e-ISSN : 2830-6252
p-ISSN : 2830-6562

JCI, Vol. 4 No. 1 (2024) 20 – 28 | https://doi.org/10.54066/jci.v2i1.176

Jurnal Cakrawala Informasi

Journal Homepage: http://www.itbsemarang.ac.id/sijies/index.php/jci
e-Mail: jci@itbsemarang.ac.id

Optimasi Prediksi Pemasaran Nasabah Deposito Bank dengan Metode Klasifikasi Logistic Regression

Atika Mutiarachim [1*]
Jaluanto Sunu Punjul Tyoso [2]

[1, 2] Bisnis Digital, Universitas 17 Agustus 1945 Semarang

INFO ARTIKEL

Histori artikel:

Diterima : 6 Juni 2024
Revisi : 14 Juni 2024
Disetujui : 25 Juni 2024
Publikasi : 30 Juni 2024

Kata kunci:

Logistic Regression
Klasifikasi
Deposito
Performance

ABSTRACT

The study aims to determine the impact of the Logistic Regression *method on the classification of customer bank deposits, using a public UCI Bank Marketing dataset, which contains customer-specific information of bank deposit telemarketing activities. Data has a binomial label consisting of 'yes' for subscribers and 'no' for non-subscribers. The data preprocessing phase is done with downsampling to make the amount of data more symmetrical, then data selection and data transformation to ensure that the data used values are consistent, attribute selection to select the attributes most accurately used and give significant influence. Classification is done using the* Logistic Regression *algorithm. Data is shared using a split method with 90% training data and 10% testing data, with the aim of optimizing the training process. The performance result consists of an accuracy of 88.53%, a classification error value of 11.4%, can be categorized as low, showing only a few errors produced by the algorithm model, a kappa value of 0.68 close to 1, so it is categorized well, a low RMSE rating of 0.3 indicates a model accurate, and a high AUC percentage of 93.4% indicates the correct algority used in this dataset, because it produces a good performance value.*

ABSTRAK

Penelitian ini bertujuan mengetahui pengaruh metode Logistic Regression pada klasifikasi nasabah deposito bank, menggunakan dataset publik UCI Bank Marketing, yang berisi informasi karakteristik nasabah dari aktivitas telemarketing deposito bank. Data memiliki label binomial yang terdiri dari '*yes'* untuk nasabah berlangganan deposito dan '*no'* untuk nasabah tidak berlangganan deposito. Tahap *preprocessing* data dilakukan dengan *downsampling* untuk membuat jumlah data lebih simetris, kemudian seleksi data dan transformasi data untuk memastikan data yang digunakan nilainya konsisten, seleksi atribut untuk memilih atribut yang paling tepat digunakan dan memberi pengaruh signifikan. Klasifikasi dilakukan dengan algoritma Logistic Regression. Data dibagi menggunakan metode *split* dengan 90% data *training* dan 10% data *testing*, dengan tujuan mengoptimalkan

* Korespondensi penulis: atikamutiarachim@untagsmg.ac.id

proses *training*. Hasil *performance* terdiri dari akurasi 88,53%, nilai *classification error* 11,4%, dapat dikategorikan rendah, menunjukkan hanya sedikit kesalahan yang dihasilkan model algoritma, nilai kappa 0,68 mendekati 1, sehingga dikategorikan baik, nilai RMSE rendah yaitu 0,3 menunjukkan model akurat, dan prosentase nilai AUC yang tinggi 93,4% menunjukkan algoritma tepat digunakan pada dataset ini, karena menghasilkan nilai *performance* yang baik.

PENDAHULUAN

Penelitian mengenai pemasaran deposito masih menjadi fokus untuk dilakukan karena bank membutuhkan strategi efektif untuk memasarkan deposito. Deposito merupakan produk simpanan bank yang menjanjikan *return* terhadap nasabah, merupakan sumber pendanaan yang stabil serta dapat diprediksi bagi pihak bank. Prediksi calon nasabah deposito merupakan aspek penting untuk mengoptimalkan pencapaian target *funding*, menyusun strategi dan aktivitas operasional bank serta penentuan strategi pemasaran produk perbankan. Model yang akurat dibutuhkan untuk memprediksi kemungkinan seorang calon nasabah tertarik membeli produk deposito atau tidak.

Faktor yang mempengaruhi ketertarikan nasabah terhadap deposito salah satunya aktivitas pemasaran yang dilakukan bank. Penelitian ini bertujuan mengkaji pengaruh Logistic Regression pada klasifikasi nasabah deposito untuk mengoptimalkan efektivitas pemasaran deposito yang sekaligus berdampak terhadap peningkatan jumlah rekening maupun nominal deposito suatu bank. Dataset rekening maupun nominal deposito yang digunakan merupakan dataset publik Bank Marketing UCI Machine Learning repository (https://archive.ics.uci.edu/dataset/222/bank+marketing). Dataset ini mengandung informasi riwayat pemasaran deposito bank pada sekitar 45.211 nasabah melalui telemarketing, dengan 16 atribut dan satu label binomial berisi yes dan no, yang menerangkan apakah nasabah bersedia berlangganan produk deposito bank atau tidak. Proses preprocessing data dilakukan guna mengetahui variabel mana yang paling efektif untuk meningkatkan akurasi pemodelan klasifikasi. Algoritma klasifikasi yang digunakan adalah Logistic Regression.

Penelitian [1] menggunakan dataset UCI Bank Marketing dengan membandingkan algoritma NN, SVM, Decision Tree, Rules Based, Fuzzy Logic, Regression Model, KNN, Random Forest, Naïve Bayes. Dataset berjumlah 45212 dengan 16 atribut dan label berisi *yes* dan *no*. Akurasi yang dihasilkan masing-masing metode adalah NN memperoleh akurasi tertinggi 87.54%, kemudian Naïve Bayes sebesar 83.78%, di urutan ketiga Regression Model sebesar 82.32%, Decision Tree sebesar 75.64%, SVM sebesar 73.45%, Fuzzy Logic sebesar 71.34%, Random Forest sebesar 68.77%, Rules Based sebesar 65.32%, KNN sebesar 62.54%.

Penelitian [2] menggunakan data bank deposito KTM dari Kaggle, terdiri dari 11162 data dengan 17 atribut, seleksi fitur PCA, hasil akurasi menunjukkan SVM RBF kernel dengan parameter C 80,51%, dan ANN 80,78%.

Penelitian [3] menggunakan data *car loans* sebanyak 25601 dengan 7 atribut, pembagian data split 80:20, metode klasifikasi Logistik Regression menghasilkan akurasi 85%.

Penelitian terdahulu menggunakan algoritma Logistic Regression dengan data non perbankan adalah [4] menggunakan dataset Kaggle prima-indians-diabetes-database dengan 768 data dengan 268 ya dan 500 tidak, melakukan klasifikasi data diabetes untuk memperoleh model dengan akurasi terbaik. Pembagian data 10 *folds cross validation*, hasil akurasi menunjukkan

Logistic Regression memperoleh hasil terbaik dengan akurasi 75,78% AUC 0,801, kemudian Naive Bayes dengan akurasi 74,87% AUC 0,799, dan terakhir Neural Network dengan akurasi 69,27% AUC 0,736.

Penelitian [5] menggunakan data HIV, dengan 15 atibut, total 1757 data, pembagian *split* 70:30, hasil akurasi naive bayes 81,6%, *modified* dan *traditional* Logistic Regression 84,9%.

Penelitian [6] klasifikasi ketepatan pemberian kartu keluarga Semarang dengan data sekunder hasil Survey Sosial Ekonomi Nasional (SUSENAS) tahun 2018, pembagian data metode split 60:40, hasil akurasi yang diperoleh Regresi Logistik Biner akurasi 88% kesalahan 12% dan metode CHAID akurasi 90,2% kesalahan 9,8%.

Prediksi kematian pasien *cardiovascular disease* [7]. Total 507 data pasien *cardiovascular* Rumah Sakit Mostar sejak 2011 sampai dengan 2017, terdiri dari 123 *died* / meninggal dan 384 *survived* / bertahan. Preprocessing *replace missing value* dengan metode k-NN, seleksi fitur hapus atribut yang mengandung lebih dari 60% *missing value*, terpilih 33 atribut. Pembagian data metode *split* 70:30, hasil akurasi Neural Networks 83,12%, Logistic Regression 80,17% dan Decision Tree 76,65%.

Penelitian [8] dengan 11.406 data pembayaran pajak Sistem Pendapatan Daerah (SIMPENDA) Kota Cirebon, dengan 12 atribut label patuh dan tidak patuh. Pembagian data *split* 80:20. Hasil akurasi klasifikasi Regresi Logistik sebesar 93,97%

Penelitian terdahulu menunjukkan metode Logistic Regression diterapkan pada beberapa data yang berbeda, memberikan hasil akurasi yang cukup tinggi diatas 60%, akurasi diatas 90% biasanya dianggap terlalu bagus, dibawah 60% dianggap buruk [9]–[11]. Hasil penelitian diharapkan dapat menjadi acuan bank dalam memprediksi calon nasabah deposito, serta menyusun strategi pemasaran deposito yang efektif.

TINJAUAN PUSTAKA

A. Pemasaran Bank Deposito

Deposito merupakan simpanan yang hanya dapat dicairkan pada waktu tertentu, dengan syarat-syarat tertentu. Deposito menguntungkan bagi perbankan karena dana nasabah tersimpan lebih lama, jangka waktu relatif panjang dan frekuensi penarikan yang jarang, sehingga bank leluasa memanfaatkan dana deposito untuk produk kredit. Pemasaran merupakan strategi untuk meningkatkan penjualan produk dan jasa termasuk deposito. Bank melakukan pemasaran dengan banyak cara *online* dan *offline*. Media, teknik dan segmen promosi harus tepat agar pencapaian target lebih optimal [12]. Strategi pemasaran yang tepat dapat meningkatkan minat nasabah melakukan deposito, sekaligus meningkatkan modal agar tidak mengalami ancaman krisis keuangan.

Bank menyimpan data nasabah untuk mengetahui profil, aliran dana, menjaga hubungan dengan nasabah dan melakukan penawaran produk bank secara individual. Data yang wajib ada antara lain nama, jenis kelamin, usia, pekerjaan, status pendidikan, no telepon, saldo, mutasi dana, data pinjaman, BI *checking* sampai dengan tipe komunikasi dalam menyampaikan promosi. Pengolahan data nasabah dapat membantu pihak bank mengetahui dan menentukan klasifikasi segmen nasabah secara akurat, sehingga target deposito tercapai optimal.

B. Data Mining

Data mining merupakan proses analisis dan mengolah data dengan jumlah besar, untuk menemukan pola, hubungan, wawasan dan informasi berharga, yang berguna bagi *user* [11][13][14]. Pengolahan data dilakukan dengan kombinasi teknik statistik, *artificial intelligence* dan *machine learning.* Metode dalam data mining antara lain klasifikasi, *clustering*, asosiasi, prediksi, deteksi anomali, dan *forecasting. Data mining* banyak diaplikasikan dalam bidang kehidupan seperti medis, kondisi alam, sistem keamanan, bisnis termasuk perbankan.

C. Klasifikasi Logistic Regression

Klasifikasi merupakan teknik *data mining* dengan melihat atribut dari kelompok yang telah didefinisikan (*supervised*) [11]. Logistic Regression merupakan metode klasifikasi untuk mengetahui kemungkinan kegagalan atau keberhasilan suatu peristiwa, memprediksi apakah suatu *output* akan terjadi atau tidak. Logistic Regression dikenal juga sebagai *Binomial* Logistic Regression karena data yang digunakan memiliki label binomial, bersifat dikotomi, yang berarti hanya ada dua hasil, dapat dikategorikan sebagai 0 atau 1, benar atau salah, ya atau tidak [15][16][17][18].

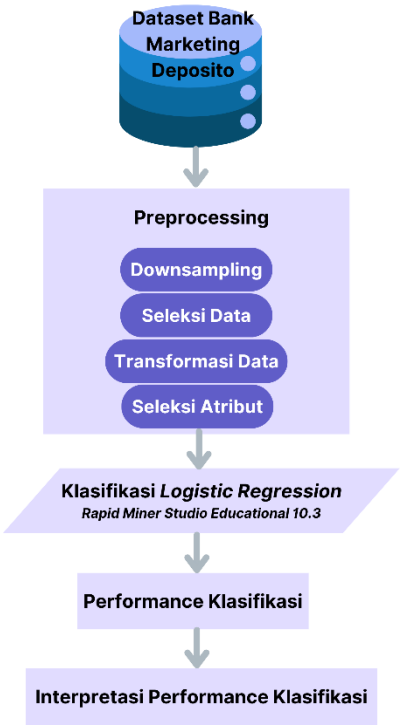
Kelebihan algoritma Logistic Regression antara lain :

1. Mempertahankan fitur regresi linier dan dapat menganalisis hasil biner.
2. Tidak terlalu rentan terhadap *over-fitting* bahkan pada data berdimensi rendah, karena menggunakan fungsi logistik sigmoid.
3. Tidak memerlukan asumsi normalitas pada variabel bebasnya, sehingga variabel bebas tidak harus berdistribusi normal
4. Cocok untuk pemodelan hubungan non-linear
5. Dapat menafsirkan koefisien model sebagai indikator pentingnya suatu fitur, tidak hanya memberikan ukuran seberapa tepat suatu prediktor koefisien, tetapi juga arah asosiasinya (positif atau negatif).
6. Dapat menghitung probabilitas keanggotaan suatu kelas
7. Model mudah diinterpretasikan karena menghasilkan koefisien yang dapat langsung diartikan sebagai odds ratio
8. Variabel dependen tidak harus kontinu, dapat dikotomi, ordinal maupun nominal.
9. Mampu menangani data dengan banyak variabel independen.
10. Algoritma *regresi logistik* efisien secara komputasi [19].

Kekurangan algoritma Logistic Regression memerlukan pemilihan variabel / atribut independen yang cermat dan pilihan strategi pembangunan model, sehingga diperlukan *preprocessing* data yang teliti untuk mengoptimalkan hasil klasifikasi Logistic Regression. Alasan para peneliti memilih metode Logistic Regression karena akurasi AUC dan CA yang dihasilkan lebih baik serta lebih hemat biaya [20].

METODE PENELITIAN

Penelitian ini dilakukan dengan mengolah dataset agar konsisten, memilih atribut yang tepat kemudian diproses pada algoritma Logistic Regression. Proses penelitian digambarkan pada alur penelitian.



Gambar 1. Alur Penelitian

PEMBAHASAN DAN HASIL

Penelitian ini menggunakan dataset publik Bank Marketing dari UCI Dataset bank-full.csv, dengan jumlah total data 45.211, 16 atribut dan 1 label binomial *yes* dan *no*. Atribut terdiri dari *age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome*, dan y (label).

A. *Preprocessing* Data

Tahap *preprocessing* bertujuan untuk mengolah data agar lebih lengkap dan konsisten [19].

1. Data *Downsampling*

Dataset bank marketing memiliki ketidakseimbangan jumlah antara data berlabel *yes* dan *no*, dari 45211 data terdapat 5289 data *yes* dan 39922 data *no*. *Imbalance dataset* adalah kondisi dimana salah satu data klasifikasi memiliki jumlah data yang lebih kecil dibandingkan kelompok klasifikasi lainnya, sehingga beresiko menyebabkan model menjadi salah klasifikasi karena sebagian besar data kelompok minoritas lebih berharga dibandingkan yang mayoritas [21]. Tiga tingkat *imbalance dataset* dilihat dari jumlah dataset minoritas yaitu *mild* 20-40%, *moderate 1-20%* dan *extreme* kurang dari 1% . Cara untuk mengatasi *imbalance dataset* adalah dengan melakukan *downsampling* dan *upweighting* [22]. Pada penelitian ini dilakukan *downsampling* data dengan melakukan hapus data secara random pada data mayoritas, dengan beberapa syarat yang telah ditentukan.

2. Seleksi Data

Proses seleksi data dilakukan dengan *cleansing data* yaitu menghapus data dengan nilai *coefficient* negatif. Hapus data dilakukan pada atribut job dengan jenis pekerjaan *unknown, housemaid, entrepreneur, blue-collar, self-employed, technician, services*. Pada atribut month, hapus data dilakukan pada data jam yang berarti bulan januari. Data *unknown* dihapus dari atribut contact, sekaligus bertujuan agar jenis kontak saat promosi lebih jelas yaitu *cellular* atau *telephone*.

3. Transformasi Data

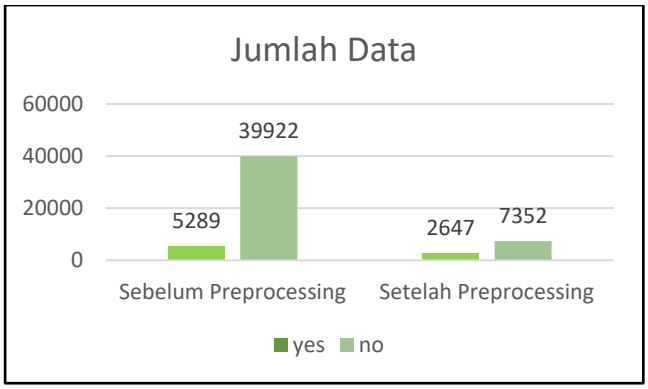
Atribut *balance* dan pdays memiliki data *outlier* yang sangat banyak, selanjutnya diubah menggunakan fungsi LN. Fungsi LN pada excel bertujuan untuk mengembalikan logaritma natural dari suatu bilangan. Seleksi atribut dilakukan guna memilih atribut yang paling tepat agar akurasi yang dihasilkan semakin baik.

4. Seleksi Atribut

Atribut yang mengandung nilai bias yaitu *marital, loan* dan *housing* dihapus, sehingga dihasilkan 9.999 data dengan 2647 data *yes* dan 7352 data *no*. Tipe data atribut *default, contact* dan *label* yang semula polynominal diubah menjadi binomial, sesuai dengan isi data. Hasil seleksi atribut ditampilkan pada tabel atribut.

Tabel 1. Atribut

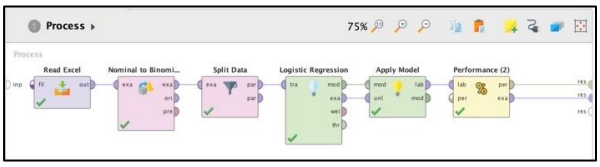
No	Atribut	Tipe Data
1	*age*	Integer
2	*job*	Polynominal
3	*education*	Polynominal
4	*default*	Binomial
5	*balance*	Real
6	*contact*	binomial
7	*day*	integer
8	*month*	polynominal
9	*duration*	integer
10	*campaign*	integer
11	*pdays*	integer
12	*previous*	integer
13	*potcome*	polynominal
14	y (label)	binomial



Gambar 2. Jumlah Data Sebelum dan Setelah *Preprocessing*

B. Klasifikasi Logistic Regression

Data dibagi dengan metode split. Metode split 50:50 memberikan hasil akurasi tertinggi yaitu 88,60%, namun pada penelitian ini digunakan hasil performance dari metode split 90% data training dan 10% data testing, dengan tujuan proses learning pada dataset dapat lebih optimal [23]. Klasifikasi Logistic Regression diterapkan dengan performance yang akan dihasilkan berupa nilai *accuracy, classification error*, kappa, *absolute error*, RMSE, *correlation*, AUC, *precision*, *recall*, F *measure*, *sensitivity* dan *specificity*.



Gambar 3. Proses pada Rapid Miner

C. *Performance* Klasifikasi

Penelitian menggunakan *tool Rapid Miner Studio Educational* 10.3. Hasil keseluruhan *performance* klasifikasi disampaikan pada tabel 3. *Confusion matrix* merupakan matrik dua dimensi yang menggambarkan perbandingan hasil prediksi dengan kenyataan. Empat kondisi dalam *confusion matrix* yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP) dan *false negative* (FN).

Tabel 2. *Confusion Matrix*

	true yes	***true no***	***class precision***
pred yes	**TP** 1600	**FP** 250	86,49%
pred no	**FN** 782	**TN** 6367	89,06%
class recall	67,17%	96,22%	

Nilai *accuracy* adalah tingkat kedekatan antara nilai yang dapat terhadap nilai sebenarnya.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{1600+6367}{1600+6367+250+782}$$

$= 0{,}8853 = 88{,}53\%$

Nilai *precision* untuk mengukur ketepatan data yang relevan.

$$Precision = \frac{TP}{TP+FP} = \frac{1600}{1600+250} = 0{,}8649 = 86{,}49\%$$

Nilai *recall* untuk mengukur proporsi positif sebenarnya yang diklasifikasi dengan benar.

Nilai *recall* sama dengan nilai *sensitivity*.

Recall & Sensitivity $= \frac{\text{TP}}{\text{TP+FN}} = \frac{1600}{1600+782} =$ 0,6717 = 67,17%

Nilai *recall* tinggi berarti *false positive* lebih banyak terjadi dibandingkan *false negative*. Nilai *precision* tinggi berarti *true positive* lebih banyak dibandingkan terjadinya *false negative* [24][25][25]. Hasil perhitungan menunjukkan nilai *precision* lebih tinggi dibandingkan nilai *recall*, hal ini berarti baik karena penelitian bertujuan untuk mengoptimalkan hasil klasifikasi data berlabel yes, yang berarti nasabah bersedia berlangganan deposito.

Nilai *specificity* merupakan kebenaran memprediksi negatif dibandingkan dengan keseluruhan data negatif.

Specificity $= \frac{\text{TN}}{\text{TN+FP}} = \frac{6367}{6367+250} = 0{,}9622 =$ 96,22%

F-Measure sering disebut sebagai F-score atau F1-score merupakan bobot perbandingan rata-rata dari *precision* dan *recall*. Jika dataset memiliki jumlah data *false negative* dan *false positive* yang mendekati / simetris, nilai akurasi sangat baik digunakan sebagai acuan *performance* algoritma, namun jika jumlah data tidak simetris, nilai F-Measure yang digunakan sebagai acuan ukuran *performance* [26].

F1-Measure= $2 \times \frac{Recall \times Precision}{Recall + Precision} = 2 \times \frac{67{,}17\% \times 86{,}49\%}{67{,}17\% + 86{,}49\%} = 75{,}61\%$

Nilai F-Measure 75,61 % menunjukkan bahwa algoritma Logistic Regression melakukan prediksi klasifikasi dengan baik.

Tabel 3. *Performance* Klasifikasi

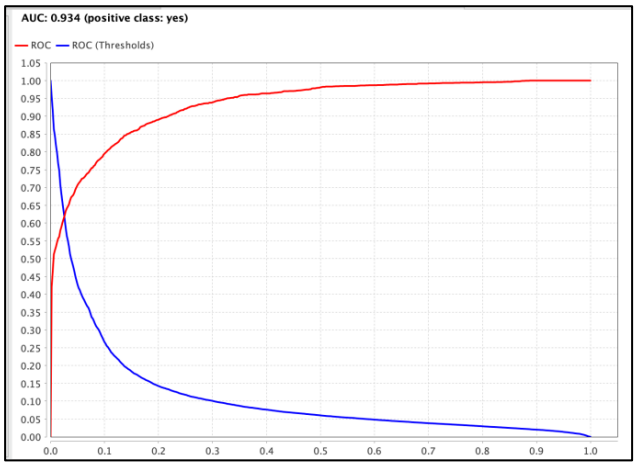
Criterion	***Performance***
Accuracy	88,53%
Precision	86,49%
Recall	67,17%
F-Measure	75,61%
Sensitivity	67,17%
Specificity	96,22%
Classification Error	11,47%
Kappa	0,68
RMSE	0,290
AUC	93,4%

Nilai *classification error* merupakan performa kesalahan dari algoritma klasifikasi yang digunakan [27]. Nilai *classification error* yang dihasilkan sebesar 11,4%, nilai akurasi yang dihasilkan adalah 88,53%, sehingga dapat dikatakan menghasilkan sangat sedikit kesalahan yang terjadi pada model yang digunakan.

Nilai kappa merupakan normalisasi dari nilai *accuracy*. Nilai kappa mendekati 1 berarti model dikatakan sempurna. Nilai kappa yang dihasilkan sebesar 0,68 jika dibulatkan menjadi 0,7 maka model dikatakan baik.

Nilai RMSE merupakan akar kuadrat dari rata-rata kesalahan kuadrat, sehingga lebih sensitif terhadap *outliers* dibandingkan metrik akurasi lainnya. Jika nilai RMSE rendah maka model dikatakan akurat. Nilai RMSE yang diperoleh rendah yaitu 0,29 jika dibulatkan menjadi 0,3 maka model dikatakan akurat.

Prosentase AUC 93,4%. Nilai AUC yang tinggi menunjukkan model klasifikasi semakin baik dan tepat digunakan pada data tersebut.



Gambar 4. Grafik AUC

KESIMPULAN

Hasil penelitian menunjukkan performa yang dihasilkan dapat dikategorikan baik dari segala aspek yaitu akurasi 88,53%, *classification error* rendah 11,4% yang artinya hanya sedikit kesalahan yang dihasilkan model algoritma, nilai kappa 0,68 mendekati 1 yang artinya baik, nilai RMSE rendah yaitu 0,3 menunjukkan model akurat, dan prosentase nilai AUC yang tinggi 93,4%. Hal ini menunjukkan bahwa algoritma Logistic Regression tepat digunakan pada dataset ini, karena menghasilkan nilai performance yang baik.

DAFTAR PUSTAKA

[1] F. Izhari, "Teknik Machine Learning untuk Bank Marketing Dataset," in *Prosiding SENATIKA (Seminar Nasional Informatika) 2021*, 2021.

[2] V. Bunga Tiara, A. M. Siregar, D. S. K. Kusumaningrum, and T. Rohana, "Bank Customer Segmentation Model Using Machine Learning," *J. Nas. Pendidik. Tek. Inform.*, vol. 13, no. 1, pp. 66–79, 2024.

[3] S. Zahi and B. Achchab, "Modeling Car Loan Prepayment Using Supervised Machine Learning," *Procedia Comput. Sci.*, vol. 170, pp. 1128–1133, 2020.

[4] D. Y. Utami, E. Nurlelah, and F. N. Hasan, "Comparison of Neural Network Algorithm, Naive Bayes and Logistic Regression To Find the Highest Accuracy in Diabetes," vol. 5, no. July, pp. 53–64, 2021.

[5] J. Shepherd, D. Candia, and F. Fuller Bbosa, "A Comparison of Logistic Regression, Modified Logistic Regression and Naïve Bayes Models for Classifying HIV Viral Load Suppression: The Case of Zombo District in Uganda," *London J. Med. Heal. Res.*, vol. 23, no. 13, 2023.

[6] M. A. Suhendra, "Ketepatan Klasifikasi Pemberian Kartu Keluarga Sejahtera di Kota Semarang Menggunakan Metode Regresi Logistik Biner dan Metode Chaid," vol. 9, pp. 64–74, 2020.

[7] D. Imamovic, E. Babovic, and N. Bijedic, "Prediction of mortality in patients with cardiovascular disease using data mining methods," *2020 19th Int. Symp. INFOTEH-JAHORINA*, no. March, pp. 1–4, 2020.

[8] M. Ripai, U. Hayati, W. Widyawati, and H. Susana, "Pengklasifikasian Surat Pemberitahuan Pajak Daerah Menggunakan Metode Regresi Logistik Biner Untuk Mengetahui Patuh Dan Tidak Patuh Dalam Pembayaran Pajak Daerah," vol. 06, no. 01, pp. 27–33, 2022.

[9] R. S. Koszalinski, A. Khojandi, and X. Li, "Missing Data, Data Cleansing, and Treatment From a Primary Study: Implications for Predictive Models," *Comput. Informatics Nurs.*, no. August, pp. 367–371, 2018.

[10] M. S. Paolella, *Linear Models and Time-Series Analysis*. usa: Wiley, 2019.

[11] S. Sarosa, *Eksplorasi dan Analisis Data Bisnis*, 1st ed. Daerah Istimewa Yogyakarta: Penerbit PT Kanisius (Anggota IKAPI), 2023.

[12] A. Mutiarachim and J. Tyoso, "Pelatihan Pembuatan Media Promosi Mudah dan Menarik dengan Aplikasi Canva untuk UMKM di Desa Blerong Kabupaten Demak," *J. Pengabdi. Masy. Nusant.*, vol. 4, no. 1, pp. 1–8, 2024.

[13] M. North, *Data mining for the masses*. .

[14] S. Fatima *et al.*, "Data Mining Methods and Obstacles: A Comprehensive Analysis," *J. Comput. Biomed. Informatics*, vol. 6, no. 1, 2024.

[15] R. C. Hill, W. E. Griffiths, and G. C. Lim, *Principles of Econometrics*, 5th ed. Wiley, 2018.

[16] K. N. R. Kumar, *Econometrics*, 1st ed. USA: CRC Press LLC, 2020.

[17] J. M. Wooldridge, *Introductory Econometrics A Modern Approach*, 7th ed. Boston: Cengage, 2020.

[18] J. H. Stock and M. W. Watson, *Introduction to Econometrics*, 4th ed. Pearson, 2020.

[19] M. North, *Data Mining for the Masses*. 2012.

[20] M. Ismail, H. Abas, R. Ramli, and R. L. Yussof, "A Review of Classification on Credit Repayment Default Behaviour using Machine Learning Algorithms," *Proc. Comput. Sci.*, vol. 2023, no. December, pp. 1–7, 2024.

[21] D. Ramyachitra and P. Manikandan, "Imbalanced Dataset Classification and Solutions : A Review," *Int. J. Comput. Bus. Res.*, vol. 5, no. 4, 2014.

[22] Google, "Imbalanced Data," *Google*. [Online]. Available: https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data. [Accessed: 22-Jan-2024].

[23] A. Bisri and R. Rachmatika, "Prediksi Kelulusan Mahasiswa Menggunakan Metode Machine Learning Pada Level Data untuk Menangani Ketidakseimbangan Kelas," 2019.

[24] D. Martin and W. Powers, "Evaluation : From precision , recall and F-measure to ROC , informedness , markedness & correlation EVALUATION : FROM PRECISION , RECALL AND F-MEASURE TO ROC , INFORMEDNESS , MARKEDNESS & CORRELATION," no. May, 2015.

[25] S. Setiawan, "Membicarakan Precision, Recall, dan F1-Score," *Medium.com*, 2020. [Online]. Available: https://stevkarta.medium.com/membicarakan-precision-recall-dan-f1-score-e96d81910354. [Accessed: 22-Jan-2024].

[26] P. Christen, D. J. Hand, and N. Kirielle, "A Review of the F-Measure : Its History , Properties , Criticism ," vol. 56, no. 3, 2023.

[27] R. T. Silangen and Y. Matdoan, "Klasifikasi Hasil Seleksi Kompotensi Dasar CPNS Menggunakan Metode Decision Tree," vol. 5, no. 2, pp. 69–75, 2022.