Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging

[1]Tesha Tasmalaila Hanifa, [2]Adiwijaya, [3]Said Al-Faraby

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

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

Abstrak

Perkembangan teknologi sekarang ini semakin pesat. Kebutuhan akan informasi dan komunikasi bertambah. Persaingan untuk mendapatkan pendapatan antar perusahaan telekomunikasi menimbulkan adanya Churn. Churn adalah pindahnya pelanggan dari satu provider ke provider lainnya. Perusahaan lebih memilih untuk mempertahankan customer, karena dibutuhkan biaya yang lebih sedikit daripada menambah customer yang baru. Namun dalam permasalahan ini churn memiliki data yang tidak seimbang dan ekstrim dibanding dengan non-churn, sehingga perlu adanya penanganan pada distribusi kelas mayoritas (non-churn) dan minoritas (churn). Pada metode pendekatan dengan data mining, data yang tidak seimbang akan mengakibatkan proses klasifikasi yang cenderung keliru pada kelas minoritas (churn). Oleh karena itu, diperlukan penanganan kelas tidak seimbang dengan teknik sampling. Metode yang digunakan pada penelitian Tugas Akhir ini adalah metode Underbagging untuk menangani imbalance data yang dikombinasikan dengan metode klasifikasi menggunakan Logistic Regression (LR). Pengujian dilakukan dengan menggunakan dataset pelanggan WITEL PT. Telekomunikasi Regional 7 dengan 53 atribut. Jumlah data churn 7.513 record dan data non-churn 192.848 record. Penelitian ini menghasilkan nilai performansi akurasi tertinggi sebesar 85,531% dan meningkatkan nilai f1-measure lebih dari 20% terhadap hasil klasifikasi tanpa penanganan imbalance data.

Kata Kunci: klasifikasi, data mining, churn prediction, logistic regression, imbalance data, underbagging.

1. Pendahuluan

Perkembangan teknologi sekarang ini semakin pesat. Kebutuhan akan informasi dan komunikasi bertambah. Telah ditemukan berbagai perangkat teknologi yang memudahkan manusia dalam menyelesaikan masalah informasi dan komunikasi. Perusahaan telekomunikasi di Indonesia sekarang sedang melakukan banyak inovasi untuk melakukan persaingan yang semakin ketat karena customer berhak memilih dari banyaknya penyedia layanan yang ada. Persaingan ini penting bagi kelangsungan perusahaan karena berpengaruh kepada pendapatan perusahaan.

Persaingan tersebut mengakibatkan beralihnya customer ke perusahaan telekomunikasi lain atau bisa disebut Churn. Churn adalah pemutusan jasa suatu perusahaan oleh pelanggan karena pelanggan tersebut lebih memilih menggunakan layanan jasa perusahaan kompetitor. Pada persaingan pasar ini dilihat dari pengalaman bahwa setiap tahunnya sekitar 30-35% laju churn dan membutuhkan 5-10 kali usaha dan biaya untuk menambah customer baru daripada mempertahankan yang sudah ada. Untuk itu industri telekomunikasi lebih memilih untuk mempertahankan customer. [1]

Dalam mempertahankan customer, perusahaan telekomunikasi membutuhkan cara untuk memprediksi dalam mengetahui resiko customer kapan akan menjadi churn. Ketika memprediksi churn, terdapat berbagai teknik data mining dapat diterapkan. Salah satunya adalah model prediksi churn. Perusahaan besar harus mengimplementasikan model churn prediction tersebut untuk dapat mendeteksi atau mengetahui kemungkinan adanya churn sebelum mereka secara efektif meninggalkan perusahaan tersebut yang bisa berpengaruh terhadap pendapatan perusahaan. Namun, pada dataset pelanggan yang didapat dari perusahaan tersebut seringkali masih tidak seimbang. Data tidak seimbang merupakan jika kelas mayoritas (not churn) lebih banyak dari kelas minoritas (churn). Data yang tidak seimbang akan mengakibatkan proses klasifikasi yang cenderung keliru pada kelas minoritas (churn). Oleh karena itu, untuk

keberhasilan pada proses klasifikasi tersebut diperlukan adanya penanganan masalah kelas tidak seimbang dahulu sebelum membuat model prediksi.

Pada permasalahan imbalance data dan churn diatas, penulis menerapkan teknik underbagging dan logistic regression. Metode underbagging merupakan metode penggabungan (ensemble) antara undersampling dan bagging. Metode Logistic Regression termasuk dalam predictive modeling, karena dalam statistika digunakan untuk prediksi probabilitas kejadian suatu peristiwa dengan mencocokkan data pada fungsi logit kurva logistik dalam suatu kelas. [2] Metode ini bisa digunakan untuk penanganan kasus prediksi churn karena pada kasus data pelanggan PT. Telekomunikasi dapat diketahui bahwa tipe data variabel respon (Y) adalah nominal. Terdapat dua kemungkinan, yaitu churn dan not churn yang menghasilkan binary values seperti angka 0 adalah non churn dan angka 1 mempresentasikan churn. [2]

Dengan penelitian mengunakan metode penanganan kelas tidak seimbang underbagging dan teknik klasifikasi logistic regression, diharapkan dapat melihat pengaruh metode tersebut pada akurasi dan F1-measure agar perusahaan Telekomunikasi dapat mencegah terjadinya churn pelanggan.

2. Metode Penelitian

2.1 Dataset

Dataset yang digunakan adalah data pelanggan PT. Telekomunikasi Indonesia Regional 7. Data yang digunakan adalah gabungan dari data diskrit dan kontinu, yang memiliki 53 atribut dengan 2 kelas label *churn* dan *non churn* dengan jumlah 200.381 *record*. Dengan rincian data *churn* 7.513 *record* dan data *non-churn* 192.848 *record*. Data *churn* dinyatakan dengan angka 1 dan angka 0 jika *non churn*. Jumlah data *churn* hanya berkisar 3% dari keseluruhan data. Data yang akan diolah dideskripsikan pada tabel 1.

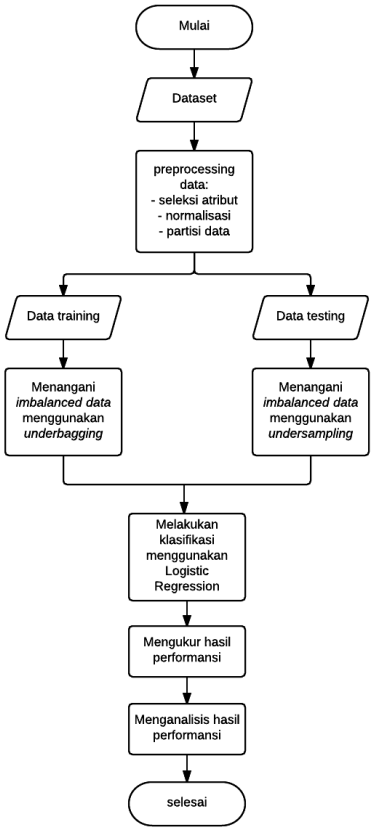
Tabel 1. Data Pelanggan

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 merupakan gambaran umum sistem yang dibangun dari penelitian Tugas Akhir ini.



Gambar 1. Gambaran Umum

2.3 Deskripsi Tahapan Proses

Gambaran mengenai masing-masing proses yang akan dilakukan pada saat membangun model klasifikasi akan dijelaskan pada sub bab-sub bab dibawah ini.

2.3.1 Data Preprocessing

Terdapat beberapa tahapan preprocessing data yang dilakukan pada penelitian ini, yaitu:

- Seleksi atribut

Penelitian Tugas Akhir ini digunakan atribut asli 53 atribut dan seleksi atribut. Pada tahap ini digunakan korelasi Pearson. 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.

- Normalisasi

Pada tahap ini dilakukan normalisasi min-max pada seluruh atribut. Normalisasi dilakukan agar proses latih data menjadi lebih cepat atau meningkatkan kinerja model klasifikasi dan membantu model dalam memahami data untuk proses pengklasifikasian. Data atribut diskalakan agar sesuai dengan rentang 1-0 karena output prediksi yang hendak didapatkan harus berada pada rentang tersebut. Misalnya, dalam dataset ini, Data atribut diskalakan agar sesuai dengan rentang 1-0. Dengan demikian, adanya pemahaman jenis pelanggan dengan nilai 1 untuk melakukan keputusan *churn*.

- Normalization Min-Max Method
  Metode Normalisasi Min-Max merupakan salah satu metode mengubah data yang kompleks dengan tidak menghilangkan isi, sehingga lebih mudah diolah. Dilakukan dengan cara standarisasi data dengan menempatkan data dalam *range* 0 sampai 1, dengan nilai terkecil sebagai 0, dan nilai terbesar sebagai 1. Metode ini memberikan keseimbangan pada data satu dengan yang lainnya.
  Rumus perhitungan metode normalisasi min-max:

$$v' = \frac{v - min_A}{max_A - min_A} \ (range.max - range.min) \ + \ range.min$$

v' merupakan data setelah normalisasi, v data sebelum normalisasi, min_A nilai minimal pada kolom/atribut sebelum normalisasi, max_A nilai maksimal pada kolom/atribut sebelum normalisasi, *range min* adalah 0, *range max* adalah 1.

- Partisi Data

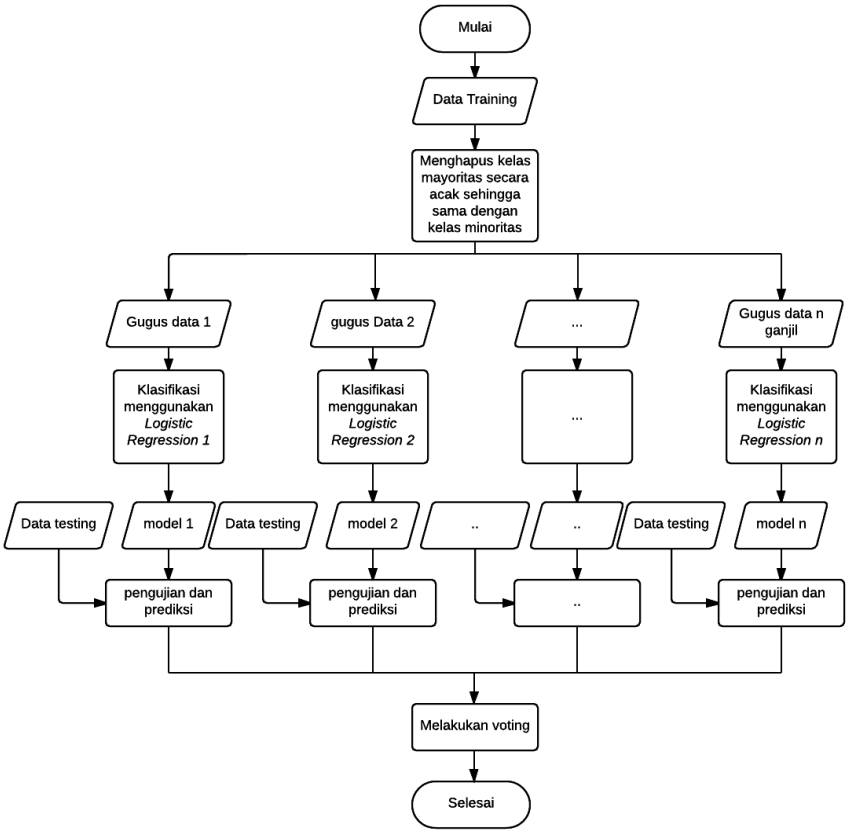
Pada tahap ini dilakukan partisi data untuk membuat data *training* dan data *testing*. Data *training* menyediakan bahan baku untuk membentuk model prediktif. Data *testing* digunakan sebagai pengukuran kinerja model klasifikasi. Pada tugas akhir ini menggunakan komposisi data training dan testing dengan perbandingan 70:30. Penggunaan data testing yang tidak terlalu sedikit yaitu 30% akan menghasilkan hasil performansi yang baik.

Tabel 2. Jumlah Data

Komposisi Data	Jenis Data	Jumlah Data
70:30	Data training	140.253 *record*
	Data testing	60.108 *record*

2.3.2. Menangani Imbalance Data menggunakan Underbagging

Pada proses ini, mengubah dataset yang digunakan untuk membangun model prediksi *logistic regression* untuk memiliki data yang lebih seimbang. Perubahan ini disebut *sampling*. Pada Tugas Akhir ini menggunakan teknik underbagging. *Underbagging* adalah metode gabungan dari *undersampling* dan *bagging*.



Gambar 2. Flowchart Underbagging

Penjelasan proses flowchart diatas adalah dari data *training* akan dibuat gugus data baru sebanyak n ganjil. Masing-masing gugus data akan dilakukan pengambilan secara acak dari kelas *non churn* dan diletakkan disetiap gugus data yaitu p. Setiap gugus data (p) ini berisi kelas *churn* dan *non churn* dengan jumlah yang sama. Pada data *training* terdapat 135.053 kelas *non churn* dan 5200 kelas *churn.* Untuk mendapatkan data yang seimbang maka data baru yang akan dibuat terdiri dari 5200 *churn* dan secara random diambil 5200 data *non churn* sehingga pada setiap *bags* yang dibuat terdapat 10.400 data. Kemudian setelah gugus data dibangkitkan berlanjut ke proses klasifikasi dengan *logistic regression*. Hasil output prediksi yang dihasilkan dari masing-masing model di gugus data akan dilakukan *voting.* Kemudian hasil v*oting* disesuaikan dengan data yang sebenarnya.

Berikut adalah tabel hasil komposisi data setelah dilakukan proses *underbagging*:

Tabel 3. Komposisi Data Underbagging

Komposisi Data	UNDERBAGGING		Jumlah data tiap bags
	Kelas *Non Churn*	Kelas *Churn*	
70:30	135,053 *record*	5,200 *record*	10,400 record

Selain melakukan penanganan *imbalance data* pada data *training*, dilakukan pula terhadap data *testing*. Namun tidak menggunakan metode *bagging*, melainkan hanya metode *Undersampling* saja. Seperti tersaji pada Gambar 3, langkah pertama *Undersampling* yaitu menghitung selisih antara kelas mayoritas dan minoritas pada data *testing*. Kemudian menghapus kelas mayoritas secara acak sehingga jumlah kelas mayoritas sama dengan minoritas

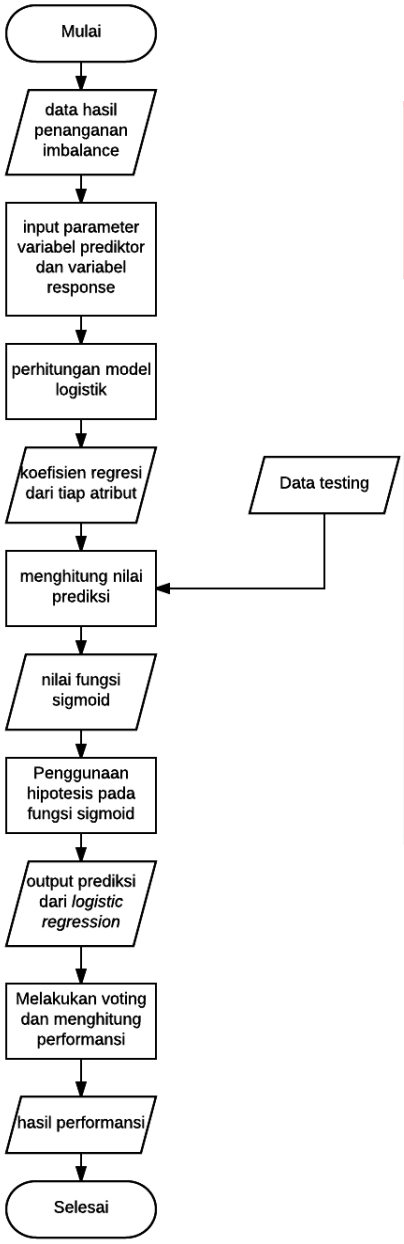
Gambar 3. Tahapan Undersampling

2.3.3. Membuat Model Prediksi dengan Menggunakan Logistic Regression

Pada tahap ini dilakukan prediksi *churn* dengan metode *logistic regression.* Dari data hasil penanganan *imbalance data*, setiap *subset* yang telah dibuat dari 53 atribut, 31 atribut dan 7 atribut akan diprediksi dengan *logistic regression.* Metode *logistic regression* adalah metode yang menghasilkan dua kemungkinan yaitu y=1 atau y=0. y=1 menyatakan *churn* dan y=0 *non* churn. Keluaran dari model ini berupa sebuah persamaan yang diantaranya adalah nilai $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_{53} X_{53}$. Dimana $\beta_0, \beta_1, \beta_2, \ldots\ldots, \beta_{53}$ merupakan koefisien regresi untuk setiap atribut yang digunakan dalam Tugas Akhir ini. Salah satu model yang dibentuk dari model terbaik dengan 53 atribut di gugus data 1 adalah:

$$Pr(Y=1) = \frac{1}{1+\exp[-(3.218X_1+1.789X_2 + -2.023X_3+\cdots.+40.291X_{53})]} \quad (3.1)$$

Model yang telah didapatkan akan dimasukkan Untuk menentukan nilai hipotesis digunakan fungsi sigmoid atau sering disebut dengan fungsi logit. Fungsi logit mempunyai rentang antara 0-1. Prediksi 1 ketika hipotesis ≥ 0.5 dan prediksi 0 ketika hipotesis ≤ 0.5. Nilai hipotesis ini menghasilkan output prediksi berupa 1 dan 0 yang kemudian akan dibandingkan dengan data aktual atau data *testing* dengan data hasil prediksi. Proses ini akan digambarkan pada Gambar yang merupakan flowchart dari *logistic regression* pada Tugas Akhir ini.



Gambar 4. Flowchart Logistic Regression

2.3.4. Evaluasi Model Klasifikasi

Dalam mengevaluasi performasi *underbagging* dan *logistic regression* dapat diteliti dengan *confusion matrix*. Dengan matriks evaluasi, dapat diketahui beberapa perhitungan untuk mengevaluasi hasil *Recall*, *Precision*, dan *F1-measure*.

a. Confusion Matrix

Pada penanganan imbalance data dibuat *confusion matrix* untuk mempresentasikan hasil prediksi sebuah *classifier*. *Confusion matrix* menghasilkan akurasi, *precision* dan *recall*. Tabel klasifikasi adalah tabel yang terdiri dari data aktual dan data prediksi, tabel ini digunakan dengan tujuan untuk mengukur kinerja suatu model klasifikasi. Berikut ini merupakan tabel klasifikasi dengan dua kelas: [21]

Tabel 4. Confusion matrix untuk dua kelas

Count		Aktual	
		Churn	**Non Churn**
Prediksi	**Churn**	TP (True Positif)	FN (False Negatif)
	Non Churn	FP (False Positif)	TN (True Negatif)

Churn adalah kondisi pindahnya pelanggan dari satu provider ke provider lain. *Non churn* merupakan kondisi pelanggan tetap menggunakan layanan pada *provider* tersebut.

- *True Positive* (TP) adalah jumlah pelanggan dengan status *churn*, ketika diprediksi hasil prediksi menunjukan *churn*.
- *False Positive* (FP) adalah jumlah pelanggan dengan status *churn*, ketika diprediksi hasil prediksi menunjukan *non churn*.
- *False Negative* (FN) adalah jumlah pelanggan dengan status *non churn*, ketika diprediksi hasil prediksi menunjukan *churn*.
- *True Negative* (FN) adalah jumlah pelanggan dengan status *non churn*, ketika diprediksi hasil prediksi menunjukan *non churn*.

True positive rate (TPR) atau sensitivitas didefinisikan sebagai kelas contoh positif diprediksi dengan benar oleh model.

$$TPR = \frac{\text{TP}}{(\text{TP+FP})} \quad (2.4)$$

True negative rate (TNR) atau spesifisitas didefinisikan sebagai kelas contoh negatif diprediksi dengan benar oleh model.

$$TNR = \frac{\text{TN}}{(\text{TN+FN})} \quad (2.5)$$

False positive rate (FPR) adalah kelas contoh negatif diprediksi sebagai kelas positif.

$$FPR = \frac{\text{FP}}{(\text{FP+TN})} \quad (2.6)$$

False negative rate (FNR) adalah kelas contoh positif diprediksi sebagai kelas negatif.

$$FNR = \frac{\text{FN}}{(\text{FN+TP})} \quad (2.7)$$

Pengukuran performansi lainnya adalah Recall, Precision, dan F1-measure.

b. 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].

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (2.8)$$

Recall bisa diberi nilai dalam bentuk presentase 1 sampai 100% dengan hasil pembagian jika nilai 1 berarti relevan.[12]

c. Precision

Precision dihitung untuk mengevaluasi seberapa baik ketepatan model dapat memprediksi suatu kelas. *Precision* didapatkan dengan menghitung perbandingan antara jumlah data untuk satu kelas tertentu yang diprediksi dengan benar dibagi jumlah total prediksi kelas tersebut[10].

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (2.9)$$

Precision juga bisa diartikan sebagai kepersisan atau kecocokan dari permintaan informasi dengan jawaban terhadap permintaan tersebut. Perhitungan ini bisa diberi nilai dalam bentuk presentase 1 sampai 100% dengan hasil pembagian jika nilai 1 berarti relevan.[12]

d. F1-measure

F1-measure adalah perhitungan kombinasi antara *recall* dan *precision.* Merupakan ukuran pada akurasi pengujian.

$$\boldsymbol{F-Measure} = \frac{\boldsymbol{2*Precision*Recall}}{\boldsymbol{Precision+Recall}} \quad (2.10)$$

Skor *F1-measure* dapat diartikan sebagai rata-rata tertimbang (*weighted average*) dari presisi dan *recall*, di mana skor *F1-measure* mencapai nilai terbaik pada 1 dan terburuk pada 0.[12]

2.3.5 Skenario Pengujian

Skenario pengujian pada penelitian Tugas Akhir ini dilakukan menggunakan *dataset* PT. Telekomunikasi Indonesia Regional 7 yang telah dijelaskan pada bab sebelumnya. Pengujian dilakukan dengan membandingkan pengaruh pada hasil dari proses penanganan *imbalance data* dengan *undersampling* dan *bagging*, dan membandingkan pengaruh *logistic regression* terhadap performansi model klasifikasi.

1. Pengujian dilakukan dengan percobaan seleksi atribut yang berjumlah 7 atribut, 31 atribut, dan 53 atribut dengan menggunakan metode klasifikasi menggunakan Logistic Regression (LR) tanpa penanganan *imbalance data* dan dengan penanganan *imbalance data* menggunakan *Underbagging* dari komposisi data 70:30. Tujuan dari skenario ini adalah untuk mendapatkan jumlah atribut terbaik.
2. Pengujian dilakukan dengan percobaan penanganan *imbalance data* menggunakan *Underbagging* dengan banyak gugus data (p) sebesar 3, 7 dan 15. Tujuan dari skenario ini adalah untuk melihat pengaruh jumlah gugus data pada hasil pengujian yang dilakukan.

3. Pengujian dilakukan dengan percobaan penanganan *imbalance data* dengan *underbagging* dan pengaruh performansi klasifikasi Logistic Regession (LR). Tujuan dari skenario ini adalah untuk melihat pengaruh dari Logistic Regression (LR).

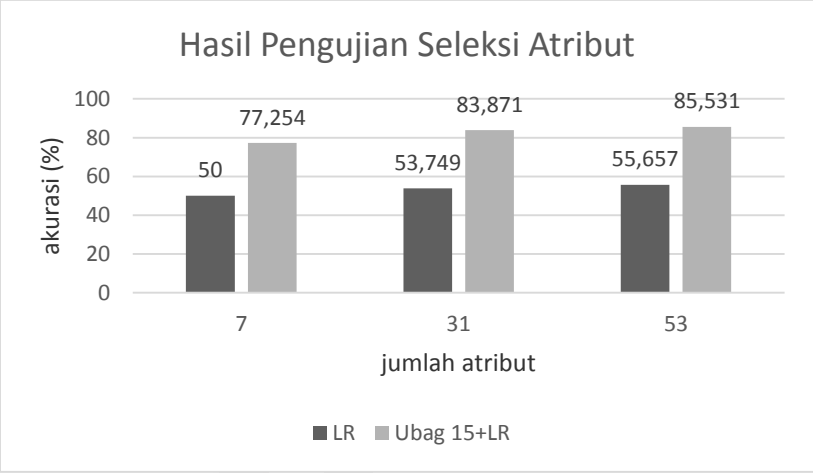
Pada masing masing skenario akan dianalisis performansi model klasifikasi yang dibangun kemudian dinyatakan dengan akurasi, dan nilai *f1-measure* pada prediksi *churn.*

3. Hasil dan Pembahasan

Dari penerapan skenario pengujian yang dijelaskan pada bab sebelumnya akan didapatkan hasil pengujian. Pada setiap skenario akan menggunakan parameter input pada penanganan imbalance data dengan jumlah gugus data (p) = 3, 7, dan 15. Berikut adalah hasil dan analisis pengujian dari skenario yang telah diterapkan.

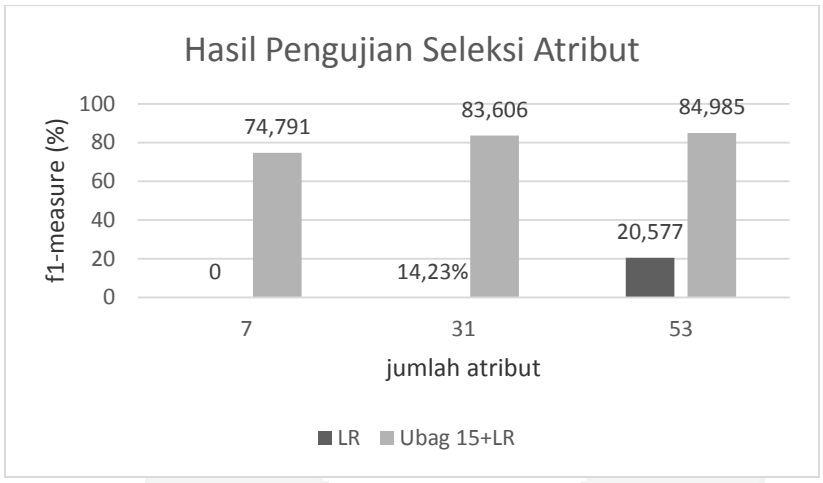
a. Pengaruh seleksi atribut terhadap hasil klasifikasi

Analisis dilakukan dari hasil pengujian model klasifikasi menggunakan Logistic Regression. Pengujian ini menggunakan hasil seleksi atribut yang berasal dari penggunaan korelasi pearson sehingga menjadi 7, 31 dan 53 atribut dengan komposisi data 70% data training dan 30% data testing. Selain itu, dilakukan analisis terhadap Logistic Regression dengan penanganan imbalance data pada 7, 31 dan 53 atribut. Pada penanganan imbalance data menggunakan metode Underbagging dengan jumlah gugus data p=15. Kemudian nilai akurasi yang didapat akan dinyatakan pada gambar 6 dan nilai f1-measure pada gambar 7.



Gambar 5. Hasil nilai akurasi pada pengujian pengaruh seleksi atribut

Pada grafik diatas memperlihatkan bahwa nilai akurasi yang diperoleh dari hasil pengujian seleksi atribut dengan hanya menggunakan metode klasifikasi *Logistic Regression* mengalami peningkatan dari seleksi atribut hingga saat menggunakan atribut asli yaitu berjumlah 53 atribut dengan nilai akurasi 55,657%. Hasil akurasi pada gambar 6 disebabkan oleh setiap atribut/variabel yang digunakan bisa mempengaruhi pada hasil performansi. Sedangkan, pada hasil akurasi dengan seleksi 31 atribut hanya terdapat selisih 1,9% dari atribut asli. Hal ini menunjukkan penggunaan seleksi 31 atribut memiliki keterkaitan antar antribut yang cukup kuat. Pada atribut dengan jumlah 7, hanya menghasilkan nilai akurasi yang rendah karena pemakaian atribut yang sedikit, karena ada variabel lain yang dibutuhkan untuk meningkatkan hasil akurasi.
Namun untuk kasus *imbalance class*, perhitungan akurasi tidak cocok karena *error* pada kelas *churn* tidak akan memberikan dampak yang berarti pada perhitungan akurasi. Sehingga perlu adanya penanganan *imbalance data.*



Gambar 6. Hasil nilai f1-measure Pengujian pengaruh Seleksi Atribut

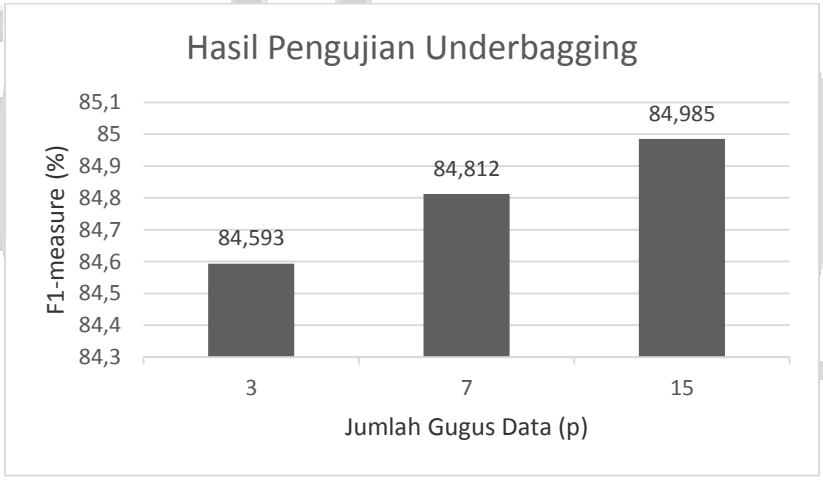
Gambar 7 merupakan hasil performansi pada pengujian pengaruh seleksi atribut diukur dari nilai *f1-measure*. Dapat dilihat dari seleksi 7 atribut nilai *f1-measure* = 0 dikarenakan true positive bernilai 0, ini berarti model dengan hanya penggunaan 7 atribut akan melakukan kesalahan pada prediksi data aktual churn yang seharusnya diklasifikasikan sebagai *churn*, karena adanya ketidakseimbangan kelas.

Karena nilai *f1-measure* yang sangat kecil maka perlu dilakukan penanganan kasus *imbalance data* agar hasil *f1-measure* yang didapatkan lebih baik lagi. Dari hasil grafik diatas, didapatkan nilai *f1-measure* tertinggi pada atribut asli tanpa penanganan *imbalance data* yaitu 20,577%. Setelah penanganan imbalance data pada penelitian ini menghasilkan nilai performansi f1-measure tertinggi sebesar 84,985%.

Melakukan seleksi atribut dapat disimpulkan memberi pengaruh terhadap hasil performansi model yang dibangun, berdasarkan pengujian bahwa hasil performansi tertinggi selalu dimiliki oleh atribut asli. Karena atribut asli sudah memiliki seluruh atribut yang dibutuhkan untuk menghasilkan performansi yang baik.

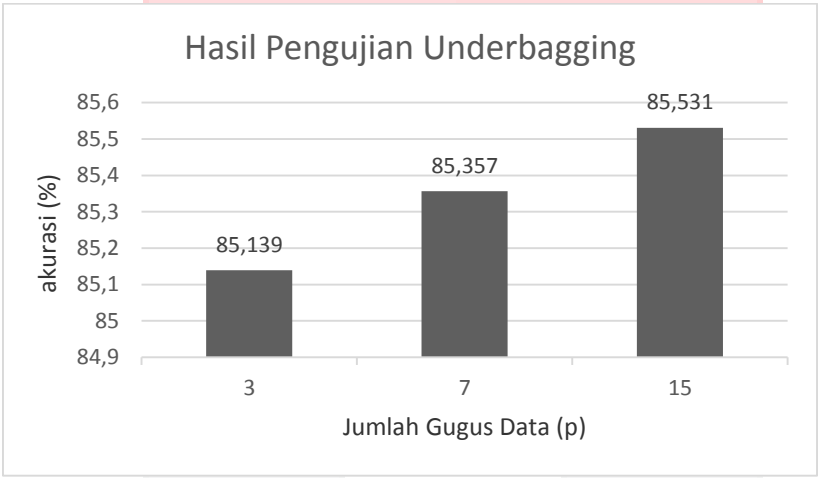
b. Pengaruh jumlah nilai gugus data pada metode Underbagging

Pada pengujian ini dilakukan penanganan imbalance data dengan menggunakan metode underbagging. Jumlah gugus data yang digunakan adalah 3, 7 dan 15 pada atribut asli 53. Kemudian dari hasil penanganan imbalance data akan dilakukan pengujian dengan menggunakan model klasifikasi *logistic regression* pada semua atribut yang dinyatakan dalam bentuk persen (%).



Gambar 7. Hasil nilai f1-measure pada pengujian pengaruh underbagging

Dari gambar 8 dapat dianalisis bahwa pada penanganan imbalance data menggunakan *underbagging* dengan p = 3, 7 dan 15 menghasilkan nilai *f1-measure* yang terus meningkat dan tertinggi pada p = 15 yaitu 84,985%. Hasil ini didapatkan karena semakin banyak jumlah gugus data p yang diterapkan, maka akan semakin meningkat pula nilai ketepatan prediksinya.



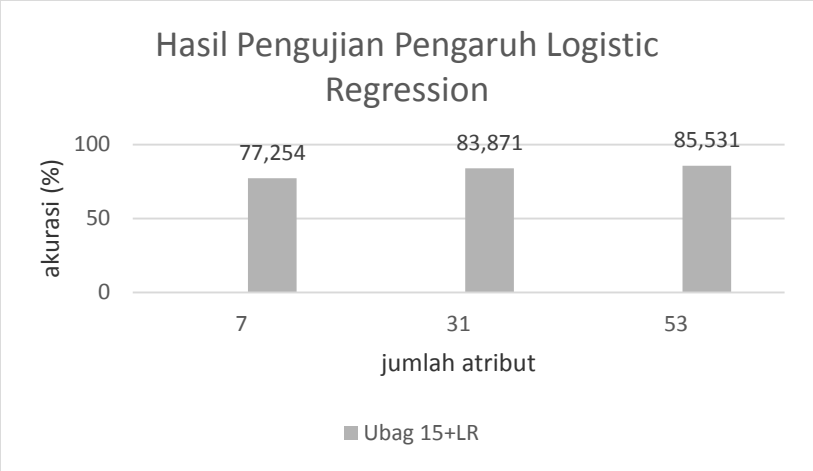
Gambar 8. Hasil nilai akurasi pada pengujian pengaruh underbagging

Gambar 9 menunjukkan hasil akurasi prediksi *churn* dari penanganan *imbalance data* menggunakan *underbagging* dengan p = 3, 7 dan 15, dengan nilai akurasi semakin meningkat dan tertinggi pada jumlah p = 15 yaitu 85,531%. Tingkat akurasi dapat meningkat dikarenakan penggunaan jumlah gugus data atau p pada penanganan imbalance data memberi pengaruh pada hasil akurasi model klasifikasi.

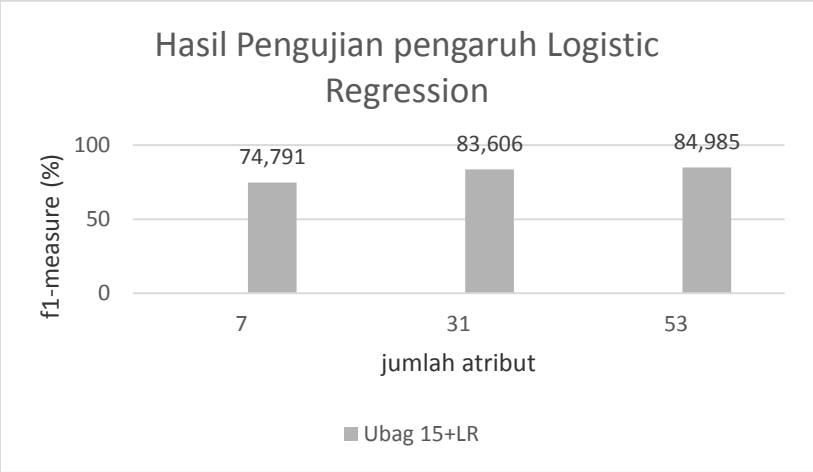
Karena data pada teknik *bagging* dilakukan *sampling* dengan atau tanpa pengembalian, sehingga distribusi data dari tiap data *bagging* berbeda. Beberapa data dari data training bisa saja muncul beberapa kali atau mungkin tidak muncul sama sekali. Hal inilah yang menjadi kunci bagaimana *bagging* bisa meningkatkan akurasi karena dengan *sampling* dengan atau tanpa pengembalian dapat memperkecil *variance* dari dataset sehingga *bagging* dapat mengurangi kesalahan dalam proses klasifikasi. *Variance* adalah perbedaan antara data aktual dengan prediksi yang diharapkan. Sehingga akurasi dapat naik.

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

Pada pengujian ini dilakukan analisis pengaruh klasifikasi *Logistic Regression* terhadap hasil performansi yang telah dilakukan penanganan imbalance data dengan menggunakan metode *underbagging* p=15. Setelah hasil didapatkan maka dilakukan analisis hasil performansi terhadap nilai akurasi dan nilai *f1-measure*. Logistic regression memprediksi nilai Y atau yang nanti disebut output prediksi, dari setiap nilai (*values*) estimasi koefisien regresi yang terdapat pada variabel X. Variabel X merupakan variabel prediktor yaitu 52 atribut.



Gambar 9. Hasil nilai akurasi pada pengujian pengaruh logistic regression



Gambar 10. Hasil nilai f1-measure pada pengujian pengaruh logistic regression

Berdasarkan gambar 11 yaitu dari hasil pengujian menggunakan 7, 31 dan 53 atribut didapatkan nilai *f1-measure* tertinggi pada atribut asli. Yaitu penggunaan seleksi 53 atribut. Hal ini dikarenakan model klasifikasi LR melakukan perhitungan menggunakan estimasi koefiesien regresi yang didapatkan dari masing-masing atribut berdasarkan persamaan 2.1. Koefisien regresi yang dihasilkan juga berjumlah 53. Dari estimasi koefiisen ini kemudian akan didapatkan nilai (z) yang didapatkan dari fungsi sigmoid $f(z)$. Setelah dilakukan penentuan nilai prediksi menggunakan hipotesis fungsi sigmoid yaitu 0.5 maka dihasilkan nilai Y atau output prediksi yang akan dibandingkan dengan data aktual.

Pada pengujian ini didapatkan variabel X yang memberikan pengaruh tinggi adalah atribut pertama sampai atribut ke 52.

Sehingga yang memberikan pengaruh pada hasil performansi klasifikasi untuk memprediksi *churn* menggunakan *logistic regression* adalah penggunaan estimasi koefisien regresi dari nilai atau isi dari 52 atribut ini.

4. Kesimpulan

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

1. Melakukan seleksi atribut dapat disimpulkan memberi pengaruh terhadap hasil performansi model yang dibangun karena hasil performansi tertinggi selalu dimiliki oleh atribut asli, walaupun atribut seleksi menghasilkan nilai yang terus meningkat. Sehingga jika menghasilkan performansi buruk terjadi karena ada variabel lain (atribut) yang dibutuhkan.
2. Melakukan penanganan imbalance data dengan *underbagging* bisa meningkatkan akurasi karena data pada teknik *bagging* dilakukan *sampling* dengan atau tanpa pengembalian, sehingga distribusi data dari tiap data *bagging* berbeda. *Sampling* ini dapat memperkecil *variance* dari dataset sehingga *bagging* dapat mengurangi kesalahan dalam proses klasifikasi.
3. Penambahan jumlah gugus data (p) pada metode *underbagging*, dapat meningkatkan nilai performansi yang dihasilkan oleh model klasifikasi yang dibangun yaitu *logistic regression.*
4. Hasil pengujian model klasifikasi *logistic regression* tanpa penanganan *imbalance data* mencapai nilai f1-measure 20,577% Hal ini menunjukkan bahwa harus dilakukan adanya penanganan *imbalance data*. Setelah dilakukan imbalance data maka f1-measure meningkat menjadi 85,531%.
5. Penggunaan estimasi koefisien nilai atau isi dari setiap atribut (variabel *predictor)* ini adalah yang memberikan pengaruh pada hasil performansi klasifikasi untuk memprediksi *churn* menggunakan *logistic regression*

5. Saran

Saran untuk pengembangan Tugas Akhir ini adalah parameter inputan gugus data (p) pada *Underbagging* dapat ditambah agar hasil performansi lebih baik

6. Referensi

[1] J. Lu, "Predicting Customer Churn in the Telecommunications Industry," *An Application of Survival Analysis Modeling Using SAS,* pp. 114-27, 2002.

[2] Y. Firdaus. I. Atastina. Prajunianto, "Churn Prediction pada Telekomunikasi Seluler dengan Metoda Logistic Regression.," *Departemen, Teknik Informatika, Institut Teknologi Telkom.*

[3] Fernandez. A. Barrenechea. E. Bustince. H. Herrera. F. Galar M, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems,* vol. 42, pp. 463-484, 2011.

[4] B. Alexander Yun-chung Liu, "The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets," *University of Texas. Austin,* 2004.

[5] E. B. Setiawan. I. Atania. Ferancis Leonardo, "Analisis Pengaruh Bagging pada Algoritma Klasifikasi Data Mining CART dan C.45," *Tugas Akhir,* 2012.

[6] P. Ning. M. Steinbach. d. V. Kumar. Tan, "Introduction To Datamining," *Boston. Pearson Addison Wesley.,* 2006.

[7] K. Coussement, "Employing SAS Text Miner Methodology to Become a Customer Genius in Customer Churn Prediction and Complaint E-mail Management," *Ghent University, Faculty of Economics and Business Administration, Department of Marketing, Belgium.,* 2008.

[8] G. Nie. Y. Chen. L. Zhang. Y. Guo., "Credit Card Customer Analysis based on Panel Data Clustering.," *Fictitious Economy and Data Science. Beijing, China.,* 2010.

[9] J. Sarwono, "Statistik itu mudah: panduan lengkap untuk belajar komputasi statistik menggunakan SPSS 16," *Yogyakarta: Andi.,* 2009.

[10] Emzir, "Metodologi Penelitian Pendidikan Kualitatif dan Kuantitatif," *Jakarta: PT Raja Grafindo Pergoda,* 2009.

[11] Berson, A., Smith, S., & Thearling, K., "Building data mining applications for CRM," *New York, NY: McGraw-Hill.,* 2000.

[12] E. Shaaban, Y. Helmy, A. Khedr, dan M. Nasr, "A Proposed Churn Prediction Model.," *The International Journal of Engineering Research and Applications (IJERA).,* vol. 2, no. 4, pp. 693-697, 2012.

[13] Abraham, Shaza M. Abd Elrahman. Ajith, "A Review of Class Imbalance Problem," *Journal of Network and Innovative Computing,* vol. 1, no. ISSN 2160-2174, pp. 332-340, 2013.

[14] J. W. d. Z.-H. Z. Xu-Ying Liu, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS,* vol. 39, APRIL 2009.

[15] D. L. W. G. X. &. J. H. Zhang, "A Novel Improved SMOTE Resampling Algorithm Based on Fractal," *Computational Information Systems,* pp. 2204-2211, 2011.

[16] E. G. M. &. G. N. Alfaro, "adabag: An R Package for Classification with Boosting and Bagging," *Journal of Statistical Software,* pp. 1-35, 2013.

[17] S. J. V. R. Barandela R, "New Appllications of Ensembles of Classifiers.," *Pattern Anal Applic 6,* pp. 245-256, 2003.

[18] S. a. Y. X. Wang, "Diversity analysis on imbalanced data sets by using ensemble models. In Computational Intelligence and Data Mining," *CIDM'09. IEEE Symposium,* pp. 324-331, 2009.

[19] J. K. M. Han, "Data Mining: Concepts and Techniques," *Morgan Kaufmann Publisher,* 2006.

[20] A. H. Karp, "USING LOGISTIC REGRESSION TO PREDICT CUSTOMER RETENTION," *Sierra Information Services, Inc.*.

[21] J. N. K. A. Chawla NV, "Editorial: Special Issue on Learning from Imbalance Data Sets.," *ACM SIGKDD Explorations.,* vol. 6, pp. 1-6, 2004.

[22] Mohammed J. Zaki, Wagner Meira JR., "Data Mining and Analysis," dalam *Data Mining and Analysis Fundamental Concepts and Algorithms*, New York, Cambridge University Press, 2014, p. 1.

[23] A. L. R. Ginanjar, "Penerapan Data Mining untuk Memprediksi Kriteria Nasabah Kredit," *Jurnal Komputer dan Informatika. Universitas Komputer Indonesia,* 2012.

[24] Z. A. M. A. B. Angelina S., "Analisis Pengaruh Metode Combine Sampling untuk Churn Prediction," *Institut Teknologi Telkom. Bandung,* 2010.

[25] F. P. Manurung. Adiwijaya. A. Aditsania, "Handling Imbalance Data pada Prediksi Churn menggunakan Underbagging dan Logistic Regression.," *Tugas Akhir. Telkom University. Bandung,* 2017.

[26] Y. Permatasari, "Penanganan Masalah Kelas Tidak Seimbang dengan RUSboost dan Underbagging (Studi kasus: Mahasiswa Drop Out).," *Sekolah Pascasarjana IPB. Bogor,* 2016.

[27] R. S. W. Aries Saifudin, "Penerapan Teknik Ensemble untuk Menangani Ketidakseimbangan," *Journal of Software Engineering,* vol. 1, April 2015.

[28] Dwiyanti, E., Adiwijaya, and Ardiyanti, A, "Handling Imbalanced Data in Churn Prediction Using RUSBoost and Feature Selection (Case Study: PT. Telekomunikasi Indonesia Regional 7).," *In International Conference on Soft Computing and Data Mining,* pp. 376-385. Springer, Cham., 2016.

[29] Effendy, V., Adiwijaya, and Baizal, Z.A., 2014. Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest. In Information and Communication Technology (ICoICT), 2014 2nd International Conference on (pp. 325-330). IEEE.

[30] Adiwijaya, Wisesty, U.N. and Nhita, F., 2014. Study of Line Search Techniques on the Modified Backpropagation for Forecasting of Weather Data in Indonesia. Far East Journal of Mathematical Sciences, 86(2), p.139.

[31] Adiwijaya, 2014, Aplikasi Matriks dan Ruang Vektor, Yogyakarta: Graha Ilmu

[32] Adiwijaya, 2016, Matematika Diskrit dan Aplikasinya, Bandung: Alfabeta