert Systems with Applications,* vol. 36, p. 5445–5449, 2009.

[5] Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *Sigkdd Explorations.,* vol. 6, no. 1, p. 1, 2004.

[6] Haibo He, Member, IEEE, and Edwardo A. Garcia, "Learning from Imbalanced Data," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,* vol. 21, no. 9, pp. 1041-4347, 2009.

[7] NCSS, Logistic Regression, NCSS.

[8] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, p. 321–357, 2002.

[9] Angelina Sagita Sastrawan, ZK. Abdurahman Baizal, Moch. Arif Bijaksana, "ANALISIS PENGARUH METODE COMBINE SAMPLING DALAM CHURN PREDICTION UNTUK PERUSAHAAN TELEKOMUNIKASI," *Seminar Nasional Informatika 2010 (semnasIF 2010)*, 2010.