Jurnal Komputer dan Informatika
Vol 17 No 1, April 2022: hlm 9-13

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

KLASIFIKASI PEMASARAN BANK PORTUGAL MENGGUNAKAN METODE LOGISTIC REGRESSION

Joshua. A. Pratama
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat, DKI Jakarta, Indonesia 11410
e-mail: *joshua.535200028@stu.untar.ac.id*

ABSTRAK

Penelitian ini membahas penggunaan tiga metode klasifikasi yang berbeda, yaitu Regresi Logistik, Random Forest, dan Decision Tree, dalam konteks strategi pemasaran yang diterapkan oleh sebuah bank di Portugal. Tujuannya adalah untuk mengelompokkan pelanggan bank ke dalam dua kategori, yaitu yang berlangganan (1) atau tidak berlangganan (0) produk atau layanan yang disediakan oleh bank. Data yang digunakan dalam penelitian ini mencakup beragam informasi mengenai pelanggan, termasuk usia mereka, saldo rekening, durasi kontak terakhir dalam kampanye pemasaran, hasil kampanye sebelumnya, dan atribut-atribut lain yang relevan.. Penelitian ini diawali dengan melibatkan pemrosesan data, termasuk penanganan *missing values*, transformasi variabel kategorikal, dan pembagian Kemudian, tiga metode klasifikasi, yaitu *Logistic Regression*, *Random Forest*, dan *Decision Tree*, diimplementasikan dan diuji pada data pelatihan. Hasilnya dievaluasi menggunakan berbagai metrik evaluasi, termasuk akurasi, presisi, *recall*, dan *F1-score*. data menjadi data pelatihan dan pengujian.

Kata kunci: Bank, Klasifikasi, *Logistic Regression*, *Random Forest*, *Decision Tree*

ABSTRACT

The research discusses the utilization of three different classification methods: Logistic Regression, Random Forest, and Decision Trees, in the context of a marketing strategy implemented by a bank in Portugal. The aim is to categorize bank customers into two groups, those who subscribe (1) and those who do not subscribe (0) to the products or services offered by the bank. The data used in this study comprises various customer attributes, including their age, account balance, the duration of the last contact in the marketing campaign, the outcomes of previous campaigns, and other relevant attributes. The research commences with data preprocessing, which involves handling missing values, transforming categorical variables, and splitting the data into training and testing sets. Subsequently, the three classification methods, namely Logistic Regression, Random Forest, and Decision Trees, are implemented and tested on the training data. The results are evaluated using various evaluation metrics, including accuracy, precision, recall, and F1-score.

Keywords: *Bank, Classification, Logistic Regression, Random Forest, Decision Tree*.

1. PENDAHULUAN

Pemasaran memegang peranan sentral dalam strategi bisnis bank, dengan tujuan merambah pangsa pasar, meningkatkan loyalitas pelanggan, dan mengoptimalkan portofolio produk dan layanan. Di tengah perubahan era digital saat ini, bank harus mengadopsi pendekatan cerdas dan efisien guna mencapai target pemasaran yang ditetapkan. Salah satu pendekatan yang menarik perhatian adalah pemanfaatan metode analisis data untuk memahami perilaku pelanggan dan mengidentifikasi mereka yang berpotensi menjadi pelanggan produk atau layanan bank.

2. TINJAUAN LITERATUR

Beberapa penelitian sebelumnya menunjukkan penerapan berbagai metode klasifikasi dalam berbagai bidang, termasuk penggunaan *logistic regression* dan algoritma lainnya sebagai pembanding. Penelitian oleh [1] mengkaji Klasifikasi Persepsi Pengguna *Twitter* terhadap Kasus

Covid-19 menggunakan metode *logistic regression* dengan hasil akurasi 77% pada hyperparameter L2 dan 74% pada *hyperparameter none*. Sementara itu, penelitian oleh [2] melakukan penelitian terkait Klasifikasi *Breast Cancer* menggunakan metode yang sama dan memperoleh akurasi 76,04% pada data latih serta 83,33% pada data uji. Selain itu, penelitian lain oleh [3] menerapkan *logistic regression* untuk klasifikasi SMS *Spam* dengan rasio data latih dan uji sebesar 80:20 serta akurasi mencapai 95%.

Metode *decision tree* juga banyak digunakan dalam penelitian klasifikasi. Penelitian oleh [4] menggunakan *decision tree* untuk menentukan jadwal kerja karyawan dan memperoleh akurasi 87%. Penelitian serupa oleh [5] membandingkan metode *KNN*, *Naïve Bayes*, dan *Decision Tree* untuk klasifikasi kualitas air, di mana *KNN* menghasilkan akurasi tertinggi sebesar 86,88%. Selanjutnya, penelitian dari [6] menerapkan *decision tree* untuk klasifikasi status gizi balita di Kabupaten Simalungun dan berhasil mencapai akurasi sempurna, yaitu 100%.

Dalam ranah *machine learning*, metode *Random Forest* juga menunjukkan performa yang kompetitif. Penelitian oleh [7] membandingkan algoritma *Random Forest* dan SVM dalam analisis sentimen PSBB dengan hasil akurasi masing-masing sebesar 57,8% dan 55,7%. Penelitian dari [8] melakukan analisis optimasi pada *Random Forest* untuk data *Bank Marketing* dan mendapatkan akurasi 88,30%. Penelitian oleh [9] juga menunjukkan bahwa *Random Forest* mampu mencapai akurasi 86,56% dalam prediksi pengobatan penyakit kutil.

Selain itu, penelitian oleh [10] membandingkan berbagai metode *supervised learning* pada data *bank customers* menggunakan *Python* dengan hasil akurasi *logistic regression* sebesar 82%, *decision tree* sebesar 79,1%, dan *random forest* sebesar 86,2%. Dalam penelitian lain oleh [11] menganalisis data penyakit jantung koroner dan menunjukkan bahwa *random forest* memiliki akurasi 85,67%, lebih baik dibanding *decision tree* yang hanya mencapai 80,33%. Sementara itu, penelitian lainnya dari [12] membandingkan *decision tree*, *naïve bayes*, dan *random forest* untuk klasifikasi penyakit jantung, dengan hasil terbaik diperoleh *random forest* sebesar 75%.

Penelitian lain oleh [13] membandingkan metode *KNN*, *Decision Tree*, dan *Random Forest* pada data *imbalanced class* untuk klasifikasi promosi karyawan. Hasilnya menunjukkan bahwa *random forest* memiliki akurasi tertinggi yaitu 86,37%, sedikit lebih tinggi dibanding *decision tree* sebesar 85,29%. Selain itu, penerapan metode regresi logistik biner juga dilakukan oleh [19] untuk mengetahui faktor-faktor yang memengaruhi kesiapsiagaan rumah tangga dalam menghadapi bencana alam, sedangkan oleh [20] menggunakan regresi logistik ordinal dan probit ordinal untuk mengestimasi probabilitas lama masa studi mahasiswa IST AKPRIND Yogyakarta.

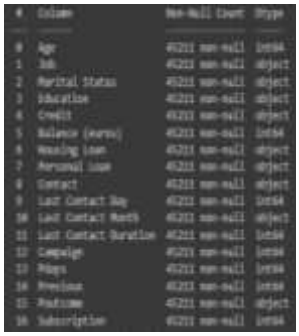
Berdasarkan berbagai penelitian tersebut, dapat disimpulkan bahwa metode *logistic regression*, *decision tree*, dan *random forest* memiliki performa yang baik dalam menyelesaikan berbagai masalah klasifikasi di bidang kesehatan, sosial, maupun ekonomi. Oleh karena itu, penelitian klasifikasi pemasaran bank Portugal dengan menggunakan metode *logistic regression* menjadi relevan untuk dilakukan sebagai bagian dari pengembangan dan pembandingan performa model klasifikasi dalam konteks data keuangan.

3. METODE PENELITIAN

Data yang digunakan berasal dari situs Kaggle dengan total 16 fitur dan 1 target. Fitur-fitur tersebut meliputi age (umur nasabah), job (jenis pekerjaan), marital (status perkawinan: married, divorced, single), education (tingkat pendidikan: unknown, secondary, primary, tertiary), default (status gagal bayar kredit), balance (rata-rata saldo tahunan dalam euro), housing (kepemilikan pinjaman perumahan), loan (pinjaman pribadi), contact (jenis komunikasi: telephone, cellular, unknown), day (hari kontak terakhir), month (bulan kontak terakhir), duration (durasi kontak dalam detik), campaign (jumlah kontak selama kampanye), pdays (hari sejak kontak terakhir dari kampanye

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

sebelumnya), *previous* (jumlah kontak sebelum kampanye saat ini), dan *poutcome* (hasil kampanye sebelumnya: success, *failure*, *other*, *unknown*). Target variabelnya adalah *subscription*, yang menunjukkan apakah nasabah berlangganan deposito berjangka (1 = *yes*, 0 = *no*). Pada Gambar 1 dibawah ini ditampilkan contoh dari info dataset.



Gambar 1. Info Dataset

3.1 Pengolahan Data

Dalam tahap pra-pemrosesan akan dijalankan tahap pengecekan dimana data yang telah di input akan di cek terlebih dahulu apakah ada *missing value* pada data yang akan diteliti atau tidak. Kemudian lanjut ke tahap berikutnya dimana akan di cari apakah ada data yang sama atau *duplicate* dengan data lain. Lalu lanjut ke tahap berikutnya yaitu mengkategorikan data yang bertipe angka agar bisa menghitung unique values dari data tersebut. *Unique values* adalah istilah yang digunakan dalam berbagai situasi untuk menunjukkan nilai atau entitas yang tidak memiliki salinan atau kesamaan dalam satu kumpulan data atau himpunan.

Ini menggambarkan fakta bahwa setiap nilai di dalam himpunan tersebut hanya muncul sekali atau hanya ada satu entitas dengan karakteristik tertentu. Lalu lanjut ke tahap *Label Encoding*. *Label encoding* merupakan tahap mengubah data yang bersifat teks atau string menjadi data dengan bentuk *numerical*. Setelah melakukan tahap preprocessing data, kemudian lanjut ke tahap membagi data menjadi *train* dan test, *train* dilambangkan dengan (x) dan *test* akan dilambangkan dengan (y). Berikut pada Gambar 2 ditunjukkan potongan kode dalam bentuk gambar yang menampilkan pembagian data.

Gambar 2. Tahap Pembagian Data Fitur dan Target

4. HASIL DAN PEMBAHASAN

Hasil dan pembahasan berupa ketiga metode yang akan dicari tahu manakah model yang lebih baik untuk klasifikasi penelitian yaitu metode *logistic regression* adalah metode statistik yang digunakan untuk memodelkan hubungan antara variabel *dependen* biner (variabel terikat) dan satu atau lebih variabel *independen* (variabel bebas) yang bersifat kontinu atau kategorikal [16]., *Random Forest* adalah metodologi ilmu data dan pembelajaran mesin yang digunakan untuk membangun model prediktif yang kuat [17]. dan metode *decision tree* adalah model prediksi ilmu data dan pembelajaran mesin yang membuat keputusan berdasarkan aturan hierarki yang didefinisikan dalam bentuk struktur pohon [18].

Berikut pada Gambar 3, Gambar 4, dan Gambar 5, masing-masing adalah potongan kode yang memuat hasil dari setiap algoritma yang digunakan.

Gambar 3. Penerapan Metode *Logistic Regression*

Gambar 4. Penerapan Metode *Random Forest*

Gambar 5. Penerapan Metode *Decision Tree*

5. KESIMPULAN

Untuk ketiga metode yang telah diuji telah didapat hasil 97% untuk nilai akurasi pada metode *logistic Regression* dan hasil nilai akurasi pada metode *random forest* dan *decision tree* mendapatkan hasil yang sama yaitu 100% untuk nilai akurasi, maka bisa disimpulkan untuk penelitian klasifikasi data kali ini metode *random forest* dan *decision tree* lebih tepat digunakan dibandingkan dengan metode *logistic regression* untuk klasifikasi pemasaran bank Portugal.

DAFTAR PUSAKA

[1] Santoso, A. K. S., Noviriandini, A., Kurniasih, A., Wicaksono, B. D., & Nuryanto, A. (2021). Klasifikasi Persepsi Pengguna *Twitter* Terhadap Kasus *Covid-19* Menggunakan Metode *Logistic Regression*. Jurnal Informatika Kaputama (JIK), 5(2), 234-241.

[2] Achmad, A. D. (2022). Klasifikasi *Breast Cancer* Menggunakan Metode *Logistic Regression*. JTRISTE, 9(1), 143-148.

[3] Suprihati, F. R. (2021). Analisis Klasifikasi SMS *Spam* Menggunakan *Logistic Regression*. Jurnal Sistem Cerdas, 4(3), 155-160.

[4] Achmad, B. D. M., Slamat, F., & ITATS, F. T. I. (2012). Klasifikasi data karyawan untuk menentukan jadwal kerja menggunakan metode *decision tree*. Jurnal Iptek, 16(1).

[5] Tangkelayuk, A. (2022). Klasifikasi Kualitas Air Menggunakan Metode KNN, *Naïve Bayes*, dan *Decision Tree*. JATISI (Jurnal Teknik Informatika Dan Sistem Informasi), 9(2), 1109- 1119.

[6] Hafizan, H., & Putri, A. N. (2020). Penerapan Metode Klasifikasi *Decision Tree* Pada Status Gizi Balita Di Kabupaten Simalungun. Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen), 1(2), 68-72.

[7] Adrian, M. R., Putra, M. P., Rafialdy, M. H., & Rakhmawati, N. A. (2021). Perbandingan Metode Klasifikasi *Random Forest* dan SVM Pada Analisis Sentimen PSBB. Jurnal Informatika Upgris, 7(1).

[8] Religia, Y., Nugroho, A., & Hadikristanto, W. (2021). Analisis Perbandingan Algoritma Optimasi pada *Random Forest* untuk Klasifikasi Data *Bank Marketing*. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), 5(1), 187-192.

[9] Erdiansyah, U., Lubis, A. I., & Erwansyah, K. (2022). Komparasi Metode *K-Nearest Neighbors* dan *Random Forest* Dalam Prediksi Akurasi Klasifikasi Pengobatan Penyakit Kutil. Jurnal Media

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Informatika Budidarma, 6(1), 208-214.
[10] Pamungkas, F. S., Prasetya, B. D., & Kharisudin, I. (2020, March). Perbandingan Metode Klasifikasi *Supervised Learning* pada Data *Bank Customers* Menggunakan *Python*. In PRISMA, Prosiding Seminar Nasional Matematika (Vol. 3, pp. 692-697).
[11] Wibisono, A. B., & Fahrurozi, A. (2020). Perbandingan Algoritma Klasifikasi Dalam Pengklasifikasian Data Penyakit Jantung Koroner. jurnal ilmiah teknologi dan rekayasa, 24(3), 161-170.
[12] Depari, D. H., Widiastiwi, Y., & Santoni, M. M. (2022). Perbandingan Model *Decision Tree*, *Naive Bayes* dan *Random Forest* untuk Prediksi Klasifikasi Penyakit Jantung. Informatik: Jurnal Ilmu Komputer, 18(3), 239-248.
[13] Sotarjua, L. M., & Santoso, D. B. (2022). Perbandingan Algoritma KNN, *Decision Tree* dan *Random Forest* Pada *Data Imbalanced Class* Untuk Klasifikasi Promosi Karyawan. Jurnal INSTEK (Informatika Sains dan Teknologi), 7(2), 192-200.
[14] Guan, K., Wu, J., Kimball, J. S., Anderson, M. C., Frolking, S., Li, B., ... & Lobell, D. B. (2017). *The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields*. *Remote sensing of environment*, 199, 333-349.
[15] Budiarti, L., Tarno, T., & Warsito, B. (2013). Analisis Intervensi Dan Deteksi *Outlier* Pada Data Wisatawan Domestik (Studi Kasus Di Daerah Istimewa Yogyakarta). Jurnal Gaussian, 2(1), 39-48.
[16] Ramadhy, I. F., & Sibaroni, Y. (2022). Analisis *Trending* Topik Twitter dengan Fitur Ekspansi FastText Menggunakan Metode *Logistic Regression*. JURIKOM (Jurnal Riset Komputer), 9(1), 1- 7.
[17] Primajaya, A., & Sari, B. N. (2018). *Random forest algorithm for prediction of precipitation*. Indonesian Journal of Artificial Intelligence and Data Mining, 1(1), 27-31.
[18] Achmad, B. D. M., Slamat, F., & ITATS, F. T. I. (2012). Klasifikasi data karyawan untuk menentukan jadwal kerja menggunakan metode *decision tree*. Jurnal Iptek, 16(1).
[19] Faruk, F. M., Doven, F. S., & Budyanra, B. (2019). Penerapan Metode Regresi Logistik Biner Untuk Mengetahui Determinan Kesiapsiagaan Rumah Tangga Dalam Menghadapi Bencana Alam. In Seminar nasional official statistics (Vol. 2019, No. 1, pp. 379-389).
[20] Daga, E. K. N., & Suryowati, K. (2017). Penerapan Metode Regresi Logistik Ordinal dan Regresi Probit Ordinal untuk Mengestimasi Probabilitas Lama Masa Studi Mahasiswa IST AKPRIND Yogyakarta. Jurnal Statistika Industri dan Komputasi, 2(02), 104-114.