PENGGUNAAN *RANDOM OVER SAMPLING EXAMPLE* (ROSE) UNTUK PENANGANAN KETIDAKSEIMBANGAN KELAS PADA KASUS *CREDIT RISK SCORING* BERBASIS *DECISION TREE* DAN REGRESI LOGISTIK DENGAN *10-FOLD REPEATED CROSS VALIDATION*

Aynayatul Khoiriyah (G14160006), Audhi Aprilliant (G14160021), Rhesa Amalia Cempaka (G14160022), Nadya Paramita (G14160029)
Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Institut Pertanian Bogor, Jl. Meranti Wing 22 Lantai 4,
Dramaga, Bogor, Jawa Barat 16680

ABSTRAK

Credit risk scoring analysis merupakan analisis data *imbalanced* yang biasanya dilakukan oleh pihak perbankan untuk menentukan status dari seorang kreditur dalam mempertanggung jawabkan pinjamannya. Hal ini dilakukan untuk meminimalisir kerugian yang akan ditanggung oleh instansi keuangan seperti bank. Secara umum, ketidakseimbangan kelas dapat ditangani dengan dua pendekatan, yaitu level data dan level algoritma. Pada kasus ini, pendekatan level data yang digunakan adalah *over sampling*, *under sampling*, *Random Over Sampling Example* (ROSE), dan SMOTE. Sedangkan pendekatan level algoritma menggunakan metode klasifikasi dengan algoritma *decision tree* dan metode regresi logistik. Penelitian ini bertujuan melakukan prediksi terhadap *credit condition* seorang *customer* menggunakan metode yang sesuai. Hasil penelitian menunjukkan bahwa model *decision tree* dengan metode sampling ROSE merupakan model pendekatan terbaik dalam memprediksi *credit condition* seorang *customer* dengan nilai AUC yang diperoleh sebesar 0.8611 dan *specificity* sebesar 0.8338.

Kata Kunci: *credit risk scoring analysis*, data *imbalanced*, *decision tree*, ROSE

PENDAHULUAN

Latar Belakang

Menurut Ganganwar V (2012), dalam mengajukan aplikasi permohonan kredit kepada instansi keuangan, pemohon harus melalui sebuah tahap analisis yang dilakukan oleh pihak perbankan yang dikenal sebagai *credit scoring analysis*. Tujuan utama dari analisis yang dilakukan adalah untuk mengelompokkan pihak pemohon sesuai dengan kriteria yang telah ditentukan sebelumnya (umumnya *"Bad Creditor"* dan *"Good Creditor"*). Selain itu, tujuan dari analisis yang dilakukan adalah untuk meminimalisir kerugian yang akan ditanggung oleh instansi keuangan apabila melakukan kesalahan dalam mengelompokkan pihak pemohon.

Sebuah data disebut *imbalanced* apabila memiliki kelas yang banyak pada salah satu kelas target dibandingkan dengan kelas lainnya. Salah satu permasalahan *data imbalanced* sering ditemui pada kasus *credit risk scoring*. Banyak penelitian tentang *data imbalanced* sepakat bahwa dikarenakan sebaran target yang tidak merata menyebabkan kinerja algoritma klasifikasi akan cenderung bias terhadap kelas mayoritas. Ada dua pendekatan yang dapat dilakukan untuk mengatasi kondisi *data imbalanced*. Pendekatan pertama yaitu pada level data, menggunakan teknik pengambilan contoh (*sampling technique*) dan pendekatan kedua yaitu pada level algoritma. Pendekatan *sampling technique* tersebut terdiri dari dua cara yaitu *over-sampling* kelas terkecil, *under-sampling* kelas terbesar, dan gabungan dari kedua teknik sampling. Sedangkan dari level algoritma dengan menerapkan algoritma yang sering

digunakan untuk kasus *imbalanced data*, seperti *Decision Tree*, *Support Vector Machine* (SVM), *K-Nearest Neighbor (KNN), Neural Network*, dan regresi logistik.

Oleh karena itu dibutuhkan sebuah model *credit scoring analysis* yang dapat digunakan pada peubah-peubah dependen dan independen yang memiliki hubungan yaang kompleks (non-linear), tidak terdistribusi secara Normal, mudah digunakan (tidak memakan waktu yang lama), serta memberikan hasil yang mudah diinterpretasikan. Dalam penelitian ini digunakan metode *Decision Tree* dan regresi logistik sebagai model *credit scoring analysis* yang baru dengan teknik sampling *over sampling* dan *under sampling*.

Tujuan

Tujuan dari penelitian mengenai Penggunaan *Random Over Sampling Example* (ROSE) untuk Penanganan Ketidakseimbangan Kelas pada Kasus *Credit Risk Scoring* berbasis *Decision Tree* dan Regresi Logistik adalah sebagai berikut: (1) Melakukan eksplorasi terhadap peubah-peubah dalam terkait *credit condition* suatu pelanggan; (2) Mengidentifikasi faktor-faktor yang berpengaruh terhadap penentuan *credit condition* suatu pelanggan; (3) Menangani data dengan *target imbalanced* menggunaakan metode ROSE; (4) Menentukan metode klasifikasi terhadap data *Banking Credit Risk* yang memiliki metrik evaluasi prediksi yang paling tinggi; (5) Melakukan prediksi terhadap *credit condition* suatu *customer* menggunakan metode yang sesuai; dan (6) Memberikan rekomendasi kepada lembaga bank dalam melakukan *credit scoring analysis* menggunakan metode klasifikasi yang sesuai.

Manfaat

Manfaat yang dapat diperoleh dari penelitian mengenai Penggunaan *Random Over Sampling Example* (ROSE) untuk Penanganan Ketidakseimbangan Kelas pada Kasus *Credit Risk Scoring* berbasis *Decision Tree* dan Regresi Logistik adalah sebagai berikut: (1) Mengetahui faktor-faktor yang berpengaruh terhadap penentuan *credit condition* suatu *customer* dan (2) Mengetahui status prediksi konfisi kredit seorang nasabah baru.

Ruang Lingkup

Ruang lingkup pada penelitian mengenai Penggunaan *Random Over Sampling Example* (ROSE) untuk Penanganan Ketidakseimbangan Kelas pada Kasus *Credit Risk Scoring* berbasis *Decision Tree* dan Regresi Logistik adalah sebagai berikut: (1) Data merupakan data sekunder yang diunduh dari laman *Kaggle* dengan nama data *Credit Bank Scoring*.

TINJAUAN PUSTAKA

Credit Scoring Analysis

Secara umum *credit scoring analysis* merupakan sebuah sistem analisis yang digunakan oleh instansi keuangan dalam menentukan status kredit dari calon kreditur berdasarkan pada data kredit historis dari kreditur-kreditur saat ini. Dalam *credit scoring analysis* yang dilakukan, calon kreditur umumnya akan dikelompokkan menjadi dua kelompok yaitu kelompok "*good creditor*", kelompok kreditur yang mampu membayar tagihan kredit dengan baik, dan kelompok "*bad creditor*", kelompok kreditur yang tidak mampu membayar tagihan kredit dengan baik (Yap, Ong & Husein 2011).

Klasifikasi

Klasifikasi merupakan proses menemukan model (atau fungsi) yang menggambarkan dan membedakan konsep atau kelas-kelas data, dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu objek atau data yang label kelasnya tidak diketahui (Han & Kamber 2001).

Klasifikasi terdiri dari dua tahap yaitu pelatihan dan prediksi (klasifikasi). Pada tahap pelatihan, dibentuk sebuah model domain permasalahan dari setiap *instance* yang ada. Penentuan model tersebut berdasarkan analisis pada sekumpulan data pelatihan, yaitu data yang label kelasnya sudah diketahui. Pada tahap klasifikasi, dilakukan prediksi kelas dari instance (kasus) baru dengan menggunakan model yang telah dibuat pada tahap pelatihan (Guvenir et al. 1998).

***Decision Tree* (Pohon Keputusan)**

Pohon keputusan merupakan pohon yang dalam analisis pemecahan masalah pengambilan keputusan adalah pemetaan mengenai alternatif-alternatif pemecahan masalah yang dapat diambil dari masalah tersebut. Pohon tersebut juga memperlihatkan faktor-faktor kemungkinan/ probabilitas yang akan mempengaruhi alternatif-alternatif keputusan tersebut, disertai dengan estimasi hasil akhir yang akan didapat bila kita mengambil alternatif keputusan tersebut. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki.

Regresi Logistik

Regresi Logistik adalah suatu metode analisis statistika untuk mendeskripsikan hubungan antara peubah respon yang memiliki dua kategori atau lebih dengan satu atau lebih peubah penjelas berskala kategori atau kontinu. Model regresi logistik biner digunakan untuk menganalisis hubungan antara satu peubah respon dan beberapa peubah penjelas, dengan peubah responnya berupa data kualitatif dikotomi yaitu bernilai 1 untuk menyatakan keberadaan sebuah karakteristik dan bernilai 0 untuk menyatakan ketidakberadaan sebuah karakteristik. Berikut persamaannya regresi logistik :

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1X$$

Dimana:

ln : Logaritma Natural.

$B_0 + B_1X$: Persamaan yang biasa dikenal dalam OLS.

$\hat{p}$: probabilitas logistik yang didapat rumus probabilitas regresi logistik sebagai berikut:

$$\hat{p} = \frac{e^{B_0+B_1X}}{1 + e^{B_0+B_1X}}$$

Dimana "e" adalah fungsi exponen. (Perlu diingat bahwa exponen merupakan kebalikan dari logaritma natural. Sedangkan logaritma natural adalah bentuk logaritma namun dengan nilai konstanta 2.71828182845904 atau biasa dibulatkan menjadi 2.72).

Imbalanced Data

Sebuah himpunan data dikatakan menjadi tidak seimbang (*imbalanced*) jika terdapat satu kelas yang direpresentasikan dalam jumlah *instance* yang kecil bila dibandingkan dengan kelas lainnya. Kondisi tersebut dapat menimbulkan masalah pada klasifikasi data yang kasusnya jarang terjadi akan tetapi sangat penting, contohnya pada pengklasifikasian

data kecurangan transaksi telepon, pengenalan citra satelit untuk pendeteksian tumpahan minyak, deteksi kegagalan mesin suatu pabrik, deteksi penyakit langka tetapi berbahaya (Barandela et al. 2002). Kondisi *data imbalanced* dapat terlihat secara nyata pada himpunan data yang memiliki dua kelas. Kelas yang jumlah *instance* terkecil (*minority class*) disebut kelas positif dan kelas yang jumlah *instance* terbesar (*majority class*) disebut kelas negatif. Rasio jumlah *instance* antara dua kelas tersebut sebesar 1:100, 1:1.000, dan 1:10.000 atau lebih.

Ada dua pendekatan yang dapat dilakukan untuk mengatasi kondisi *data imbalanced*. Pendekatan pertama yaitu pada level data, menggunakan teknik pengambilan contoh (*sampling technique*) dan pendekatan kedua yaitu pada level algoritma. Pendekatan *sampling technique* tersebut terdiri dari dua cara yaitu *over-sampling* kelas terkecil dan *under-sampling* kelas terbesar.

Over Sampling

Over sampling merupakan teknik pengambilan contoh meningkatkan jumlah kelas terkecil dengan cara mereplikasi data secara acak sehingga jumlahnya sama dengan kelas terbanyak. Penggunaan metode *over-sampling* yang berlebihan dapat menyebabkan *overfitting*.

Under Sampling

Under sampling merupakan teknik pengambilan contoh mengurangi jumlah data terbesar secara acak sehingga jumlahnya sama dengan kelas terkecil. Penggunaan metode *under-sampling* yang berlebihan dapat menyebabkan hilangnya beberapa informasi penting yang terdapat pada dataset.

SMOTE (*Synthetic Minority Over-Sampling Technique*)

Tujuan utama dari teknik SMOTE adalah untuk menangani *imbalance class*. Penggunaan teknik SMOTE (*Synthetic Minority Over-Sampling Technique*) (Chawla et al. 2002) menghasilkan hasil yang baik dan efektif untuk menangani *imbalance class* yang mengalami *overfitting* pada proses teknik *over-sampling* untuk kelas minoritas (positif) (Riquelme et al. 2008).

SMOTE menciptakan sebuah contoh dari kelas minoritas sintetis yang beroperasi di ruang fitur daripada ruang data. Dengan menduplikasi contoh kelas minoritas, teknik SMOTE menghasilkan contoh sintetis baru dengan melakukan ekstrapolasi sampel minoritas yang ada dengan sampel acak yang diperoleh dari nilai k tetangga terdekat. Dengan hasil sintetis pada contoh yang lebih dari kelompok minoritas, sehingga mampu memperluas area keputusan mereka untuk minoritas (Chawla et al. 2002). Seleksi atribut (pilihan fitur) adalah bagian dari atribut dari proses seleksi yang relevan untuk membangun model pembelajaran yang baik (Guyon 2003).

ROSE (*Random Over Sampling Example*)

ROSE meningkatkan ukuran kelas minoritas dengan mensintesis sampel baru atau langsung mereplikasi secara acak dataset training (Yu et al., 2013). ROSE meningkatan jumlah kelas mayoritas sehingga meningkatkan waktu prediksi.

AUC (*Area Under Curve*)

Area Under the ROC (*Receiver Operating Characteristic*) *Curve* (AUROC atau AUC) adalah ukuran numerik untuk membedakan kinerja model, dan menunjukkan seberapa

sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif (Attenberg & Ertekin 2013). *Area Under Curve* (AUC) digunakan untuk mengukur hasil kinerja model prediksi dapat dilihat. Hasilnya dapat dilihat dari hasil *confusion matrix* secara manual dengan perbandingan klasifikasi menggunakan kurva *Receiver Operating Characteristic* (ROC).

$$TP_{rate} = \frac{TP}{TP+FN} \; ; \qquad FP_{rate} = \frac{FP}{FP+TN} \; ; \qquad AUC = \frac{1+TP\ rate - FP\ rate}{2}$$

METODE

Data Penelitian

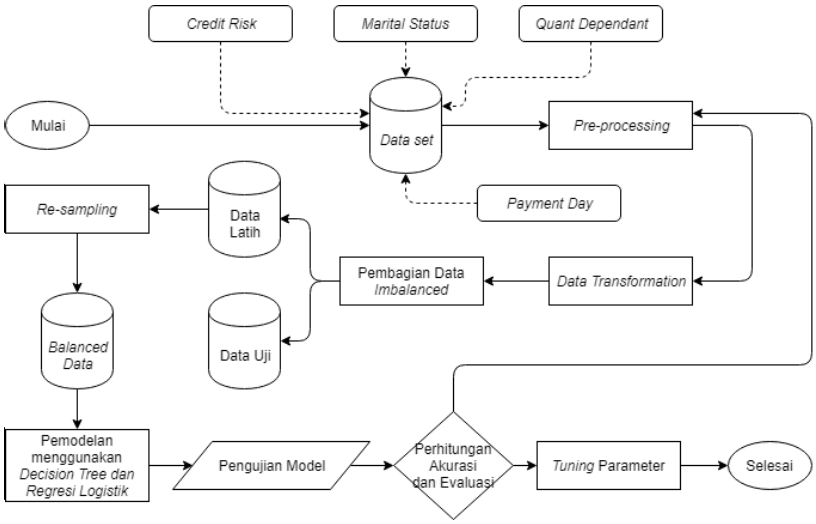
Data yang digunakan adalah data sekunder yang berasal dari situs web *kaggle*. Data terbagi menjadi 2, yaitu data latih dan data uji dengan masing-masing data berformat *csv* (*comma delimited*). Data ini berisi beberapa parameter yang digunakan dalam menentukan skor dari suatu kartu kredit. Data ini terdiri dari 31 parameter/ atribut dan 40000 amatan. Beberapa peubah yang terdapat di dalam data sebelum dilakukan pra-proses data ditampilkan di dalam tabel 1 sebagai berikut:

Tabel 1. Penjelasan Beberapa Peubah Sebelum dilakukan Tahap Pra-proses

No.	Nama Peubah	Tipe	Penjelasan
1	ID CLIENT	Faktor	Sebagai identitas dari pemilik kartu kredit
2	ID SHOP	Faktor	Kode pembuatan kartu kredit
3	SEX	Faktor	Jenis kelamin (“male”, ”female”)
4	MARITAL STATUS	Faktor	Status (“single”, “married”, “divorced”, “widow”, “other”)
5	AGE	Numerik	Usia pemilik kartu kredit
6	QUANT DEPENDANTS	Numerik	Jumlah tanggungan dalam keluarga
7	EDUCATION	Faktor	Tingkat pendidikan
8	FLAG RESIDENTIAL PHONE	Faktor	Apakah memiliki telpon rumah (“yes”, “no”)
9	TARGET LABEL	Faktor	Label atau kelas: Bad = 1, Good = 0

Tahapan Kegiatan

Tahapan yang dilakukan pada penelitian terdiri atas *data selection*, data *pre-processing*, *data transformation*, klasifikasi menggunakan algoritma *Decision Tree* dan regresi logistik, interpreasi hasil dan evaluasi. Secara sistematis, alur dari tahapan-tahapan dapat dilihat pada gambar 1.



Gambar 1 *Flow chart* tahapan kegiatan

Lingkungan Pengembangan

Perangkat keras yang digunakan dalam penelitian ini adalah komputer personal dengan spesifikasi sebagai berikut:

a. Prosesor : AMD A8 – 7410
b. Memory : 4GB
c. VGA : Radeon (TM) R5 Graphics

Perangkat lunak yang digunakan dalam penelitian ini adalah sebagai berikut:

a. Sistem operasi Microsoft Windows Pro 10 (64-bit) dan Ubuntu Bionic Breaver
b. Bahasa pemograman R dengan packages *caret, ggplot, ROSE, DMwR, SMOTE*
c. Microsoft Excel 2016 sebagai media pengolah data tambahan, media penggabungan data, dan transformasi data

HASIL DAN PEMBAHASAN

Pra-proses dan Transormasi Data

Tahapan awal dalam penelitian mengenai klasifikasi *Banking Credit Risk* adalah pra-proses data. Pra-proses yang dilakukan sebagai berikut:

1. Mereduksi atribut yang dianggap tidak relevan dengan *credit scoring.*
   Atribut yang direduksi adalah ID CLIENT, ID SHOP, dan AREA RESIDENCIAL PHONE.
2. Mereduksi atribut yang bernilai *null* untuk seluruh amatan.
   Atribut yang direduksi adalah EDUCATION, COD APPLICATION BOTH, QUANT BANKING ACCOUNT, FLAG CARD INSURANCE OPTION, QUANT DEPENDANT, FLAG OTHER CARD, FLAG MOBILE PHONE, dan FLAG CONTACT PHONE.
3. Mereduksi beberapa observasi yang atribut tertentunya tidak bernilai.
   a. Sebanyak 3 observasi direduksi karena atribut SEX nya tidak diisi
   b. Sebanyak 4 observasi direduksi untuk atribut MARITAL STATUS nya bernilai S, tetapi MATE INCOME nya bernilai > 0, yang kemungkinan disebabkan oleh *human error*.
   c. Atribut PERSONAL NET INCOME yang bernilai lebih besar dari 80.000 sebanyak 15 amatan.
4. Mereduksi atribut PROFESSION CODE karena atribut tersebut berisi 291 kategori pekerjaan, tanpa adanya keterangan.
5. Recode atribut nominal dan kategorik menjadi peubah numerik.
   a. Atribut PERSONAL REFERENCE1 dan PERSONAL REFERENCE2 bernilai 1 untuk yang mengisi dan bernilai 0 untuk yang tidak mengisi.
   b. Atribut PAYMENT DAY bernilai 1 untuk yang mengisi di bawah 20 dan bernilai 0 untuk yang mengisi di atas tanggal 20.
6. Atribut MONTH IN THE JOB dan MONTH IN RESIDENCE diubah menjadi data tahunan serta berganti menjadi atribut YEAR IN THE JOB dan YEAR IN RESIDENCE.
7. Penanganan pada beberapa data yang dianggap tidak wajar.
   a. Atribut YEAR IN RESIDENCE yang bernilai lebih besar dari atribut AGE, disesuaikan dengan nilai AGE.
   b. Atribut YEAR IN THE JOB yang benilai lebih besar daripada AGE, disesuaikan dengan rata-rata usia produktif/ kerja di Brazil yaitu 15 tahun.
   c. Atribut MARITAL STATUS bernilai *single*, tetapi MATE INCOME > 0, maka dihapus

8. Membuat peubah *dummy* untuk data kategorik.
    a. Atribut baru untuk MARITAL STATUS, yaitu MARITAL STATUS S, MARITAL STATUS C, MARITAL STATUS O, MARITAL STATUS V
    b. Atribut baru untuk RESIDENCE TYPE yaitu RESIDENCE TYPE P, RESIDENCE TYPE O, RESIDENCE TYPE A.

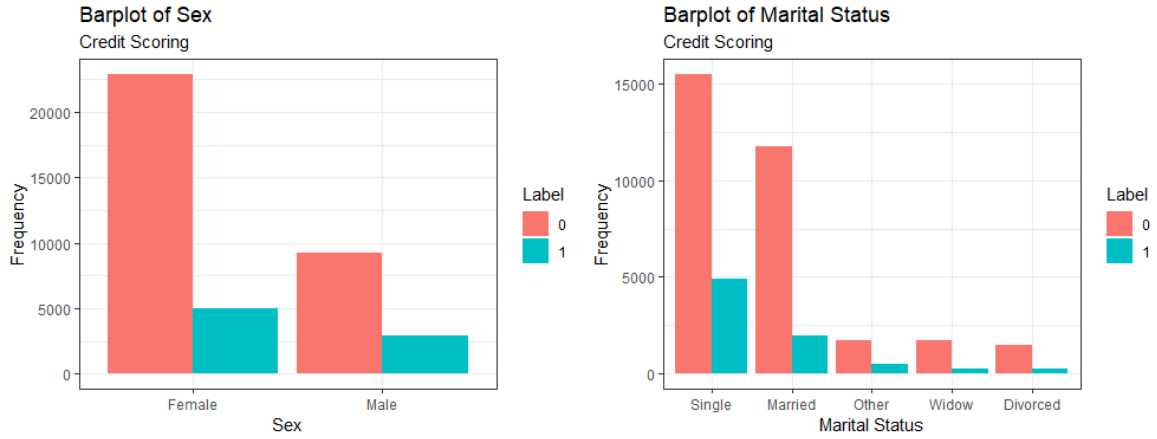
Tabel 2. Penjelasan Peubah Setelah dilakukan Tahap Pra-proses

No.	Nama Peubah	Tipe	Penjelasan
1	SEX	Faktor	Jenis kelamin (bernilai 1 apabila berjenis kelamin M dan bernilai 0 apabila berjenis kelamin F)
2	MARITAL STATUS S	Faktor	Status (bernilai 1 apabila berstatus Single dan 0 apabila berstatus lainnya)
3	MARITAL STATUS C	Faktor	Status (bernilai 1 apabila berstatus Married dan 0 apabila berstatus lainnya)
4	MARITAL STATUS O	Faktor	Status (bernilai 1 apabila berstatus Other dan 0 apabila berstatus lainnya)
5	MARITAL STATUS V	Faktor	Status (bernilai 1 apabila berstatus Divorced dan 0 apabila berstatus lainnya)
6	AGE	Numerik	usia pemilik kartu kredit
7	FLAG RESIDENCIAL PHONE	Faktor	Telepon rumah (bernilai 1 apabila memiliki telepon rumah dan bernilai 0 apabila tidak memiliki telepon rumah)
8	PAYMENT DAY	Faktor	Pembayaran bulanan (bernilai 1 apabila pembayaran sebelum tanggal 20 dan bernilai 0 apabila pembayaran setelah tanggal 20)
9	SHOP RANK	Faktor	Peringkat perusahaan untuk suatu toko dalam istilah komersial
10	RESIDENCE TYPE P	Faktor	Kepemilikan tempat tinggal (bernilai 1 apabila Owned dan bernilai 0 apabila lainnya)
11	RESIDENCE TYPE O	Faktor	Kepemilikan tempat tinggal (bernilai 1 apabila Other dan bernilai 0 apabila lainnya)
12	RESIDENCE TYPE A	Faktor	Kepemilikan tempat tinggal (bernilai 1 apabila Rented dan bernilai 0 apabila lainnya)
13	YEAR IN RESIDENCE	Numerik	Lama tinggal di tempat tinggal saat ini (dalam tahun)
14	FLAG MOTHER NAME	Faktor	Mencantumkan nama ibu kandung saat pembuatan akun kartu kredit (bernilai 1 apabila Yes dan bernilai 0 apabila No)
15	FLAG FATHER NAME	Faktor	Mencantumkan nama ayah kandung saat pembuatan akun kartu kredit (bernilai 1 apabila Yes dan bernilai 0 apabila No)
16	FLAG RESIDENCE TOWN	Faktor	Bekerja dan tinggal di kota yang sama (bernilai 1 apabila Yes dan bernilai 0 apabila No)
17	FLAG RESIDENCE STATE	Faktor	Bekerja dan tinggal di state yang sama (bernilai 1 apabila Yes dan bernilai 0 apabila No)
18	YEAR IN THE JOB	Numerik	Lama bekerja di perkerjaan saat ini (dalam tahun)
19	MATE INCOME	Numerik	Pendapatan per kapita suami-istri dalam sebulan (dalam mata uang Brazil)
20	FLAG RESIDENCIAL ADDRESS	Faktor	Menerima paket di alamat yang sama dengan alamat tempat tinggal
21	PERSONAL REFERENCE 1	Faktor	Nama depan berdasarkan referensi pribadi (dalam bahasa Portugal)
22	PERSONAL REFERENCE 2	Faktor	Nama depan berdasarkan referensi pribadi (dalam bahasa Portugal)

No.	Nama Peubah	Tipe	Penjelasan
23	PERSONAL NET INCOME	Numerik	Pendapatan per kapita individu dalam sebulan (dalam mata uang Brazil)
24	QUANT ADDITIONAL CARDS IN THE APPLICATION	Faktor	Jumlah kartu tambahan yang diminta dalam formulir pembuatan
25	TARGET LABEL	Faktor	Label atau kelas: *Bad* = 1, *Good* = 0

Eksplorasi Data

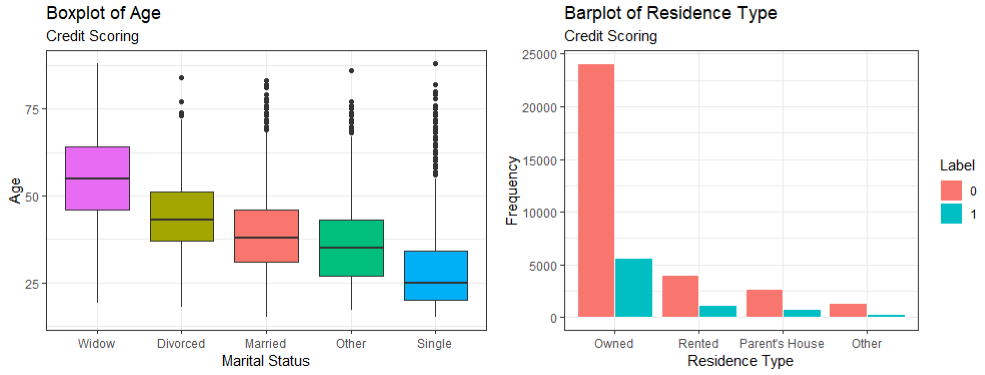
Eksplorasi data digunakan untuk mendapatkan informasi dari data *credit risk scoring*. Eksplorasi dilakukan untuk beberapa peubah yang dianggap cukup menarik untuk dibahas. Peubah tersebut seperti SEX, MARITAL STATUS, AGE, APPLICANT'S AGE, FLAG RESIDENCIAL PHONE, RESIDENCE TYPE, APPLICANT'S PAYMENT DAYS, APPLICANT'S YEAR IN RESIDENCE, dan APPLICANT'S YEAR IN THE JOB.



(a) (b)

Gambar 2 (a) Barplot peubah SEX berdasarkan Peubah LABEL dan (b) Barplot peubah MARITAL STATUS berdasarkan Peubah LABEL

Berdasarkan Gambar 2 (a) dapat disimpulkan bahwa baik dari jenis kelamin laki-laki maupun perempuan, lebih banyak pemilik kartu kredit yang berkategori "*good creditor*" daripada "*bad creditor*". Pada data yang kami miliki pemilik kartu kredit dengan jenis kelamin perempuan memiliki jumlah yang lebih banyak daripada laki-laki, sehingga wajar apabila dalam plot terlihat frekuensi perempuan jauh lebih tinggi. Pada Gambar 2 (b) dapat dilihat bahwa dari kelima status perkawinan yaitu *single, married* (menikah), *widow* (janda/ duda karena pasangan meninggal), *divorced* (bercerai), dan *other* frekuensi pemilik kartu kartu kredit berkategori *good creditor* lebih banyak dari kategori *bad creditor*. Selain itu, dapat dikatakan juga bahwa pemilik kartu kredit mayoritas berstatus *single*, selanjutnya didominasi pemilik kartu kredit berstatus *married*, kemudian *other*, *widow*, dan *divorced*.



(a) (b)

Gambar 3 (a) Boxplot peubah AGE berdasarkan Peubah MARITAL STATUS dan (b) Barplot peubah RESIDENCE TYPE berdasarkan Peubah LABEL

Dilihat dari Gambar 3 (a) dapat dikatakan bahwa pemilik kartu kredit dengan status *widow* rata-rata berusia di atas 50 tahun. Pemilik kartu kredit dengan status *single* rata-rata berusia 25 tahun, namun terdapat banyak pencilan yang usianya di atas 50 tahun bahkan di atas 75 tahun masih berstatus *single*. Kemudian pemilik kartu kredit dengan status *married* rata-rata berusia lebih dari 30 tahun, namun tetap saja terdapat pencilan pemilik kartu kredit yang berstatus *married* berusia di atas 60 tahun. Untuk pemilik kartu kredit dengan status *other* rata-rata berusia 30 tahun, dan terdapat beberapa pencilan yang berusia diatas 60 tahun. Sedangkan pemilik kartu kredit yang berstatus *divorced* rata-rata berusia 40 tahun, dan tetap terdapat pencilan yang berusia di atas 60 tahun. Dari Gambar 3 (b) terlihat bahwa pemilik kartu kredit yang memiliki rumah sendiri lebih banyak dibandingkan pemilik kartu kredit dengan status kepemilikan tempat tinggal yang lain. Selain itu untuk semua status kepemilikan tempat tinggal, kategori *good creditor* selalu memiliki frekuensi yang lebih tinggi dibandingkan dengan kategori *bad creditor*.

Re-sampling

Untuk mengatasi masalah imbalanced data, diterapkan 4 macam metode *re-sampling*, yaitu *over sampling*, *under sampling*, SMOTE (*Synthetic Over Sampling Technique*), dan ROSE (*Random Over Sampling Example*). Jumlah data baru ditampilkan pada tabel 3.

Tabel 3. Jumlah Target pada Data Latih dan Data Uji berbagai Metode *Re-sampling*

Data Credit Risk	**Total Data**		**Data Latih**		**Data Uji**	
	Good	**Bad**	**Good**	**Bad**	**Good**	**Bad**
Data Original	32083	7894	22512	5472	9571	2422
Data Over Samp	32083	32189	22355	22636	9571	2422
Data Under Samp	7863	7894	5559	5471	2304	2423
Data SMOTE	31576	23682	22182	16499	9394	7183
Data ROSE	27600	27658	19383	19298	8217	8360

Klasifikasi Data

Menurut Ganganwar (2012), beberapa metode yang dapat memiliki performa yang baik untuk kasus *data imbalanced* yaitu *Support Vector Machine*, *k-Nearest Neighbor* (KNN), *Neural Networks*, Regresi Logistik, dan *Decision Tree*. Selain menggunakan beberapa algoritma, peneliti menggunakan empat metode sampling, yaitu *under sampling, over sampling, Synthetic Oversampling Technique* (SMOTE), dan *Random Over Sampling Example* (ROSE).

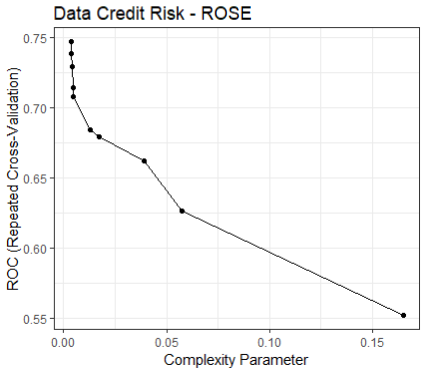
Tabel 4. Akurasi Klasifikasi *10-Fold Repeated Cross Validation Decision Tree* untuk Data Latih

No.	**Algoritma**	**Accuracy**	**AUC**	**Sensitivity**	**Specificity**	**G-Mean**
1	DT	0.8117	0.6155	0.9906	0.0843	0.2889
2	DT – OS	0.6246	0.6472	0.5849	0.6641	0.6232
3	DT – US	0.6260	0.6401	0.5949	0.6569	0.6251
4	DT – SMOTE	0.7194	0.7471	0.8452	0.5517	0.6828
5	DT – ROSE	0.7734	0.8211	0.7334	0.8133	0.7723
6	LR	0.8032	0.6585	0.9990	0.0076	0.0871
7	LR – OS	0.6173	0.6633	0.5873	0.6471	0.6164
8	LR – US	0.6165	0.6632	0.5895	0.6435	0.6159
9	LR – SMOTE	0.7284	0.7892	0.8183	0.6085	0.7056
10	LR – ROSE	0.7137	0.7838	0.7460	0.6816	0.7130

Tabel 5. Akurasi Klasifikasi *10-Fold Repeated Cross Validation Decision Tree* untuk Data Uji

No.	Algoritma	Accuracy	AUC	Sensitivity	Specificity	G-Mean
1	DT	0.7999	0.5195	0.9828	0.0561	0.2348
2	DT – OS	0.6192	0.6244	0.5789	0.6595	0.6178
3	DT – US	0.5950	0.5949	0.5645	0.6254	0.5941
4	DT – SMOTE	0.7097	0.6888	0.8349	0.5428	0.6731
5	DT – ROSE	0.7654	0.7653	0.7238	0.8069	0.7642
6	LR	0.8028	0.5035	0.9981	0.0088	0.0937
7	LR – OS	0.6182	0.6181	0.5885	0.6477	0.6173
8	LR – US	0.6090	0.6089	0.5755	0.6423	0.6079
9	LR – SMOTE	0.7209	0.7058	0.8115	0.6001	0.6978
10	LR – ROSE	0.7145	0.7145	0.7432	0.6858	0.7139

Berdasarkan tabel 4 dan tabel 5, kita dapat mengetahui bahwa algoritma *Decision Tree* dengan metode *re-sampling Random Over Sampling Example* (ROSE) memiliki metrik evaluasi AUC dan *Specificity* yang paling tinggi dibandingkan dengan metode yang lain. Sehingga metode tersebut dipilih sebagai model untuk menangani kasus *credit risk scoring*. Selanjutnya dilakukan *tuning* parameter *Decision Tree* untuk meningkatkan performa dari model tersebut. Parameter yang digunakan adalah *Complexity Parameter* (CP). Nilai dari parameter yang akan di-*tuning* didasarkan pada nilai parameter pada pembelajaran data latih. Nilai parameter CP pada pembelajaran data latih ditampilkan pada gambar 4 dan tabel 6.

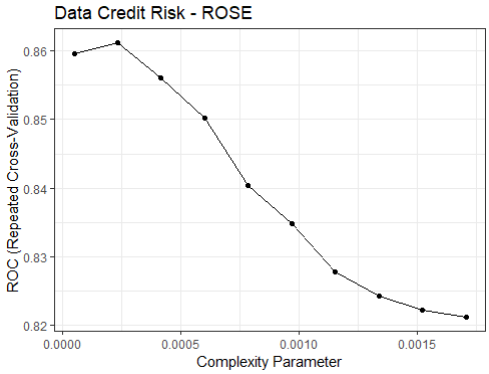


Gambar 4 Scatterplot antara *Complexity Parameter* dengan ROC

Tabel 6. CP dan AUC pada Data ROSE

Compexity P	AUC
0.001708075	0.8211959
0.001811594	0.8209881
0.002639752	0.8194456
0.002691511	0.8187812
0.002846791	0.8159654
0.003053830	0.8157833
0.004710145	0.8149570
0.005124224	0.8146032
0.009575569	0.7616219
0.217261905	0.6400299

Tuning parameter dilakukan dengan mempertimbangkan nilai dari CP dan AUC di tabel 5. Secara matematis, nilai AUC semakin meningkat dengan berkurangnya nilai CP. Sehingga parameter di-*tuning* untuk nilai di bawah nilai 0.001708075. *Tunning* parameter sebanyak 10 nilai parameter dari 0.00005 – 0.001708075. Selanjutnya hasil dari *tuning* parameter ditampilkan pada tabel 6.



Gambar 5 Scatterplot antara *Complexity Parameter* dengan ROC

Tabel 7. CP dan ROC Tuning ROSE

Compexity P	AUC
0.0000500000	0.8594917
0.0002342306	0.8611909
0.0004184611	0.8559946
0.0006026917	0.8502280
0.0007869222	0.8404080
0.0009711528	0.8348941
0.0011553833	0.8278051
0.0013396139	0.8243261
0.0015238444	0.8221928
0.0017080750	0.8211959

Berdasarkan pada gambar 5 dan tabel 7 diperoleh bahwa nilai untuk nilai *Complexity Parameter* di bawah 0.001708075 memiliki titik puncak AUC, yaitu pada CP 0.0002342306 menghasilkan AUC 0.8611909. Sehingga nilai CP tersebut dipilih sebagai parameter *Decision Tree* yang memiliki AUC terbaik untuk data latih. Selanjutnya model tersebut diterapkan pada data uji.

Tabel 8. Akurasi *Decision Tree* dengan *Tuning* Parameter CP 0.0002342306

No.	Algoritma	Accuracy	AUC	Sensitivity	Specificity	G-Mean
1	Data Latih ROSE	0.8208	0.8611	0.8077	0.8338	0.8206
2	Data Uji ROSE	0.7895	0.7895	0.7738	0.8052	0.7893

Pemilihan Peubah Berpengaruh

Berdasarkan hasil pra-proses yang dilakukan pada peubah data *credit risk scoring* diperoleh 24 peubah bebas yang akan digunakan untuk melakukan klasifikasi. Melalui algoritma *Decision Tree* dengan nilai AUC dan *specificity* tertinggi, kita memperoleh 20 peubah dari 25 peubah yang dianggap berpengaruh terhadap peubah target. Peubah tersebut ditampilkan pada tabel 9.

Tabel 9. Peubah berpengaruh pada TARGET LABEL (Decision Tree CP 0.0002342306)

No.	Nama Peubah	Persentase Kumulatif	Persentase
1	MATE INCOME	100.000	56.458
2	RESIDENCE TYPE P1	43.542	0.438
3	FLAG FATHERS NAME1	43.104	2.427
4	PERSONAL REFERENCE .21	40.677	0.007
5	FLAG RESIDENCIAL PHONE1	40.670	3.608
6	AGE	37.062	19.413
7	YEAR IN THE JOB	17.649	0.952
8	SEX1	16.697	0.116
9	PERSONAL NET INCOME	16.581	0.605
10	RESIDENCE TYPE A1	15.976	3.281
11	PAYMENT DAY1	12.695	6.405
12	YEAR IN RESIDENCE	6.290	0.229
13	FLAG RESIDENCIAL ADDRESS.POSTAL ADDRESS1	6.061	0.052
14	RESIDENCE TYPE O1	6.009	1.002
15	MARITAL STATUS O1	5.007	1.278
16	MARITAL STATUS S1	3.729	0.336
17	MARITAL STATUS C1	3.393	1.408
18	QUANT ADDITIONAL CARDS IN THE APPLICATION1	1.985	0.281
19	FLAG RESIDENCE TOWN.WORKING TOWN1	1.704	1.069
20	MARITAL STATUS V1	0.635	0.635

SIMPULAN DAN REKOMENDASI

Simpulan

Simpulan yang didapatkan berdasarkan penelitian mengenai penggunaan *Random Over Sampling Example* (ROSE) untuk penanganan ketidakseimbangan kelas pada kasus *Credit Risk Scoring* berbasis *Decision Tree* dan Regresi Logistik adalah bahwa klasifikasi menggunakan algoritma *Decision Tree* dengan metode *sampling Random Over Sampling Example* (ROSE) memiliki metrik evaluasi AUC dan *specificity* yang tinggi dibandingkan dengan metode *sampling* lain maupun algoritma regresi logistik. Model yang diperoleh memiliki *Complexity Parameter* (CP) optimal yaitu 0.0002342306 menghasilkan AUC sebesar 0.8611909 pada data latih dan menghasilkan AUC sebesar 0.7895 pada data uji. Hal

ini menunjukkan bahwa model yang didapatkan tidak *overfitting*. Penelitian ini juga memperoleh peubah yang dianggap berpengaruh dalam penentuan kelas target. Dua di antaranya adalah peubah MATE INCOME sebesar 56.468% dan AGE sebesar 19.413%.

Rekomendasi

Rekomendasi yang dapat diberikan kepada pihak bank dalam menilai *credit condition* nasabah baru berdasarkan penelitian mengenai Penggunaan *Random Over Sampling Example* (ROSE) untuk Penanganan Ketidakseimbangan Kelas pada Kasus *Credit Risk Scoring* berbasis *Decision Tree* dan Regresi Logistik adalah: (1) Menggunakan metode *Decision Tree* dengan metode *sampling Random Over Sampling Example* (ROSE); (2) Memperbaiki proses pencatatan data nasabah calon kreditur agar lebih detail dan mengurangi data inkonsisten pada pembuatan model; dan (3) Menerapkan beberapa algoritma klasifikasi untuk *credit risk scoring* lain seperti *Support Vector Machine* (SVM), *Random Forest*, *Gradient Boosting Methods* (GBM), dan *Neural Network* dengan kombinasi metode *sampling* agar didapatkan model dengan akurasi yang lebih optimal.

DAFTAR PUSTAKA

Aritonang G. 2006.Klasifikasi Imbalanced Data Menggunakan Algoritma Klasifikasi Voting Feature Intervals [skripsi]. Bogor (ID): Institut Pertanian Bogor.

Barandela R, Sanchez JS, Garcia V, Rangel E.2002. *Strategies for Learning in class imbalance problems, Pattern Recognition* [internet]. [diunduh 2005 Des 22]; 36(3): 849-850. Tersedia pada: http://sci2s.ugr.es/keel/monografia/unbalanced/imbalance-classes.pdf.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. *SMOTE : Synthetic Minority Over-Sampling Technique*. 16:321-357.

Frendy, Surjandari I.*Pembentukan Model Credit Scoring dengan Menggunakan Metode Bayesian Network : Studi Kasus Permohonan Aplikasi Kredit Pemilikan Rumah (KPR)*. Depok (ID) : FT UI.

Gouvenir HA, Demiroz G, Ilter N. 1998. Learning Differential Diagnosis of Erythemato-Squamous Diseases using Voting Feature Intevals. *Artificial Intelligence in Medicine*. 13(3):147-165

Guyon I. 2003. An Introduction to Variable and Feature Selection 1 Introduction. *Journal of Machine Learning Research*. 3:1157-1182.

Han J, Kamber M. 2001. *Data Mining Concepts and Techniques*. San Fransisco (USA): Academic Press

Irawan E, Wahono RS. 2015. Penggunaan *Random Under Sampling* untuk Penanganan Ketidakseimbangan Kelas pada Prediksi Cacat Software Berbasis *Neural Network*. *Journal of Software Engineering*. 1(2):92-100.

Jain M, Richariya V. 2012. An Improved Techniques Based on Naivee Bayesian for Attack Detection. *International Journal of Emerging Technology and Advanced Engineering*. 1(2):324-331.

Sukmawati AP. 2015. Integrasi SMOTE dan Informastion Gain pada Naive Bayes untuk Prediksi Cacat Software. *Journal of Software Engineering*. 1(2):86-91.

Tampil YA, Komalig H, Langi Y. 2017. Analisis Regresi Logistik untuk Menentukan Faktor-Faktor yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado. *JdC*. 6(2):56-62.

Yap BW, Ong, Husain NH. 2011. Using Data Mining to Improve Assessment of Credit Worthiness Via Credit Scoring Models. *Expert Systems with Applications*. 38:13274-13283.

Yu D, Hu J, Tang Z, Shen H, Yang J, Yang J. 2013. *Neurocomputing Improving protein-ATP binding residues prediction by boosting SVMs with random under-sampling* [internet]. [diunduh 2010 Okt 10]; 104:180–190. Tersedia pada: http://doi.org/10.1016/j.neucom.