

1. Pendahuluan

1.1 Latar belakang

Churn prediction adalah salah satu aplikasi dari task *data mining* yaitu klasifikasi yang bertujuan untuk memprediksi para pelanggan yang berpotensi untuk menghentikan jasa layanannya selama ini. Didalam konsep *churn*, “lebih murah *cost* bagi perusahaan untuk mempertahankan *customer* mereka daripada mengupayakan menarik *customer* yang baru [8]”. Didalam kasus *churn prediction*, ada satu masalah bagi algoritma klasifikasi biasa, yaitu *imbalance*. *Imbalance* adalah kondisi ketika porsi data untuk keperluan *learning* di sebuah kelas lebih banyak daripada porsi data di kelas lain. “Algoritma klasifikasi biasa cenderung mem-*bias* prediksi dari *record* yang aktualnya adalah kelas minor di prediksi ke dalam kelas mayor [2]”. Dengan kondisi seperti ini, maka akurasi dari kelas mayor cenderung baik sedangkan akurasi dari kelas minor cenderung buruk.

Implementasi dari *churn prediction* pada studi kasus di sebuah perusahaan telekomunikasi yang penulis ambil, menggunakan tool data mining Clementine 9.0 di dalam pembangunan modelnya. Permasalahan data *imbalance* di atasi dengan *node balancing*, konsep dari *node balancing* Clementine 9.0 adalah melakukan *oversampling*, *undersampling* atau gabungan dari *oversampling* dan *undersampling*. Dengan cara seperti itu, maka akurasi data *training* pada kelas minor cenderung membaik dengan peningkatan akurasi yang signifikan sesuai dengan tingkat *oversampling* / *undersampling* yang dilakukannya. Tetapi permasalahannya adalah model yang dihasilkan terlalu *overfitting*, terlalu spesifik untuk data yang di *training*, sehingga menghasilkan akurasi yang cenderung menurun untuk data *testing*.

Databoost-IM adalah sebuah pengembangan dari databoost, yaitu sebuah metoda klasifikasi data mining untuk kondisi data yang *imbalance*. Databoost merupakan perpaduan dari *boosting* dengan *data generation*. *Boosting* adalah sebuah metoda *ensemble* dimana preformansi dari algoritma klasifikasi yang lemah diperbaiki dengan memfokuskan pada *hard examples* yang sulit untuk diklasifikasikan. Metoda *boosting* murni cenderung akan menurunkan akurasi dari kelas mayor karena mesin *learning* dipaksa fokus hanya pada *hard examples* saja. DataboostIM memadukan *boosting* dengan *data generation*, *data generation* adalah membuat data sintesis untuk ditambahkan ke data asli. “Tujuan utama dari *data generation* adalah memperbaiki *knowledge* yang dimiliki oleh *current classifier* [3]”. Oleh karena itu tujuan utama dari DataboostIM adalah memperbaiki akurasi prediksi terhadap kelas mayor dan kelas minor, tanpa mengorbankan salah satu dari dua kelas yang ada. Diharapkan dengan metoda DataboostIM ini dapat menjadi alternatif pembangunan model klasifikasi *churn prediction* pada study kasus yang penulis ambil.

1.2 Perumusan masalah

Berdasarkan latar belakang permasalahan di atas, dapat di rumuskan permasalahan yaitu menemukan alternatif solusi lain untuk kasus *churn prediction*

telco. Akan di implementasikan metoda DataboostIM sebagai alternatif solusi lain. Hipotesis yang akan di uji adalah sejauh mana DataboostIM mampu memperbaiki model yang dihasilkan oleh *base classifier* dalam bentuk *f-measure*, *g-mean*, *recall*, *precision* serta *lift curve* jika dibandingkan dengan metoda *balancing clementine*. Dugaan yang akan dibuktikan adalah "DataboostIM memberikan hasil yang lebih baik dikarenakan tambahan data sintetik yang diharapkan mampu menambah knowledge dari *current classifier* [3]".

Lingkup TA ini hanya mencakup kasus *churn prediction* pada sebuah perusahaan telekomunikasi. Asumsi yang dipakai adalah data yang masuk ke aplikasi sudah mengalami *preprocessing*.

1.3 Tujuan

Berdasarkan dari rumusan masalah yang telah didefinisikan di atas, maka tujuan Tugas Akhir ini yaitu :

1. Mengimplementasikan metoda DataboostIM dalam bentuk perangkat lunak yang terintegrasi dengan model-model klasifikasi di Clementine 9.0 sebagai *base classifier*-nya (meliputi Neural Network, C50, CHAID).
2. Menganalisis performansi dari metoda DataboostIM dibandingkan dengan metoda *balancing* dari Clementine dalam bentuk *f-measure*, *g-mean*, *recall*, *precision* (biasa dipakai untuk mengukur performansi pada kasus *imbalance class*) serta *lift curve* (biasa di pakai untuk mengukur performansi pada kasus *churn prediction*).

1.4 Metodologi penyelesaian masalah

Metoda penyelesaian masalah yang digunakan dapat di jelaskan sebagai berikut :

1. Studi iteratur

Pencarian referensi, mencari referensi dan sumber-sumber lain yang layak yang berhubungan dengan *data mining*, CEMI Clementine 9.0, *churn prediction*, *imbalance class* (khususnya metoda DataboostIM).

2. Pencarian dan pemahaman data

Mencari data pelanggan pada sebuah perusahaan telekomunikasi di Indonesia yang sudah dalam bentuk *artifact*, yaitu data yang sudah diproses dari data pelanggan yang berupa data demografis dan data *traffic* (data Call Detail Record) serta *payment* (data pembayaran), kedalam bentuk satu *record* satu pelanggan dengan atribut-atribut yang memanjang ke kanan. Setelah data didapatkan, maka langkah selanjutnya adalah *data understanding* yaitu memahami data yang didapat.

3. Analisis dan Perancangan Perangkat Lunak

Menganalisa permasalahan yang akan diselesaikan, menganalisa tahapan-tahapan yang digunakan untuk menyelesaikan permasalahan dengan metoda *Object Oriented*.

4. Implementasi Sistem

Melakukan implementasi sistem dengan membangun perangkat lunak sesuai dengan perancangan yang telah dilakukan dengan memakai php dan *tool data mining* Clementine 9.0.

5. Pengujian Sistem dan Analisa Hasil

Pengujian dilakukan terhadap metoda DataboostIM terhadap parameter-parameter yang telah di definisikan di awal yaitu *f-measure*, *gmean*, *recall*, *precision*, dan *lift curve* dibandingkan dengan hasil yang didapat pada metoda *balancing* Clementine 9.0. Analisa dilakukan terhadap hasil yang diperoleh dari pengujian tersebut. Disamping DataboostIM dan *balancing* Clementine 9.0, akan di uji juga terhadap algoritma-algoritma *imbalance* yang lain dengan menggunakan *tool* weka 3.5.3.

6. Pengambilan Kesimpulan dan Penyusunan Tugas Akhir

1.5 Sistematika Penulisan

Sistematika penulisan di dalam tugas akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Berisi latar belakang pembuatan tugas akhir, perumusan masalah, tujuan, metodologi penyelesaian masalah dan sistematika penulisan

BAB II LANDASAN TEORI

Pada bab ini dikemukakan berbagai teori yang mendukung penyusunan tugas akhir, antara lain meliputi teori tentang *data mining*, *imbalance class*, *churn prediction*, *ensemble methods*, *databoost* dan *databoostIM*.

BAB III ANALISIS KEBUTUHAN DAN PERANCANGAN SISTEM

Menjelaskan tentang proses analisa dan perancangan sistem yang akan di bangun..

BAB IV ANALISA HASIL PENGUJIAN

Menjelaskan tentang proses analisa terhadap beberapa skenario pengujian terhadap sistem. Pengujian juga melibatkan perbandingan dengan algoritma *boosting* yang lain dengan menggunakan tool weka 3.5.3.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan akhir dan saran pengembangan dari tugas akhir.