	JTKSI (Jurnal Teknologi Komputer dan Sistem Informasi) JTKSI, Volume 6, Nomor 3, September 2023 E ISSN: 2620-3030; P ISSN: 2620-3022, pp.252-256 Accredited SINTA 4 Nomor 200/M/KPT/2020 http://ojs.stmikpringsewu.ac.id/index.php/jtksi Received: 26 Juli 2023 Revised: 28 Agustus 2023; Accepted: 5 September 2023
---	--

Analisis Statistik Pembayaran Kredit Menggunakan Metode Random Forrest Classifier Dan LightGBM

Akni Widiyastuti¹, Chairani², Kurniawati³, Siti Aminah⁴, Muhamad Abror⁵

^{1,3,4}Fakultas Ilmu Komputer, Institut Teknologi Bisnis dan Bahasa Dian Cipta Cendikia, Lampung

^{2,5}Fakultas Ilmu Komputer, Informatics & Business Institute Darmajaya, Lampung

^{1,3,4}Jl. Cut Nyak Dien No. 65 Durian Payung Palapa Bandar Lampung, Lampung, Indonesia

^{2,5}Jl. Z.A. Pagar Alam No. 93, Bandar Lampung, Indonesia

e-mail : id.akni.widiyastuti@gmail.com¹, xkurniawatix@gmail.com², chairani@darmajaya.ac.id³, sitiaminah@gmail.com⁴, m.abror@gmail.com⁵

Abstract

Banking also takes part in marketing technology based on financial services such as the following: Banking, Credit and Loans, Financial Advisory, and Consumer Convenience. By automating most of the banking processes and adding speed, transparency, and security to them, Fintech applications also indirectly contribute to a seamless experience for customers. It is known that one of the banks in Indonesia has also implemented a fintech system, one of which is in terms of processing credit payments (online credit). Some people have applied for credit at the bank by making payments digitally or online. Banks have collected data from their customers who have made credit online. These data are collected to obtain information that helps banks see the status of credit payments. In the research results, the accuracy value was obtained by using the Random Forrest Classifier and LightGBM method comparisons, the values obtained from the AUC for the validation set were around 65% and 80%.

Keywords: Credit, Random Florest Classifier, LightGBM

Abstrak

Perbankan juga mengambil bagian dalam teknologi pemasaran berbasis pada layanan keuangan seperti berikut ini: Perbankan, Kredit dan Pinjaman, Penasihat Keuangan, dan Kenyamanan Konsumen. Dengan mengotomatiskan sebagian besar proses perbankan dan menambahkan kecepatan, transparansi, dan keamanan pada proses-proses tersebut, aplikasi Fintech juga membantu secara tidak langsung pengalaman yang memudahkan untuk pelanggan. diketahui bahwa salah satu bank di Indonesia juga telah menerapkan sistem fintech, salah satunya adalah dalam hal proses pembayaran kredit (kredit online). Beberapa orang telah menerapkan kredit di bank dengan melakukan pembayaran secara digital atau online. Bank telah mengumpulkan data dari pelanggan mereka yang telah melakukan kredit online. Data-data ini dikumpulkan untuk mendapatkan informasi yang membantu bank melihat status pembayaran kredit. Pada hasil penelitian di peroleh nilai akurasi dengan menggunakan Dengan komparasi Metode Random Forrest Classifier dan LightGBM nilai yang diperoleh dari AUC untuk set validasi sekitar 65% dan 80%.

Kata Kunci: Credit, Random Florest Classifier, LightGBM

I. PENDAHULUAN

Gagasan utama fintech adalah membuat layanan keuangan lebih baik bagi konsumen, bukan untuk mengeluarkan bank dari bisnis. Perbankan juga mengambil bagian dalam teknologi pemasaran berbasis pada layanan keuangan seperti berikut ini: Perbankan, Kredit dan Pinjaman, Penasihat Keuangan, dan Kenyamanan Konsumen. Dengan mengotomatiskan sebagian besar proses perbankan dan menambahkan kecepatan, transparansi, dan keamanan pada proses-proses tersebut, aplikasi Fintech juga membantu secara tidak langsung pengalaman yang memudahkan untuk

pelanggan. Dari penjelasan di atas, diketahui bahwa salah satu bank di Indonesia juga telah menerapkan sistem fintech, salah satunya adalah dalam hal proses pembayaran kredit (kredit online). Beberapa orang telah menerapkan kredit di bank dengan melakukan pembayaran secara digital atau online. Bank telah mengumpulkan data dari pelanggan mereka yang telah melakukan kredit online. Data-data ini dikumpulkan untuk mendapatkan informasi yang membantu bank melihat status pembayaran kredit.

Data yang diberikan sekitar 24000 pelanggan dengan karakteristik berbeda, di mana 15945 di

antaranya tidak pernah melewati pembayaran dan 8055 di antaranya setidaknya melewati pembayaran sekali. Sebagai Peneliti, saya untuk memprediksi 6000 pelanggan baru dengan jenis karakteristik tertentu yang diberikan pada Tabel apakah mereka melewati pembayaran setidaknya satu kali selama periode enam bulan.

Secara garis besar penelitian ini mencakup beberapa pembahasan sebagai berikut: Memilih variabel yang menjadi acuan untuk melakukan analisis dan memberikan justifikasi. Menentukan metodologi yang digunakan untuk analisis dan menjelaskan alasan mengapa memilih metode tersebut (disebutkan dan jelaskan teorinya secara singkat). Memperlihatkan hasil eksplorasi data dan menjelaskan hasil analisis. Diakhir pembahasan diberikan saran berdasarkan hasil penelitian.

Metode penggunaan machine learning adalah salah satu metode yang dapat digunakan untuk mengolah data yang bertujuan untuk membuat suatu keputusan maupun prediksi berdasarkan data-data yang ada sehingga dapat diterapkan dalam penelitian ini. Machine learning juga sudah dapat digunakan dan diimplementasikan di banyak aplikasi dan perangkat lunak yang berguna untuk membantu manusia dalam tugas atau peran yang dikerjakan oleh mesin dan lainnya [4]. Dalam penelitian yang akan dilakukan pada penelitian ini akan menggunakan random forest classifier pada penerapan machine learning dimana pada random forest classifier yang merupakan pemodelan klasifikasi untuk menentukan hasil biner berdasarkan rangkaian variabel independen. Menurut Undang-Undang Perbankan No. 10 Tahun 1998, pinjaman adalah uang atau surat promes yang mewajibkan peminjam untuk melunasi utangnya setelah jangka waktu tertentu berdasarkan perjanjian atau kesepakatan pinjaman antara bank dengan pihak lain yang bunganya meningkat dan sebanding (Wikisource, 2014)

II. METODE PENELITIAN

a. Random Forest Classifier Method

Random Forest adalah algoritma machine learning yang menggabungkan keluaran dari beberapa decision tree untuk mencapai satu hasil. Sesuai namanya, Forest atau 'hutan' dibentuk dari banyak tree (pohon) yang diperoleh melalui proses bagging atau bootstrap aggregating. Setiap tree pada Random Forest akan mengeluarkan prediksi kelas. Prediksi kelas dengan vote terbanyak menjadi kandidat prediksi pada model. Semakin banyak jumlah tree maka akan menghasilkan akurasi yang lebih tinggi dan mencegah masalah overfitting. Algoritma Random Forest diperkenalkan oleh Leo Breiman dan Adele Cutler. Algoritma ini didasarkan pada konsep ensemble learning, yakni proses menggabungkan beberapa pengklasifikasi untuk memecahkan masalah yang kompleks dan untuk meningkatkan kinerja model.

b. Logistic Regression

LightGBM algoritma yang dirancang oleh

Microsoft Research Asia menggunakan kerangka Gradient Boosting Decision Tree (GBDT) (Ke dkk, 2017). Tujuannya untuk meningkatkan efisiensi komputasi, sehingga masalah prediksi dengan big data dapat diselesaikan dengan efisien (Liang dkk, 2020). LightGBM memiliki beberapa keunggulan dibandingkan metode GBDT lainnya, yaitu kecepatan pelatihan lebih cepat, efisiensi lebih tinggi, penggunaan memori lebih rendah, tingkat akurasi lebih baik, kemampuan dalam menangani data dengan skala yang besar dan dukungan pembelajaran paralel dan GPU (Rufo dkk, 2021). LightGBM adalah kerangka Gradient Boosting yang cepat, terdistribusi dan berkinerja tinggi berdasarkan algoritma pohon keputusan yang dapat digunakan untuk peringkat, klasifikasi, regresi dan banyak tugas pembelajaran mesin lainnya (Rufo dkk, 2021).

c. Confusion Matrix

Confusion Matrix adalah sebuah metode yang biasa digunakan untuk perhitungan akurasi, *recall*, *precision*, dan *error rate*. Dimana, *precision* mengevaluasi kemampuan sistem untuk menemukan peringkat yang paling relevan, dan didefinisikan sebagai persentase dokumen yang di *retrieve* dan benar-benar relevan terhadap *query*. *Recall* mengevaluasi kemampuan sistem untuk menemukan semua item yang relevan dari koleksi dokumen dan didefinisikan sebagai persentase dokumen yang relevan terhadap *query*. *Accuracy* merupakan perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus dan *error rate* merupakan kasus yang diidentifikasi salah dengan jumlah seluruh kasus [16]. Berikut adalah cara pengukuran performa confusion matrix

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 - Score = \frac{2TP}{2TP + FP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

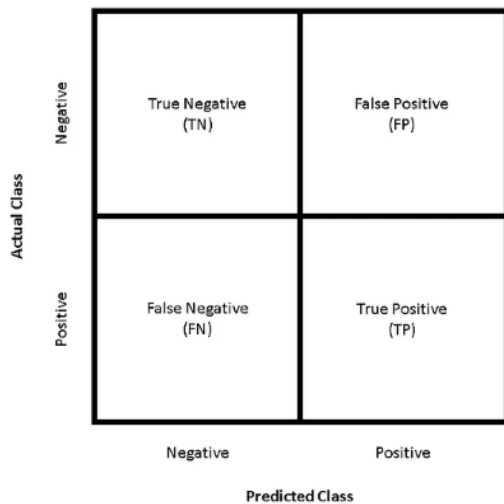
d. K-fold Cross Validation

K-Fold Cross Validation merupakan proses validasi model yang dilakukan dengan cara membagi dataset menjadi k bagian atau fold, dan dilakukan iterasi sebanyak k kali. Pada setiap iterasi, setiap bagian atau fold digunakan sebagai testing dataset sebanyak satu kali secara bergantian. Bagian yang lainnya k-1 fold digunakan sebagai training dataset. Hal ini bertujuan untuk melakukan testing terhadap model menggunakan data yang belum pernah dilihat sebelumnya. Untuk penggunaan jumlah fold terbaik untuk uji validitas, dianjurkan menggunakan 10-fold cross validation dalam model machine learning [17].

e. ROC AUC

Grafik ROC adalah grafik dua dimensi hubungan

antara True Positive Rate (TPR) atau Sensitivity (sumbu Y) dengan False Positive Rate (FPR) atau 1-Specificity (sumbu X). Beberapa hal penting yang perlu diperhatikan dalam grafik ROC adalah jika grafik menunjukkan titik kiri terendah (0,0), maka mewakili nilai probabilitas yang tidak pernah menunjukkan kondisi positif, yang artinya klasifikasi tidak menghasilkan kondisi false positive dan true positive. Kemudian terdapat *Accuracy* yang merupakan perbandingan kasus yang diidentifikasi benar dengan jumlah seluruh kasus dan *error rate* merupakan kasus yang diidentifikasi salah dengan jumlah seluruh kasus[8]. Berikut adalah pengukuran performa ROC AUC



Kurva ROC melakukan ini dengan memplot sensitivitas, probabilitas memprediksi positif nyata akan menjadi positif, terhadap spesifisitas 1, probabilitas memprediksi negatif nyata akan menjadi positif. Dalam tahapan pemrosesan algoritma setelah dilakukan preprocessing pada dataset yang telah dilakukan dan juga pemodelan data sehingga nantinya data dapat dilakukan pemrosesan algoritma dengan menggunakan logistic regression. *Model training* diawali dengan melatih data sebelum dilakukan bosting data oleh adaboost, kemudian akan dilakukan inisiasi kernel untuk pemrosesan Logistic Regression sehingga menghasilkan model dari data latih yang kemudian model tersebut di uji berdasarkan data uji yang disiapkan kemudian akan dilakukan cross validation untuk menguji nilai akurasi menggunakan ROC AUC. Sehingga nantinya dapat dikategorikan fake account dan real account berdasarkan nilai 0 dan 1 berdasarkan dataset yang telah disediakan. Berikut adalah alur pemrosesan algoritma.

III. PEMBAHASAN

Pada hasil pembahasan digunakan metode Machine Learning dengan algoritma logistic regression dengan pengumpulan dataset yang diolah menjadi example set untuk dilakukan pemrosesan menggunakan algoritma logistic regression. Berikut adalah contoh example set pada dataset fake account instagram.

profile_pic	num_posts	num_followers	num_following	num_likes	num_comments	num_reposts	num_saved	num_blocked	num_hidden	num_archived	num_deleted	num_hidden_posts	num_hidden_comments	num_hidden_reposts	num_hidden_saved	num_hidden_blocked	num_hidden_hidden	num_hidden_archived	num_hidden_deleted
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Gambar 2. Exaple set dataset fake account

Pada pengolahan data fake account Instagram ini untuk memprediksi apakah akun pengguna itu palsu atau tidak. Pada prediksi akun Instagram menggunakan kategori biner dalam dataset dimana palsu dengan menggunakan angka 1 dan 0 menunjukan akun asli. Pada pengolahan awal terdapat variable yang akan digunakan dalam penelitian ini yang terdapat didalam dataset. Berikut adalah variable dalam dataset tersebut.

```

Data columns (total 12 columns):
#   Column              Non-Null Count  Dtype
---  -
0   fake                 576 non-null    int64
1   profile_pic          576 non-null    object
2   ratio_numlen_username 576 non-null    float64
3   len_fullname         576 non-null    int64
4   ratio_numlen_fullname 576 non-null    float64
5   sim_name_username    576 non-null    object
6   len_desc             576 non-null    int64
7   extern_url           576 non-null    object
8   private              576 non-null    object
9   num_posts            576 non-null    int64
10  num_followers         576 non-null    int64
11  num_following         576 non-null    int64
dtypes: float64(2), int64(6), object(4)

```

Gambar 3. variable dan Type Data

Dari data variable diatas terdapat tiga tipe data yang berbeda diantaranya type float dengan 2 data, integer dengan 6 data dan object dengan 4 data yang akan di olah. Kemudian terdapat deskripsi dari masing masing variable yang digunakan. Berikut adalah deskripsi masing-masing variable.

No.	Column name	Data Type	Description
0	'profile_pic'	categorical	Apakah akun tersebut memiliki foto profil (1) atau tidak (0)
1	'ratio_numlen_username'	numerical (float)	Rasio karakter numerik dalam nama pengguna akun dengan panjangnya
2	'len_fullname'	numerical (int)	Jumlah karakter dalam nama lengkap pengguna
3	'ratio_numlen_fullname'	numerical (float)	Rasio karakter numerik dalam nama lengkap pengguna terhadap panjangnya
4	'sim_name_username'	categorical	Apakah nama pengguna benar-benar cocok dengan nama pengunanya (1), tidak sama sekali (0).
5	'len_desc'	numerical (int)	Jumlah karakter dalam deskripsi akun
6	'extern_url'	categorical	Apakah deskripsi akun menyertakan URL (1) atau tidak (2)
7	'private'	categorical	Apakah postingan pengguna hanya dapat dilihat oleh pengikutnya (1) atau oleh semua pengguna Instagram (2).
8	'num_posts'	numerical (int)	Jumlah posting di akun pengguna
9	'num_followers'	numerical (int)	Jumlah pengguna Instagram yang mengikuti akun tersebut
10	'num_following'	numerical (int)	Jumlah pengguna Instagram yang diikuti oleh akun tersebut
11	'fake'	categorical	Apakah akun pengguna itu asli (0) atau akun palsu (1).

Gambar 4. Deskripsi variable

Setelah ternariskan variable yang akan digunakan maka selanjutnya dilakukan pengecekan pada statistic data pada data yang akan dilatih agar tidak ada data

missing yang terdapat pada dataset sebelum dilakukan pemodelan pada algoritma.

	ratio_numlen_username	len_fullname	ratio_numlen_fullname	len_desc	num_posts	num_followers	num_following
count	576.00	576.00	576.00	576.00	576.00	576.00	576.00
mean	0.16	1.46	0.04	22.62	107.49	85307.24	508.38
std	0.21	1.05	0.13	37.70	402.03	910148.46	917.98
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	0.00	1.00	0.00	0.00	0.00	39.00	57.50
50%	0.00	1.00	0.00	0.00	9.00	150.50	229.50
75%	0.31	2.00	0.00	34.00	81.50	716.00	589.50
max	0.92	12.00	1.00	150.00	7389.00	15338538.00	7500.00

	percent_missing (%)
fake	0.00
profile_pic	0.00
ratio_numlen_username	0.00
len_fullname	0.00
ratio_numlen_fullname	0.00
sim_name_username	0.00
len_desc	0.00
extern_url	0.00
private	0.00
num_posts	0.00
num_followers	0.00
num_following	0.00

Gambar 5. Pengecekan deskripsi statistik dan missing data

Setelah pengecekan deskripsi data juga terapat hasil dari pengecekan data missing pada setiap variable dan dipastikan bahwa setiap variable tidak ada data hilang sehingga dapat diolah untuk dilakukan pemodelan pada algoritma. Pada tahapan ini dilakukan pelabelan pada dataset dengan pembelajaran learning. Dimana fungsi label encoding diwakili oleh string ke fitur biner yang hanya berisi 0 dan 1 yang dapat digunakan untuk pemodelan machine learning. Berikut adalah hasil pelabelannya sesuai dengan dataset.

fake	profile_pic	ratio_numlen_username	len_fullname	ratio_numlen_fullname	len_desc	extern_url	private	num_posts	num_followers	num_following
0	0	1	0.27	0	0.00	53	0	32	1000	995
1	0	1	0.00	2	0.00	44	0	0	296	533
2	0	1	0.10	2	0.00	0	0	1	13	159
3	0	1	0.00	1	0.00	82	0	0	679	414
4	0	1	0.00	2	0.00	0	0	1	6	151

Gambar 6. Pelabelan encoding dengan machine learning

Setelah dilakukan pelabelan model dilakukan pengimporan logistic regression model dengan hasil tanpa regularisasi dan dengan regularisasi. Pada pemodelan ini menggunakan target pada value dari pelabelan model diatas dimana akan dilakukan testing dengan dan tanpa regularisasi. Dan berikut adalah hasil presisi dan recallnya.

```

Presisi model tanpa regularisasi: 0.8666666666666667
Recall model tanpa regularisasi: 0.8666666666666667
Presisi model dengan regularisasi: 0.881355932203898
Recall model dengan regularisasi: 0.8666666666666667

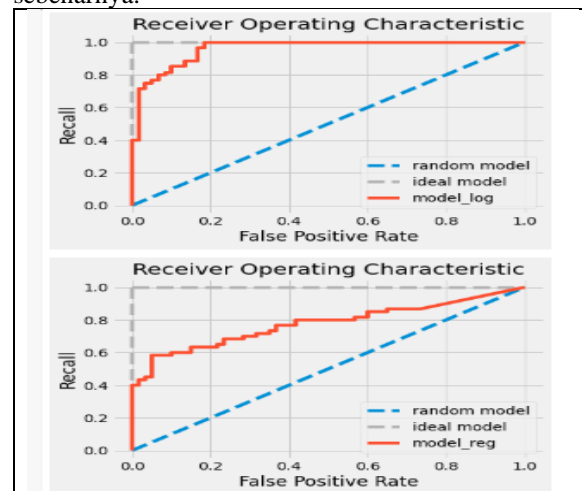
```

Gambar 7. Hasil presisi dan akurasi feature matrix and target vector

Pada hasil diatas dapat terlihat bahwa pemodelan dengan regularisasi untuk mengenali akun palsu lebih tinggi daripada pengenalan dengan pemodelan tanpa

regularisasi. Pada hasil presisi dengan menggunakan regularisasi sebesar 0,88 dan recall sebesar 0,88.

Pada logistic regression menggunakan ambang batas 0,5 (50%) untuk menentukan kategori yang di prediksi. Jika menurunkan mabang batas maka akan lebih banyak prediksi positif dengan demikian recall akan meningkat. Dalam penelitian ini dengan menggunakan ROC (receiver operating characteristic) lebih bagus dalam melihat bagaimana pengklasifikasi prediktif seperti model regresi logistic dapat membedakan antara positif dan negative yang sebenarnya.



Gambar 8. Receiver Operating Characteristic

Dapat dilihat pada hasil gambar diatas dimana pemodelan pada variable model_log lebih baik dengan menggunakan regularisasi daripada pemodelan dengan variable model_reg dengan non regularisasi. Pada gambar dengan variable model_log kurva merah menunjukkan hasil yang baik dengan kurva yang tidak menyimpang pada ideal modelnya. Berbeda dengan model_reg dimana garis kurvanya keluar terlalu jauh dari ideal model.

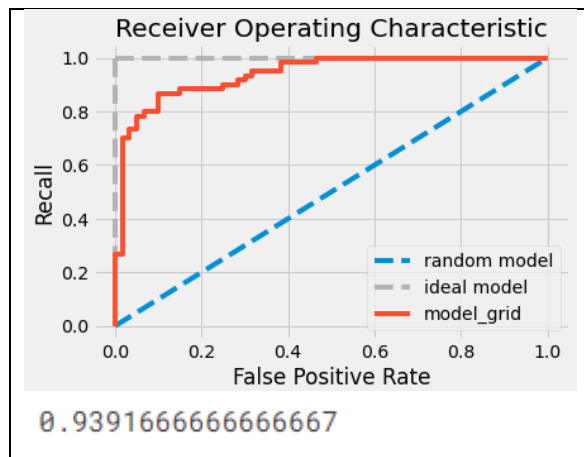
```

roc_auc_score untuk model tanpa regularisasi 0.9613888888888888
-----
roc_auc_score untuk model dengan regularisasi 0.7763888888888889

```

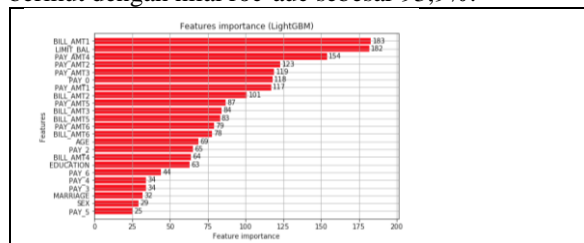
Gambar 9. ROC AUC Score

Pada hasil Roc Auc Score dapat dilihat perbandingan skor model tanpa regularisasi 96%, dengan model dengan regularisasi 77%. Setelah dilakukan validasi hasil pada model logistic regression dapat menghasilkan nilai roc auc score yang baik sebagai berikut.



Gambar 10. Hasil Nilai ROC-AUC

Nilai dari pengukuran ROC-AUC sebesar 93,9% menunjukkan hasil yang sangat baik. Dan pada hasil prediksi akun fake terdapat pada table berikut dengan nilai roc-auc sebesar 93,9%.



Gambar 11. Feature Important

Terdapat dua account Instagram yang modelnya absolut dengan kategorinya. Probabilitas palsu yang di prediksi adalah 1,0 dan 0,0. Akun pengguna di kategorikan menjadi asli dan palsu berdasarkan karakteristiknya.

IV. KESIMPULAN

Pada penelitian ini digunakan split validasi training dan validasi silang untuk mengevaluasi efektivitas model untuk memprediksi nilai target, yaitu mendeteksi apakah klien kartu kredit akan mengalami default bulan depan. Dengan kombinasi Metode Random Forrest Classifier dan LightGBM nilai yang diperoleh dari AUC untuk set validasi sekitar 0,80. Sehingga disimpulkan Metode Random Forrest Classifier dan LightGBM adalah model yang terbaik untuk melakukan prediksi apakah klien kartu kredit akan mengalami default bulan depan.

Jika Anda kesulitan dalam menemukan penelitian yang ingin anda lakukan, boleh menggunakan data yang ada pada penelitian sederhana ini. Tetapi anda harus menggunakan metode yang lainnya dan hasil prediksinya harus melebihi 80%.

DAFTAR PUSTAKA

- [1] H. Gunawan and G. S. Budhi, "Penerapan Machine Learning dalam mendeteksi Fake Account pada Instagram," 2022.
- [2] M. Aljabri, R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, and D. M. Alomari,

"Machine learning-based social media bot detection: a comprehensive literature review," *Soc. Netw. Anal. Min.*, vol. 13, no. 1, Dec. 2023, doi: 10.1007/s13278-022-01020-5.

- [3] J. Ezarfelix, N. Jeffrey, and N. Sari, "Systematic Literature Review: Instagram Fake Account Detection Based on Machine Learning," *Eng. Math. Comput. Sci. J.*, vol. 4, no. 1, pp. 25–31, Feb. 2022, doi: 10.21512/emacsjournal.v4i1.8076.
- [4] E. P. Meshram, R. Bhambulkar, P. Pokale, K. Kharbikar, and A. Awachat, "Automatic Detection of Fake Profile Using Machine Learning on Instagram," *Int. J. Sci. Res. Sci. Technol.*, pp. 117–127, May 2021, doi: 10.32628/ijrsrst218330.
- [5] A. N. Rohim and A. R. Pratama, "Analisis Sentimen Publik di Media Sosial Instagram atas Kinerja Presiden Joko Widodo," *Jurnal AUTOMATA*, vol. 3, no. 1, pp. 1–5, 2022.
- [6] N. G. Ramadhan, F. D. Adhinata, A. J. T. Segara, and D. P. Rakhmadani, "Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression," *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 2, p. 251, Apr. 2022, doi: 10.30865/jurikom.v9i2.3979.
- [7] A. A. Kurniawan and M. Mustikasari, "Implementasi Deep Learning Menggunakan Metode CNN dan LSTM untuk Menentukan Berita Palsu dalam Bahasa Indonesia," vol. 5, no. 4, pp. 2622–4615, 2020, doi: 10.32493/informatika.v5i4.7760.
- [8] W. Willy, D. P. Rini, and S. Samsuryadi, "Perbandingan Algoritma Random Forest Classifier, Support Vector Machine dan Logistic Regression Clasifier Pada Masalah High Dimension (Studi Kasus: Klasifikasi Fake News)," *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 4, p. 1720, Oct. 2021, doi: 10.30865/mib.v5i4.3177.