PENERAPAN LIGHT GRADIENT BOOSTING DALAM PREDIKSI RASIO KLIK TAYANG

Kartika Handayani, Erni
Program Studi Sistem Informasi D3 Kampus Kota Pontianak, Fakultas Teknik dan Informatika
Universitas Bina Sarana Informatika, Pontianak, Indonesia
kartika.kth@bsi.ac.id

ABSTRAK

Prediksi rasio klik tayang adalah salah satu kriteria yang paling sering digunakan untuk menentukan efektivitas suatu iklan. Dalam produksi periklanan, prediksi klik tayang sangat berpengaruh bagi perusahaan yang memasang iklan tersebut. Selain memprediksi klik tayang suatu iklan, pemakaian model atau algoritma yang digunakan juga sangat penting dalam menganalisa klik tayang yang terjadi. Penelitian ini melakukan pengujian dengan dataset social network menggunakan light gradient boosting. Dilakukan pengujuan resampling untuk mengetahui pengaruh penggunaan pada dataset social network yang tidak seimbang. Hasil pengujian prediksi rasio klik tayang memperoleh hasil terbaik pada nilai *accuracy 91.25%, recall 93.10%, precision 84.38% F1-Score 88.52%* dan *AUC 0,92*. Hasil ini diperoleh dari pengujian tanpa *resampling* dan menggunakan *resampling* ROS.

Kata kunci: *iklan, klik tayang, light gradient boosting*

1. PENDAHULUAN

Prediksi Rasio klik tayang sangat penting dan dianggap sebagai salah satu cerita paling menguntungkan dalam domain pembelajaran mesin. Dalam klik tayang perlu diprediksi dengan akurat, karena hasil prediksi yang akurat menentukan apakah Rasio klik tayang tersebut tepat diklik atau tidak oleh konsumen yang melihat. Prediksi Rasio klik tayang telah digunakan dalam sejarah di setiap jenis format iklan seperti, mesin pencari, iklan tekstual dan kontekstual, iklan video, dan lain-lain. Peningkatan pesat jaringan iklan, prediksi rasio klik tayang membutuhkan analisis data yang sangat besar [1].

Advertising dikenal sebagai iklan internet atau pemasaran *online*, dan merupakan salah satu bisnis yang paling cepat berkembang dan paling menguntungkan[2] Dataset *Social Network Ads* adalah *dataset* iklan yang diperoleh dari repositori kumpulan data *Kaggle* dan merupakan kumpulan data standar [3].

Dalam produksi periklanan, prediksi klik tayang sangat berpengaruh bagi perusahaan yang memasang iklan tersebut. Perusahaan juga harus tahu, apakah iklan tersebut benar-benar dilihat oleh pengguna, atau hanya sekedar meng klik nya saja. Selain memprediksi klik tayang suatu iklan, pemakaian model atau algoritma yang digunakan juga sangat penting dalam menganalisa klik tayang yang terjadi.

Pendekatan *machine learning* memiliki kemampuan untuk mengatasi keterbatasan tersebut. Pendekatan ini telah dilakukan dalam beberapa penelitian sebelumnya dengan membandingan beberapa metode seperti *support vector machine, logistic regression, random forest,* dan *decission tree* [4][1]. Penelitian sebelumnya dengan dataset yang sama namun menggunakan metode yang berbedac dengan tujuan membandingkan beberapa dataset dan metode yang berbeda, belum menghasilkan tingkat AUC yang optimal yakni metode *Convolutional Neural Network* (CNN) memperoleh nilai AUC yaitu 69% data Avazu, 67% data Avito, 57% data Talking, 65% data Kad dan 64% data Corpon.

Berdasarkan penjelasan di atas, penelitian ini mengusulkan penerapan *machine learning* dalam prediksi rasio klik tayang. Data yang digunakan dalam penelitian memiliki masalah dengan ketidakseimbangan data. Sehingga dilakukan pengujian dengan metode *resampling* seperti *Random Under Sampling* (RUS), *Random Over Sampling* (ROS), *Synthetic Minority Over Sampling Technique* (SMOTE), dan SMOTE-Tomek. Penggunaan pendekatan ketidakseimbangan kelas sampling untuk menguji pada dataset yang digunakan dapat memberikan hasil performa yang lebih baik atau tidak.

Penelitian ini mengusulkan model prediktif *Light Gradient Boosting* dan diharapakan dapat menghasilkan *accuracy, recall, precision, f-score* dan *AUC* yang lebih baik dari penelitian sebelumnya dan menambah kontribusi penelitian terkait data yang digunakan.

2. TINJAUAN PUSTAKA

Untuk mendukung penelitian ini, tinjauan pustaka atau kerangka pemikiran yang relevan dibutuhkan agar penelitian ini berdasarkan pada sumber yang dihasilkan terstruktur dan sesuai pada intinya, khususnya pada masalah penelitian terkait.

2.1. Data Mining

Data mining adalah proses penggalian untuk kebutuhan informasi atau pengetahuan dalam volume data yang besar. Saat kami menerapkan penambangan data dalam volume data yang besar kemudian

mengungkap pola menarik dan hubungan tersembunyi. Nama alternatifnya adalah penambangan data dikenal sebagai pengetahuan ekstraksi, Penemuan pengetahuan, Analisis Data/Pola, Analisis data dalam database dll. [5]

Data mining bertujuan untuk menambah suatu pengetahuan dari data yang kita miliki. Data mining merupakan metode yang digunakan oleh suatu organisasi untuk mendapatkan sebuah informasi yang berguna dari data mentah yang diolah. Data mining dibagi menjadi 3 tipe yaitu:

1. Tipe data numerik merupakan tipe data yang diperoleh dengan cara pengukuran.
2. Tipe data kategorial merupakan tipe data yang diperoleh dengan cara klasifikasi atau kategorisasi.
3. Tipe data rentang waktu merupakan tipe data yang diperoleh dengan cara memperlihatkan beberapa objek yang berbeda.

Menurut [6]serangkaian proses tahapan memiliki tujuh (7) tahapan yaitu :

1. Pembersihan Data (*data cleaning*)
   Pembersihan data adalah proses untuk menghilangkan data-data yang tidak relevan. Data-data yang dibuang terkadang dibandingkan terlebih dahulu dengan hipotesa yang telah dibuat. Sehingga pada proses selanjutnya dapat dengan mudah menemukan hasil yang diinginkan.
2. Integrasi data (*data integration*)
   Integrasi data merupakan proses dalam menggabungkan data dari beberapa database kedalam satu database baru. Tidak sedikit data yang dibutuhkan diambil dari berbagai database atau teks file.
3. Seleksi data (*data selection*)
   Data yang sudah ada di database seringkali tidak semuanya dibutuhkan, maka dari itu dibutuhkan penyeleksian data untuk data yang benar-benar dibutuhkan dalam proses selanjutnya.
4. Transformasi data (*data transformation*)
   Data digabung atau diubah sesuai dengan proses yang digunakan dalam data mining. Karena beberapa format data mining membutuhkan format data yang khusus dalam pemrosesannya.
5. Proses *mining*
   Adalah proses menggali data dari sebuah database atau kumpulan data untuk memperoleh informasi yang tersembunyi dari data yang diolah
6. Evaluasi Pola (*pattern evaluation*)
   Dalam proses ini adalah hasil dari teknik data mining berupa pola-pola yang akan diuji pada hipotesa yang sudah dibuat sebelumnya. Sehingga akan memperoleh kesimpulan-kesimpulan yang mendekati hasil atau hipotesa untuk proses selanjutnya.
7. Presentasi pengetahuan (*knowlegde presentation*)

Ini termasuk dalam langkah akhir dari data mining dalam tahap ini saatnya untuk mempresentasikan hasil yang telah dilakukan dengan mengimplementasikan analisis yang didapat. Sehingga akan memperoleh kesimpulan *real*.

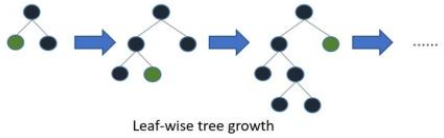
2.2. Klik Tayang

Klik tayang adalah jumlah klik pada promosi iklan yang diperoleh tergantung pada persentase klik pengunjung aktif diantisipasi yang terjadi. prediksi klik pada promosi iklan memungkinkan perusahaan Internet untuk membedakan iklan yang paling relevan untuk setiap pengguna dan lebih mengembangkan pengalaman pengguna. Lebih jelasnya, rasio klik-tayang (CTR), yang merupakan proporsi jumlah klik dengan jumlah tayangan, mungkin merupakan ukuran utama yang digunakan untuk menghitung nilai komersial sebuah iklan [7].

Rasio klik-tayang (CTR) digunakan untuk memperkirakan kemungkinan pengguna mengklik iklan atau produk yang ditampilkan kepada mereka. Iklan internet adalah pasar dengan minat yang signifikan, prediksi rasio klik-tayang adalah teknik utama periklanan internet dan merupakan metode penting untuk periklanan online dan evaluasi pemasaran [8].

2.3. *Light Gradient Boosting*

Lightgbm Menggunakan Peningkatan Gradien Dalam Konstruksinya, Tetapi GBM Ringan Tidak Membagi Nilai *Eigen* Satu Per Satu, Sehingga Perlu Untuk Menghitung Manfaat Pemisahan Dari Setiap Nilai *Eigen*. Algoritma Lightgbm Pada Model Untuk Meningkatkan Akurasi Dan Ketahanan Peramalan [9]. Dapat Menemukan Nilai *Split* Yang Optimal [10]. Jika Diilustrasikan Penambahan Pohon Pada Lightgbm Adalah Gambar 1.



Gambar 1. Algoritma Lightgbm [11]

2.4. *Resampling*

Teknik resampling akan memanipulasi kelas distribusi dengan memperbaiki data latih sehingga didapatkan kelas yang seimbang. Terdapat beberapa teknik resampling antara lain: random over sampling, random under sampling, directed over sampling, directed under sampling, kombinasi dari over sampling dan under sampling, dan Synthetic Minority Oversampling Technique (SMOTE) [12].

2.5. *Random Under Sampling* (RUS)

Teknik RUS secara acak menghilangkan sampel dari kategori massal, hingga mencapai keseimbangan kategori relatif [13]. Untuk pendekatan *under-sampling*, sebagian besar contoh kategori dibuang

sampai distribusi informasi yang seimbang tercapai. Metode data *merchandising* ini diselesaikan dengan segala cara. Mempertimbangkan kumpulan informasi dengan seratus *instance* kelas minoritas dan 1.000 *instance* kelas magnitudo, total 900 kategori yang mayoritas akan dihapus secara acak dalam teknik RUS. Di akhir proses, kumpulan data akan diseimbangkan dengan dua ratus *instance*. kelas halaman akan digambarkan dengan 100 *instance*, sedangkan minoritas juga akan memiliki 100 [14].

2.6. *Random Over Sampling* (ROS)

Algoritma ROS secara acak mereplikasi sampel dari kelas minoritas [13]. *Oversampling* dapat dilakukan dengan meningkatkan jumlah *instance* atau sampel kelas minoritas berdasarkan produksi contoh baru atau beberapa contoh berulang [15].

2.7. *Synthetic Minority Over Sampling Technique* (SMOTE)

Synthetic Minority Over Sampling Technique SMOTE menghasilkan sampel buatan dari kelas minoritas dengan menginterpolasi *instance* yang ada yang sangat dekat satu sama lain [13]. Untuk kategori minoritas dalam kumpulan informasi, SMOTE *initial* memilih *instance* data kelas minoritas secara acak. Kemudian, k tetangga terdekat Ma, berkaitan dengan kelas minoritas, ID. Contoh data kedua Mb kemudian dipilih dari k set tetangga terdekat. selama cara ini, Ma dan Mb terhubung, membentuk fase garis di ruang fitur. Data buatan baru kemudian dihasilkan sebagai kombinasi *gibbous* Ma dan Mb. Prosedur ini terjadi sampai *dataset* seimbang antara kelas minoritas dan mayoritas. berkat keefektifan SMOTE, perluasan dari banyak teknik pengambilan sampel yang dibuat secara luas [14].

2.8. SMOTE-Tomek

SMOTE dengan Tomek Link menyeimbangkan pengetahuan dan membangun instan kategori tambahan yang terpisah dengan baik. selama pendekatan ini, setiap *instance* data yang menghasilkan tautan yang dibuang Tomek, baik dari kelas minoritas atau mayoritas [14]. Apakah kumpulan sampel a dan b berpasangan dari kategori kumpulan data yang sama sekali berbeda sehingga sampel tidak ada c di mana pun d(a, c) < d(a, b) atau d2(b, c) < d(a, b), d adalah ruang antara pasangan sampel. Setelah 2 sampel (a dan b) mengetik TL, masing-masing sampel dihapus [13].

2.9. Evaluasi Model

Model Evaluasi yang bisa digunakan untuk mengukur nilai optimal dari berbagai parameter dari setiap metode, prediksi yang berasal dari semua kombinasi parameter yang mungkin dievaluasi pada dataset klasifikasi digunakan 5 nilai yaitu: *accuracy, precision, recall, f1-score* dan *AUC/ROC.* Adapun penjelasan dan rumus-rumus dalam perhitungannya sebagai berikut [16].*Accuracy*

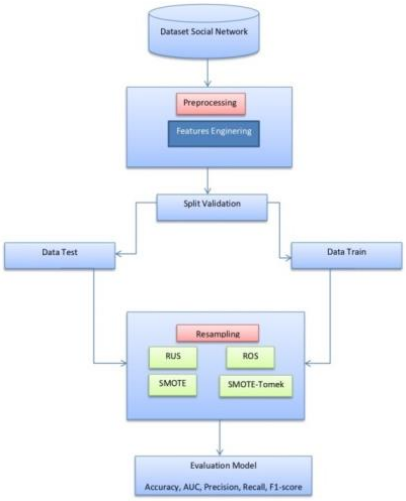
Accuracy adalah metrik evaluasi yang paling umum digunakan untuk klasifikasi. Namun, untuk masalah klasifikasi data yang tidak seimbang, akurasi mungkin bukan pilihan yang baik karena akurasi sering kali memiliki bias terhadap kelas mayoritas [17].

$$Accuracy = \frac{(TP+TN)}{Totalsample}\ x\ 100 \text{..... (1)}$$

1. *Precision*
   Precision adalah bagian dari data yang diambil sesuai dengan informasi yang diperlukan [18].
   $$Precision = \frac{(TP)}{(TP+FP)}\ x\ 100\% \text{ ..... (2)}$$
2. *Recall*
   Recall adalah pengumpulan data yang berhasil diambil dari bagian data yang relevan dengan queri [18].
   $$Recall = \frac{(TP)}{(TP+FN)}\ x\ 100\% \text{ ..... (3)}$$
3. *AUC/ROC*
   Kurva ROC (*Receiver Operating Characteristics*) diakui sebagai pilihan paling rasional untuk data yang tidak seimbang, yang menggambarkan pertukaran relatif antara manfaat dan biaya (Fawcett). Kurva ROC dihasilkan berdasarkan matriks dasar dalam *machine learning* yang disebut matriks kebingungan adalah matriks kebingungan dari masalah klasifikasi biner [17].
   Rumus menghitung AUC/ROC adalah:
   $$AUC = 1/2 \sum_{ki=1}^{n} (\text{xi} + 1 - \text{xi})(\text{yi} + 1 - \text{yi}) \text{...(4)}$$
4. *F-1 Score*
   Menggabungkan *precision* dan *recall* menjadi satu ukuran. Secara matematis *F-1 score* merupakan rata-rata harmonik dari *precision* dan *recall*. Untuk menghitung *F-1 score binary classification* [19] .
   Rumus menghitung *F-1 score* adalah:
   $$F - 1 = 2\frac{(Precision.Recall)}{(Precision+Recall)} \text{ ....(5)}$$

3. METODE PENELITIAN

Berikut gambaran tahapan penelitian yang dilakukan



Gambar 2. Tahapan Penelitian

3.1. Dataset

Social Network ads adalah kumpulan data kecil terdiri dari 400 instance berbeda dan pelanggan dalam kumpulan data terdiri dari wanita dengan usia minimum 18 tahun hingga usia maksimum 60 tahun. Kumpulan data terdiri dari pria dengan rentang usia yang sama dengan wanita. Gaji di sana berkisar antara 15000 hingga 150000. Variabel target biner mengambil nilai (0 atau 1), sementara '0' menyiratkan hasil negatif bagi pelanggan yang tidak membeli produk, dan '1' menunjukkan hasil positif. Kelima fitur tersebut dipilih untuk memprediksi jumlah pelanggan yang membeli produk baru, variabel signifikan tersebut digunakan untuk prediksi menggunakan classifier [3]. Berikut Tabel 2.2 Atribut *dataset Social Network.*

Tabel 1. Atribut *Dataset Social Network*

Nama Atribut	Tipe	Keterangan
User ID	*Float64*	ID konsumen yang melihat atau mengkik iklan
Age	*Int64*	usia konsumen
Gender	*Int64*	apakah konsumen laki-laki atau perempuan
Estimated Salary	*objek*	perkiraan gaji konsumen yang melihat atau mengklik Rasio klik tayang
Purchased	*Int64*	0 = tidak mengklik iklan dan 1= mengklik iklan

Tabel 2. *Dataset Social Network*

User ID	***Gender***	***Age***	***EstimatedSalary***	***Purchased***
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0

Data yang terdiri dari 400 data memiliki dua label *class* yaitu tidak mengklik rasio klik tayang dan mengklik rasio klik tayang diantaranya 257 data yang tidak mengklik rasio klik tayang dan 143 data yang mengklik rasio klik tayang. Berikut adalah statistik distribusi kelas dapat dilihat pada Tabel 3.

Tabel 3. Distribusi Kelas *Dataset Social network*

Nama Kelas	**Jumlah Data**
Tidak mengklik iklan	257
Mengklik iklan	143

3.2. *Preprocessing*

Dilakukan *Feature Engineering* dalam penelitian ini. Tindakan mengekstraksi fitur dari data mentah dan mengubahnya menjadi format yang sesuai untuk model pembelajaran mesin. Ini adalah langkah penting dalam alur pembelajaran mesin, karena fitur yang tepat dapat meringankan kesulitan pemodelan, dan oleh karena itu memungkinkan saluran untuk menghasilkan hasil dengan kualitas yang lebih tinggi [20].

3.3. *Split Validation*

Tahapan selanjutnya melakukan *split validation* yang bermaksud untuk melakukan pembagian data *training* dan *testing*, data *training* digunakan sebagai data pelatihan untuk pembelajaran model data *testing* digunakan sebagai data pengujian untuk validasi atau evaluasi model. Pada penelitian ini dilakukan *split* dengan menggunakan *split percentage* sebesar 80:20 yang merupakan pembagian data *training* sebanyak 80% dan data *testing* 20%.

3.4. *Resampling*

Setelah melihat statistik dari distribusi kelas data yang tidak seimbang, penulis selanjutnya melakukan eksperimen dengan menambahkan *resampling.* Dilakukan pembandingan antara *Random Under Sampling* (RUS), *Random Over Sampling* (ROS), *Synthetic Minority Over Sampling Technique* (SMOTE), dan SMOTE-Tomek.

3.5. Evaluasi Model

Evaluasi model digunakan untuk mengetahui kinerja dari hasil uji model algoritma LightGBM dengan *resampling.* Digunakan *Accuracy*, *Precision*, *Recall*, AUC dam F1-Score untuk mengevaluasi kinerja model.

4. HASIL DAN PEMBAHASAN

4.1. Analisis Statistika Dataset

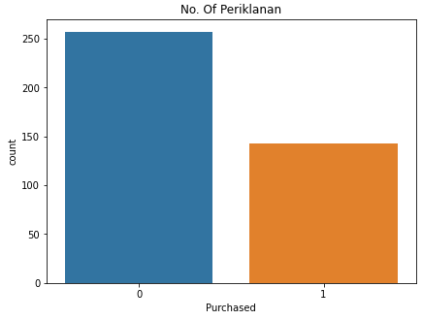
Penelitian ini membahas tentang prediksi rasio klik tayang pada iklan internet atau pemasaran *online*. Data yang digunakan diperoleh dari data publik di *Kaggle.* Dataset *social network* yang memiliki 5 fitur dengan jumlah 400 *instance*. Untuk mengetahui statistika deskriptif dataset *social network* akan dijelaskan pada Tabel 4.

Tabel 4. Nilai Statistika *Dataset Social Network*

	User ID	**Gender**	**Age**	**Estimated Salary**	**Purchased**
count	400.00	400.00	400.00	400.00	400.00
mean	15691539.76	0.49	37.66	69742.50	0.36
std	71658.32	0.50	10.48	34096.96	0.48
min	15566689.00	0.00	18.00	15000.00	0.00
25%	15626763.75	0.00	29.75	43000.00	0.00
50%	15694341.50	0.00	37.00	70000.00	0.00
75%	15750363.00	1.00	46.00	88000.00	1.00
max	15815236.00	1.00	60.00	150000.00	1.00

Statitik deskriptif dari set data tersebut, Tabel 4 memberikan informasi pada kolom *count* dapat dilihat jumlah data yang digunakan pada penelitian ini, jumlah total data yang digunakan pada penelitian ini adalah sebanyak 400 *instance*, *mean* adalah nilai tengah dari atribut, std adalah standard deviasi atau simpangan dari dataset, *min* adalah nilai paling kecil

dari dataset, 25%, 50%, 75% menginformasikan bahwa jumlah dataset pada quartil 1, quartil 2 dan quartil 3 dan nilai *max* adalah nilai paling tinggi dari dataset penelitian. Sedangkan distribusi kelas *dataset social network* Gambar 3.



Gambar 3. Distribusi Kelas Dataset *Social Network*

Berdasarkan distribusi kelas pada Gambar 5 dapat dilihat bahwa terjadinya *imbalance* pada kelas jumlah pelanggan yang tidak mengklik (0) rasio klik tayang sebanyak 257 dan yang mengklik (1) rasio klik tayang sebanyak 143.

4.2. *Preprocessing*

Preprocessing yang dilakukan dalam penelitian ada beberapa tahapan diantaranya menghapus atribut yang tidak dibutuhkan, penambahan atribut yang baru, *feature engineering dan feature correlation.*

4.3. *Feature Engineering*

Dilakukan transformasi data pada fitur *Gender*. Berikut data contoh data sebelum dilakukan transformasi data pada Tabel 5 dan pada Tabel 6 setelah dilakukan trasformasi.

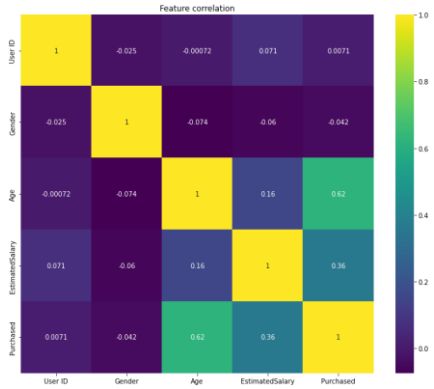
Tabel 5. Fitur *Gender*

Gender
male
male
female
female
male

Tabel 6. Fitur *Gender* Setelah Transformasi

Gender
1
1
0
0
1

4.4. *Feature Corelation*



Gambar 4. *Feature Correlation*

Berdasarkan *feature correlation* pada data *social network* diatas dapat dilihat korelasi yang paling tinggi dengan *class Purchased* adalah atribut atau *feature Age* dengan nilai korelasi 0.62 sedangkan korelasi yang paling rendah dengan class *Purchased* adalah atribut atau *feature gender* dengan nilai korelasi -0.042. Dimana ketika nilai korelasi mendekati angka 1 maka hubungan antar atribut sangat tinggi

4.5. Hasil Eksperimen dan Analisis

Setelah dilakukan tahapan penelitian, berikut hasil dari pengujian dengan *Light Gradient Boosting*:

Tabel 7. Hasil Pengujian

Metode	Accuracy	Recall	Precision	F1-Score
Tanpa Resampling	91.25%	93.10%	84.38%	88.52%
ROS	91.25%	93.10%	84.38%	88.52%
SMOTE	90%	89.66%	83.87%	86.67%
RUS	90%	89.66%	83.87%	86.67%
SMOTE-Tomek	85%	86.20%	75.75%	80.64%

Berdasarkan hasil pengujian hasil tanpa menggunakan metode resampling sama dengan hasil menggunakan metode resampling ROS. Dengan hasil accuracy 91.25%, recall 93.10%, precision 84.38% dan F1-Score 88.52%. Sedangkan hasil AUC pada tabel 9.

Tabel 8. Hasil AUC

Metode	AUC
Tanpa Resampling	0,92
ROS	0,92
SMOTE	0,90
RUS	0,90
SMOTE-Tomek	0,85

Berdasarkan hasil pengujian, AUC tanpa menggunakan metode resampling sama dengan hasil menggunakan metode resampling ROS sebesar 0,92. Hasil ini termasuk dalam *excellent classification.*

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa dalam prediksi rasio klik tayang menggunakan *dataset sosial network* menggunakan *light gradient boosting*, memperoleh hasil terbaik *accuracy* 91.25%, *recall* 93.10%, *precision* 84.38% *F1-Score* 88.52% dan *AUC* 0,92. Hasil ini diperoleh dari pengujian tanpa *resampling* dan menggunakan *resampling* ROS. Penggunaan resamping dalam pengujian *dataset sosial network* menggunakan *light gradient boosting* tidak menghasilkan performa yang berbeda. Sehingga penelitian selanjutnya dapat dilakukan dengan perbandingan metode lain dan melakukan

prerocessing data yang belum dilakukan pada penelitian ini.

DAFTAR PUSTAKA

[1] S. Saraswathi, V. Krishnamurthy, D. Venkata Vara Prasad, R. K. Tarun, S. Abhinav, and D. Rushitaa, "Machine learning based ad-click prediction system," *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 3646–3648, 2019, doi: 10.35940/ijeat.F9366.088619.

[2] Y. Xie, D. Jiang, X. Wang, and R. Xu, "Robust transfer integrated locally kernel embedding for click-through rate prediction," *Inf. Sci. (Ny).*, vol. 491, pp. 190–203, 2019, doi: 10.1016/j.ins.2019.04.006.

[3] P. Saha, "Performance analysis of the Machine Learning Classifiers to predict the behaviour of the customers, when a new product is launched in the market," vol. 5, no. 3, pp. 1907–1911, 2019.

[4] J. Mahindra Bagul and T. B. Kute, "Ad-Click Prediction using Prediction Algorithm: Machine Learning Approach," *Int. Res. J. Eng. Technol.*, 2020.

[5] K. K. Pandey and D. Shukla, "Challenges of big data to big data mining with their processing framework," *Proc. - 2018 8th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2018*, pp. 89–94, 2018, doi: 10.1109/CSNT.2018.8820282.

[6] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," *Edik Inform.*, vol. 2, no. 2, pp. 213–219, 2017, doi: 10.22202/ei.2016.v2i2.1465.

[7] B. Xia, X. Wang, T. Yamasaki, K. Aizawa, and H. Seshime, "Deep neural network-based click-through rate prediction using multimodal features of online banners," *Proc. - 2019 IEEE 5th Int. Conf. Multimed. Big Data, BigMM 2019*, pp. 162–170, 2019, doi: 10.1109/BigMM.2019.00-29.

[8] D. Jiang, R. Xu, X. Xu, and Y. Xie, "Multi-view feature transfer for click-through rate prediction," *Inf. Sci. (Ny).*, vol. 546, pp. 961–976, 2021, doi: 10.1016/j.ins.2020.09.005.

[9] Y. Ju, G. Sun, Q. Chen, M. Zhang, H. Zhu, and M. U. Rehman, "A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting," *IEEE Access*, vol. 7, no. c, pp. 28309–28318, 2019, doi: 10.1109/ACCESS.2019.2901920.

[10] Y. Su, "Prediction of air quality based on Gradient Boosting Machine Method," *Proc. - 2020 Int. Conf. Big Data Informatiz. Educ. ICBDIE 2020*, pp. 395–397, 2020, doi: 10.1109/ICBDIE50010.2020.00099.

[11] S. P. Singh, P. Singh, and A. Mishra, "Predicting Potential Applicants for any Private College using LightGBM," *2020 Int. Conf. Innov. Trends Inf. Technol. ICITIIT 2020*, 2020, doi: 10.1109/ICITIIT49094.2020.9071525.

[12] H. Hudori, "Resampling Neural Network Untuk Penanganan Class Imbalance Pada Prediksi Klaim Asuransi," *Teknois J. Ilm. Teknol. Inf. dan Sains*, vol. 10, no. 1, pp. 57–64, 2020, doi: 10.36350/jbs.v10i1.78.

[13] E. Rendón, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutiérrez, "Data sampling methods to dealwith the big data multi-class imbalance problem," *Appl. Sci.*, vol. 10, no. 4, 2020, doi: 10.3390/app10041276.

[14] M. Dorn *et al.*, "Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets," *PeerJ Comput. Sci.*, vol. 7, p. e670, 2021, doi: 10.7717/peerj-cs.670.

[15] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," *2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020*, no. May, pp. 243–248, 2020, doi: 10.1109/ICICS49469.2020.239556.

[16] N. W. Wardani, "Penerapan Data Mining Dalam Analytic CRM (Studi Kasus: Prediksi Penlanggan Churn di Perusahaan Retail dengan Algoritma Naive Bayes Berdasarkan Segmentasi Pelanggan)," *Yayasan Kita Menulis*, 2020.

[17] G. Haixiang, L. Yijing, L. Yanan, L. Xiao, and L. Jinling, "BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification," *Eng. Appl. Artif. Intell.*, vol. 49, pp. 176–193, 2016, doi: 10.1016/j.engappai.2015.09.011.

[18] H. M. Nawawi, S. Rahayu, J. J. Purnama, and S. I. Komputer, "Algoritma c4.5 untuk memprediksi pengambilan keputusan memilih deposito berjangka," vol. 16, no. 1, pp. 65–72, 2019.

[19] J. Mohajon, "Confusion Matrix for Your Multi-Class Machine Learning Model," *Towar. data Sci.*, 2020.

[20] A. Zheng and A. Casari, "Feature Engineering for Mchine Learning," *Reilly Media, Gravenstein Highw. North, Sebastopol.*, 2018.