Analisa Akurasi Dan F1 Score Pada Algoritma Smote Dan Naïve Bayes Pada Dataset Bank Direct Marketing

Amin Nur Rais[1], Warjiyono[2] , Wawan Kurniawan[3], Rian Ardianto[4]
Sistem Informasi Akuntansi Universitas Bina Sarana Informatika [1, 2, 3, 4]
Jl. Kramat Raya No 98, Jakarta Pusat
amin.arv@bsi.ac.id[1], warjiyono.wrj@bsi.ac.id[2] , wawan.wwk@bsi.ac.id[3] , rian.rao@bsi.ac.id[4]

Abstract − *Introducing products directly is done by many industries, one of the industries that utilize direct marketing is the banking industry. With the product introduction process, the amount of incoming data continues to grow, so that the data can be analyzed for bank marketing, and used to choose the type of marketing carried out. Data mining or often known as data mining is becoming a trend in processing data to get the information needed. Machine Learning becomes a model to find certain patterns to help in making decisions in direct marketing. In the study the process of processing data with machine learning by comparing the results of naïve bayes classification algogirma with the use of smote. The testing model is divided into 2 tests, with smote and without smote. The test results show that the use of smote against the naïve bayes classification algorithm has an influence on the accuracy produced. At the level of accuracy, the comparison between using smote and not using smote, is better when not using smote with an accuracy of 87.52%. Whereas in the F1 Score calculation, the F1 score is better when using a smote of 80.26%.*

Keywords: accuracy, f1 score, smote, naïve bayes

Abstrak – Mengenalkan produk secara langsung banyak dilakukan oleh berbagai industi, salah satu industry yang memanfaatkan pemasaran secara langsung adalah industry perbankan. Dengan dilakukannya proses pengenalan produk, jumlah data yang masuk terus bertambah,sehingga data dapat dianalisa terhadap pemasaran bank, dan digunakan untuk memilih jenis pemasaran yang dilakukan. Penambangan data atau yang sering dikenal dengan data mining menjadi trend dalam melakukan pengolahan data untuk mendapatkan informasi yang dibutuhkan. Machine Learning menjadi model untuk menemukan pola tertentu untuk membantu dalam mengambil keputusan dalam melakukan pemasaran secara langsung. Pada penelitian dilakukan proses pengolahan data dengan machine learning dengan membandingkan hasil pengujian algogirma klasifikasi naïve bayes dengan penggunaan smote. Model pengujian dibagi menjadi 2 pengujian, dengan smote dan tanpa smote. Hasil pengujian menunjukkan bahwa penggunaan smote terhadap algoritma klasifikasi naïve bayes memiliki pengaruh terhadap akurasi yang dihasilkan. Pada tingkat akurasi, perbandingan antara penggunaan smote dan tidak menggunakan smote, lebih baik pada saat tidak menggunakan smote dengan akurasi 87,52%. Sedangkan pada perhitungan F1 Score, nilai F1 score lebih baik pada saat menggunakan smote sebesar 80,26%..

Kata Kunci: akurasi, f1 score, smote, naïve bayes.

1.a Latar Belakang

Mengenalkan produk secara langsung banyak dilakukan oleh berbagai industi, salah satu industry yang memanfaatkan pemasaran secara langsung adalah industry perbankan. Kampanye pemasaran produk bank secara langsung sangat berguna untuk menawarkan produk baru kepada calon pelanggan. Dalam mengenalkan produk secara langsung, bank dapat melakukan analisa pasar dengan memanfaatkan ruang teknologi informasi yang dapat membantu dalam mengambil keputusan [1]. Proses analisa dapat dilakukan dengan teknologi informasi yang didukung dengan terus bertambahnya informasi data pelanggan untuk mendukung pengambilan keputusan [2]. Dengan

bertambahnya waktu, jumlah data yang masuk terus bertambah,sehingga data dapat dianalisa terhadap pemasaran bank, dan digunakan untuk memilih jenis pemasaran yang dilakukan. Kampanye pemasaran dapat dilakukan melalui email, telepon, dan email langsung kepada calon pelanggan yang memungkinkan calon pelanggan dapat memutuskan untuk mengambil produk yang ditawarkan atau tidak.

Penambangan data atau yang sering dikenal dengan data mining menjadi trend dalam melakukan pengolahan data untuk mendapatkan informasi yang dibutuhkan. Dalam melakukan penambangannya, data mining menggunakan teknik machine learning untuk membaca dan mengekstrak data yang tersedia. Salah satu data yang dapat diolah adalah data tentang perbankan. Machine Learning menjadi model untuk menemukan pola tertentu untuk membantu dalam mengambil keputusan dalam melakukan pemasaran secara langsung. Mengolah dataset yang tidak seimbang menjadi masalah renspons area dan menghasilkan kinerja model yang buruk. Tetapi dengan model klasifikasi Neural Network dapat memberikan klasifikasi yang baik. Akan tetapi perlu dipertimbangkan dalam mengevaluasi model ansambel, yaitu ukuran sample dari data training yang dapat mempengaruhi kinerja bagging, boosting dirancang untuk model klasifikasi yang buruk dari sejumlah besar data sementara bagging cenderung lebih berguna untuk mengatasi masalah klasifikasi dengan jumlah data yang terbatas [3].

Dalam mengolah data yang asimetris, digunakan algoritma SMOTE dan Rotation Forest (PCS)-J48 dimana SMOTE digunakan untuk memodifikasi data dan meningkatkan keakuratan prediksi. Dengan menggunakan PCA-J48 mendapatkan nilai akurasi yang tinggi dengan nilai 81,23% dan spesifisitas 93,17%. Namun sensifitas metode BayesNet dan PCA-RandomTree memiliki nilai sensitivitas lebih tinggi dari PCA-J48 [4].

Dalam memprediksi respon pelanggan, dengan menerapkan empat pengklasifikasian, yaitu Multilayer Perceptron Neural Network (MLPNN), Decision Tree (C4.5), Regresi Logistik, dan Random Forest (RF). Penelitian ini menggambarkan bahwa klasifikasi dengan RF menjadi klasifikasi paling produktif dalam kemampuan prediksi dengan nilai akurasi 87%. Sedangkan fitur utama dari pelanggan yang kemungkinan besar berlangganan berjangka jika pelanggan menghabiskan lebih lama dalam panggilan dengan pendidikan minimal sekolah menengah [5].

Dalam menangani dataset yang tidak seimbang, terdapat 3 pendekatan yang dapat dilakukan, yaitu pendekatan level data, level algoritmik, dan pendekatan dengan menggabungkan metode. Pada pendekatan level data, mencakup berbagai teknik resampling untuk memperbaiki distribuasi kelas pada data. Pada level algoritmik, dilakukan proses penyesuaian operasi algoritma yang ada untuk menjalankan pengklasifikasian (classifier) agar lebibih konduktif terhadap klasifikasi kelas minoritas. Sedangkan pada pendekatan gabungan (ensemble), terdapat 2 ensemble-learning paling popular, yaitu boosting dan bagging. Pada pendekatan algoritma dan ensemble memiliki tujuan yang sama, yaitu memperbaiki algoritma pengklasifikasi tanpa mengubah data, sehingga dapat dianggap ada 2 pendekatan saja, yaitu pendekatan level data dan pendekatan level algoritma. Dengan membagi menjadi 2 pendekatan dapat mempermudah dalam proses perbaikan.

Bank dalam membuat keputusan untuk menyetujui atau menolak pinjaman, perlu disajikan kredibilitas dari pelanggan berdasarkan factor yang ditentukan. Dan didapatkan bahwa factor usia, durasi, dan jumlah menjadi factor yang penting yang dapat mempengaruhi seseorang dalam hal financial. Dari data yang didapatkan, ditemukan bahwa algoritma terbaik untuk pengklasifikasian kredit beresiko adalah Random Forest dengan akurasi 79,68% meskipun lebih lambat dalam runtime untuk dataset dimensi besar [6].

Untuk meningkatkan laba dari proses pemasaran, dilakukan proses pendekatan dari dataset dengan system dua lapis. Yang pertama dengan mengelompokkan pelanggan, dan kemudian membangun klasifikasi untuk menawarkan produk dengan direct-marketing. Metode pengelompokan dan klastering menunjukkan peningkatan positif pada akurasi [7].

Untuk mengidentifikasi pelanggan yang potensial dari dataset, diselidiki dengan mengklasifikasikannya dengan SVM Classifier dan algoritma AdaBoost. Hasil diperoleh menunjukkan bahwa akurasi SVM biasa

mendapatkan akurasi 91,67% dan sensivitas 83,8% sedangkan AdaBoost SVM memberikan akurasi hingga 95,07% dan sensivitas 91,65% dalam 30 iterasi.

Paper ini memiliki motivasi untuk meneliti akurasi dan f1 score dengan pendekatan klasifikasi. Untuk itu, diusulkan model dengan menguji akurasi menggunakan algoritma SMOTE yang dikombinasikan dengan model Naïve Bayes

1.b Rumusan Masalah

Berdasarkan pada uraian yang telah dijabarkan pada latar belakang dapat diidentifikasikan bahwa bentuk dataset bank direct marketing bersifat tidak seimbang sehingga hasil yang diperolah akan cenderung ke kelas mayoritas. Dan untuk meningkatkan hasil kinerja, diperlukan algoritma yang memiliki kinerja akurasi dan f1 score tinggi dalam melakukan klasifikasi .

1.c Batasan Masalah

Penelitian ini menggunakan dataset bank direct marketing dari UCI repository. aplikasi machine learning yang digunakan adalah WEKA versi 3.8.1. Pengolahan data awal menggunakan preprocessing SMOTE (Synthetic Minority Over-sampling Technique). Pengklasifikasi menggunakan Naïve Bayes. Evaluasi kinerja model menggunakan perbandingan nilai akurasi dan f1 score yang didapat dalam setiap uji coba model.

1.d. Tujuan

Tujuan dari penelitian ini adalah untuk meningkatkan hasil kinerja, diperlukan algoritma yang memiliki kinerja akurasi dan f1 score tinggi dalam melakukan klasifikasi

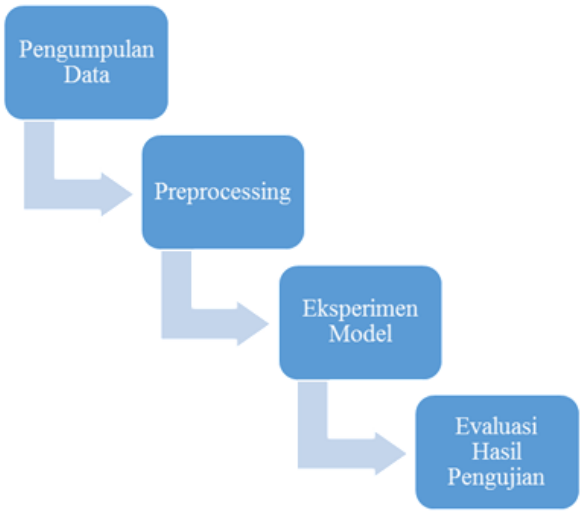
1.e. Manfaat Penelitian

Tujuan pada penelitian ini adalah menghasilkan model prediksi baru dalam menangani imbalance class pada dataset bank direct marketing dengan menerapkan model usulan yaitu dengan preprocessing SMOTE dengan algoritma klasifikasi Naïve Bayes. Penelitian ini diharapkan dapat menjadi model pembanding dengan penelitian-penelitian sejenis, terkait dengan peningkatan nilai akurasi dan f1 score pada dataset bank direct marketing.

1. f. Metode Penelitian

Pendekatan yang digunakan pada penelitian ini yakni pendekatan yang bersifat kuantitatif. Aspek kuantitatif memberikan penekanan bahwa pengukuran merupakan dasarnya, karena memberikan hubungan antara observasi dan formalisasi model, teori dan hipotesis. Dalam melakukan penelitian, dilakukan tahap – tahap dengan kerangka untuk melakukan penelitian yang dilakukan sebagaimana digambarkan pada gambar 1 sebagai acuan penelitian.

Pada proses pengumpulan data, data yang digunakan menggunakan dataset public yang diambil dari kaggle repository tentang pemasaran langsung yang dilakukan oleh bank di Portugis dengan 41.188 data yang terdiri dari 20 atribut dan 2 kelas. dataset ini berkaitan dengan pemasaran secara langsung yang dilakukan oleh lembaga perbankan Portugis kepada calon pelanggannya. Pemasaran ini didasarkan pada panggilan telepon. Biasanya perlu lebih dari satu kali proses penawaran kepada calon konsumen yang sama untuk memastikan apakah akan menggunakan prosuk yang ditawarkan atau tidak. Dataset ini terbagi menjadi 4 bagian (table 1), yaitu data client, releated with the last contact of the current campaign, other atribut, dan social and economic context attributes.



Gambar 1. Gambar Kerangka Penelitian

Dataset yang digunakan akan diuji dengan berbagai pemodelan klasifikasi dengan teknik K-Fold Validation (10). Dimana proses pengujian

akan dilakukan dengan membagi kedalam 2 segmen preprocessing. Terdapat 4 jenis preprocessing yang diujikan, yaitu tanpa preprocessing (none) dan dengan SMOTE yang digabungkan dengan algoritma naïve bayes.

Tabel 1. Ringkasan Dataset

No	Attribute Name	Attribut Type
Data Client		
1	Age	Numeric
2	Job	Categorical
3	Marital	Categorical
4	Education	Categorical
5	Default	Categorical
6	Housing	Categorical
7	Loan	Categorical
Related With The Last Contact Of The Current Campaign		
8	Contact	Categorical
9	Month	Categorical
10	Day_of_week	Categorical
11	Duration	Numeric
Other Attributes		
12	Campaign	Numeric
13	Pdays	Numeric
14	Previous	Numeric
15	Poutcome	Categorical
Social And Economic Context Attributes		
16	Emp.var.rate	Numeric
17	Cons.price.idx	Numeric
18	cons.conf.idx	Numeric
19	Euriborn3m	Numeric
20	Nr.employed	Numeric
Output		
21	Y	Binary

2.a. Dasar Teori

Dasar teori dilakukan untuk mengetahui nilai akurasi dan f1 score dalam studi kasus ketidakseimbangan kelas pada dataset bank direct marketing serta mengetahui dasar sumbernya. Tinjauan pustaka meliputi penjelasan tentang bank direct marketing, imbalance class, SMOTE dan Naive Bayes. Tinjauan studi ini akan digunakan sebagai landasan penelitian agar dapat mengetahui state of the art tentang penelitian bank direct marketing yang membahas tentang imbalance class.

2.b. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) adalah metode oversampling yang digunakan untuk menangani masalah ketidakseimbangan kelas, tetapi metode ini memiliki keterbatasan sampai batas tertentu karena sampel baru disintesia di antara contoh minoritas yang berdekatan, sehingga metode ini tidak bisa menunjukkan distribusi data yang lengkap (Sumasdhi & Hemalatha, 2013). SMOTE merupakan pendekatan oversampling pada kelas minoritas yaitu dengan melakukan oversampling untuk menciptakan sampel "sintetik".

Metode SMOTE diusulkan sebagai salah satu solusi dalam menangani data tidak seimbang dengan prinsip yang berbeda dengan Metode oversampling yang telah diusulkan sebelumnya [9]. Jika Metode oversampling memiliki dengan memperbanyak pengamatan secara acak, Metode SMOTE menambah jumlah data kelas minor agar setara dengan kelas mayor dengan cara membangkitkan data buatan. Metode ini mensintesis sampel kelas minoritas baru antara beberapa contoh minoritas yang terletak berdekatan, bukan hanya menduplikasi mereka secara acak menggunakan ROS [10]. Jumlah k-tetangga terdekat ditentukan dengan mempertimbangkan kemudahan dalam melaksanakannya. Pembangkitan data buatan yang berskala numerik berbeda dengan kategorik. Data numerik diukur jarak kedekatannya dengan jarak Euclidean sedangkan data kategorik lebih sederhana yaitu dengan nilai modus [11]. Akhirnya, data pada saat itu dimasukkan sebagai contoh dari minoritas baru. Dengan menambahkan sampel minoritas baru ke dalam data pelatihan, diharapkan over-fitting dapat diselesaikan.

2.c. Naive Bayes

Naive Bayes merupakan salah satu metode machine learning yang menggunakan metode

probabilitas. Teorema bayes dapat dijelaskan sebagai berikut:

$$P(C|x) = \frac{P(C)P(x|C)}{P(x)}$$

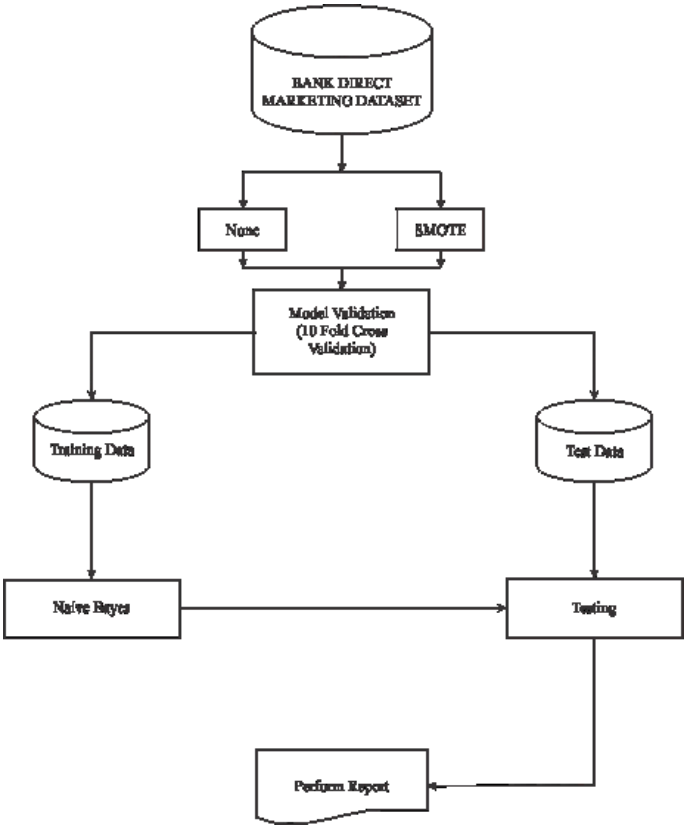
Dimana:
C : Kelas
$P(x|C)$: Kemungkinan posterior x pada kondisi kelas C
$P(C)$: Kemungkinan kelas C tanpa memandang bukti apapun
$P(x)$: Kemungkinan posterior x tanpa memandang kelas/bukti lain

Metode Naive bayes ini sangat mudah untuk dibangun, tidak memerlukan skema estimasi parameter secara iteratif yang kompleks. Hal ini berarti dapat segera diterapkan untuk kumpulan data yang besar. Berikut ini adalah algoritma Naïve Bayes:

- Masukan: Data latih T, data uji x
- Hitung mean (rata-rata) dan standar deviasi setiap kelas
- Hitung nilai probabilitas data uji untuk setiap kelas
- Klasifikasikan data uji sesuai nilai probabilitas kelas yang tertinggi
- Keluaran: Hasil klasifikasi

1. Metode Penelitian

Hasil dari penelitian ini merupakan model dalam pemodelan klasifikasi pada data imbalance class dengan menggunakan algoritma SMOTE dengan algoritma Naïve Bayes sebagai classifiernya. Dimana model yang diusulkan dilakukan percobaan – percobaan untuk mengetahui hasil akurasi dan F1 Score dari model yang diusulkan.



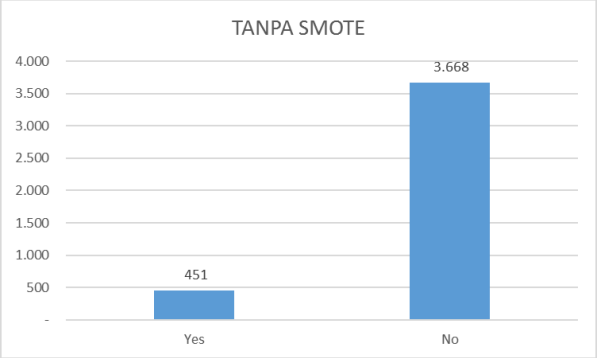
Gambar 2.Metode Penelitian

Dengan menggunakan metode penelitian yang diusulkan, proses ujicoba dapat dilakukan dengan beberapa percobaab, yaitu dengan melakukan percobaan langsung dengan algoritma klasifikasi naïve bayes, dan dengan penggabungan algoritma SMOTE dengan algoritma naïve bayes. Haril pengujian akan menghasilkan table confusion matrix yang nantinya akan digunakan sebagai landasan untuk menghitung akurasi dan f1 score.

4. Implementasi dan Hasil

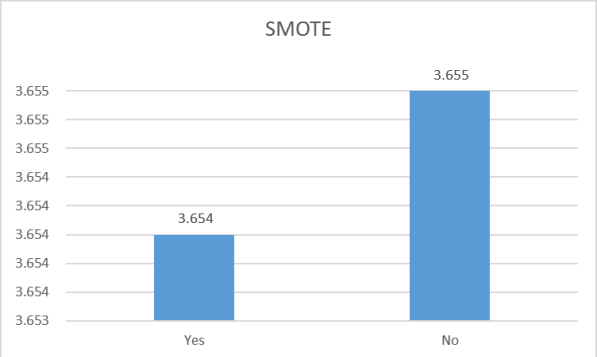
Dalam penelitian ini menggunakan dataset publik yang diambil dari UCI machile learning dengan jumlah jumlah data 41.188 yang terbagi menjadi 20 atribut dan 2 kelas. dataset yang digunakan berkaitan dengan model pemasaran secara langsung yang dilakukan oleh lembaga perbankan portugis kepada para calon nasabahnya. Untuk melakukan penelitian imbalance class menggunakan perangkat laptop dengan processor i5-2520M ram 4gb dengan sistem operasi windows 7 proffesional 64 bit.

Dalam pengolahan datanya, digunakan tools WEKA 3.8.1.



Gambar 3. Jumlah data tanpa SMOTE

Setelah data terkumpulkan, preprocessing data dilakukan untuk menangani ketidak seimbangan kelas pada datanya dengan menggunakan teknik SMOTE. Dengan menggunakan algoritma SMOTE, akan menciptakan data sintesis dari kelas minoritas. Algoritma SMOTE berguna untuk emnghasilkan data yang lebih efektif dan lebih baik dalam menangani ketidakseimbangan kelas yang over-fitting pada proses over-sampling pada kelas minoritas.



Gambar 4. Jumlah data dengan SMOTE

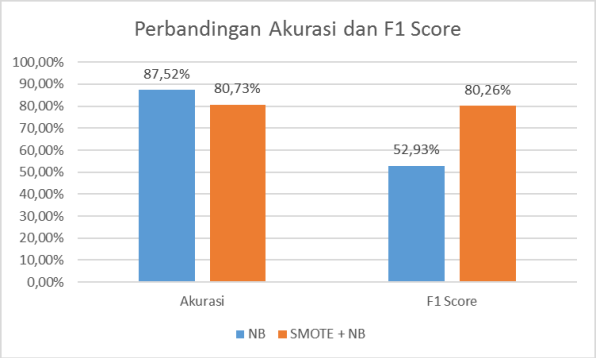
Pengujian pertama dilakukan dengan menggunakan algoritma naïve bayes, kemudian pada pengujian kedua menggunakan algoritma SMOTE dan naïve bayes. Dari kedua pengujian tersebut didapatkan nilai confusion matrix pengujian, kemudian dihitung nilai akurasi dan f1 score.

Tabel 2. Hasil confusion matrix

No	Algoritma	TP	FN	FP	TN
1	NB	289	175	339	3.316
2	SMOTE + NB	2.863	788	620	3.035

Rumus :
Akurasi = (TP+TN) / (TP+TN+FP+FN)
Recall = TP / (TP+FN)
Precision /PPV = TP / (TP+FP)
F1 Score = (2 x Recall x Precision) / (Recall + Precision)

Dari confusion matrix pada tabel 2, dapat dihitung dan diketahui bahwa nilai akurasi tanpa smote sebesar 87,52%, sedangkan jika digunakan algoritma smote menjadi 80,73%. Sedangkan pada nilai f1 score tanpa smote sebesar 52,93% dan jika menggunakan smote 80,26%. Sehingga dapat diketahui bahwa ketika pengujian tanpa menggunakan smote, memiliki akurasi lebih baik dari pada penggunaan smote, tetapi memiliki dampak f1 score yang lebih buruk dibandingkan dengan yang menggunakan smote.



Gambar 5. Perbandingan Akurasi dan F1 Score

5. Kesimpulan

Penggunaan smote terhadap algoritma klasifikasi naïve bayes memiliki pengaruh terhadap akurasi yang dihasilkan. Pada tingkat akurasi, perbandingan antara penggunaan smote dan tidak menggunakan smote, lebih baik pada saat tidak menggunakan smote dengan akurasi

87,52%. Sedangkan pada perhitungan F1 Score, nilai F1 score lebih baik pada saat menggunakan smote sebesar 80,26%.

6. Pustaka

[1] S. Abbas, "Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset," *Int. J. Comput. Appl.*, vol. 110, no. 3, pp. 1–7, 2015.

[2] S. Moro, R. M. S. Laureano, and P. Cortez, "Using Data Mining for Bank Direct Marketing: An application of the CRISP-DM methodology," *25th Eur. Simul. Model. Conf. ESM' 2011*, no. Figure 1, pp. 117–121, 2011.

[3] Y. Pan and Z. Tang, "Ensemble methods in bank direct marketing," *11th Int. Conf. Serv. Syst. Serv. Manag. ICSSSM 2014 - Proceeding*, 2014.

[4] P. Ruangthong and S. Jaiyen, "Bank direct marketing analysis of asymmetric information based on machine learning," *Proc. 2015 12th Int. Jt. Conf. Comput. Sci. Softw. Eng. JCSSE 2015*, pp. 93 – 96, 2015.

[5] J. Asare-Frempong and M. Jayabalan, "Predicting customer response to bank direct telemarketing campaign," *2017 Int. Conf. Eng. Technol. Technopreneurship, ICE2T 2017*, vol. 2017-Janua, pp. 1 – 4, 2017.

[6] A. Gahlaut, Tushar, and P. K. Singh, "Prediction analysis of risky credit using Data mining classification models," *8th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2017*, 2017.

[7] M. Mitik, O. Korkmaz, P. Karagoz, I. H. Toroslu, and F. Yucel, "Data Mining Based Product Marketing Technique for Banking Products," *IEEE Int. Conf. Data Min. Work. ICDMW*, pp. 552–559, 2017.

[8] T. Sumadhi and M. Hemalatha, "An Enhanced Approach for Solving Class Imbalance Problem in Automatic Image Annotation," *Int. J. Image, Graph. Signal Process.*, vol. 5, no. 2, pp. 9–16, 2013.

[9] N. V Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : Synthetic Minority Over-sampling Technique," vol. 16, pp. 321–357, 2002.

[10] V. García, J. S. Sánchez, R. Martín-Félez, and R. A. Mollineda, "Surrounding neighborhood-based SMOTE for learning from imbalanced data sets," *Prog. Artif. Intell.*, vol. 1, no. 4, pp. 347–362, 2012.

[11] S. Terhadap, D. Tidak, S. Pada, P. Model, and K. Jamu, "PENERAPAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE," vol. 1, no. 1, 2013.