MODEL DETEKSI PENIPUAN DALAM TRANSAKSI PERBANKAN MENGGUNAKAN REGRESI LOGISTIK

disusun oleh:

Kelompok 6
Anyelyra Kantata (2206048625)
Geraldus Harry Pascal (2206048663)
Matthew Donathan Wibiksono (2206048612)
Widya Siti Ropiah (2206048745)

PROGRAM STUDI MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS INDONESIA
DEPOK

2023/2024

ABSTRAK

Penipuan dalam transaksi merupakan masalah yang serius yang dihadapi oleh banyak perusahaan dan individu karena sangat merugikan. Untuk menangani hal tersebut, diperlukan metode yang efektif dan efisien untuk mendeteksi penipuan dalam transaksi. Dalam makalah ini, digunakan metode regresi logistik untuk mendeteksi penipuan dalam transaksi. Metode ini dipilih karena kemampuannya dalam menangani data biner. Dalam penelitian ini, dataset yang dipakai memiliki beberapa fitur penting seperti, *type*, *amout, isFraud, oldbalanceOrg,* dan *newbalanceDest* yang dapat membangun model prediksi untuk mendeteksi penipuan. Pada makalah ini, dijelaskan berbagai tahapan sebelum membangun model hingga evaluasi model menggunakan beberapa evaluasi metrik. Hasil penelitian menunjukkan bahwa regresi logistik dapat memberikan tingkat akurasi yang tinggi dalam mendeteksi penipuan transaksi. Oleh karena itu, model deteksi penipuan dalam transaksi menggunakan regresi logistik dapat dipakai dan dimanfaatkan serta dapat dikembangkan menjadi lebih baik lagi.

Kata Kunci: regresi logistik, SMOTE, *fraud detection, imbalanced data*

DAFTAR ISI

ABSTRAK **2**
DAFTAR ISI **3**
1. PENDAHULUAN **4**
1.1 Latar Belakang 4
1.2 Rumusan Masalah 4
1.3 Tujuan Penelitian 4
2. PEMBAHASAN **4**
2.1 Data 4
2.2 Regresi Logistik 5
2.3 Implementasi 6
2.3.1 Pre-processing 6
2.3.2 Eksplorasi Data 6
2.3.3 Pembentukan Model Regresi Logistik 6
2.4 Analisis Akurasi Model 7
2.4.1 Analisis Akurasi Model (Sebelum SMOTE) 8
2.4.2 Analisis Akurasi Model (Setelah SMOTE) 9
3. PENUTUP **10**
3.1 Kesimpulan 10
3.2 Saran 10
REFERENSI **11**

1. PENDAHULUAN

1.1 Latar Belakang

Penipuan dalam transaksi menjadi masalah yang sangat serius bagi lembaga keuangan dan perusahaan yang bergerak di bidang *e-commerce*. Dengan semakin berkembangnya zaman, volume transaksi semakin besar dan bervariasi, sehingga berpotensi tinggi terjadi penipuan. Oleh karena itu, deteksi dini diperlukan untuk mencegah terjadinya penipuan serta meminimalisir kemungkinan terjadinya hal tersebut.

Salah satu metode yang efektif dan efisien adalah metode regresi logistik. Regresi logistik adalah metode statistik yang mampu memodelkan hubungan antara satu variabel dependen dan satu atau lebih variabel independen. Metode ini memiliki kemampuan untuk memberikan probabilitas kejadian yang dapat digunakan untuk mengklasifikasikan suatu transaksi sebagai penipuan atau bukan.

Penelitian ini bertujuan untuk mengeksplorasi data dan mengimplementasikan model regresi logistik dalam mendeteksi penipuan. Dengan menganalisis data transaksi dan membangun model, diharapkan model ini dapat memberikan hasil yang akurat. Dengan demikian, penelitian ini dapat memberikan kontribusi yang signifikan dalam pengembangan model deteksi penipuan serta dapat membantu perusahaan untuk mengambil tindakan preventif dalam mencegah kerugian yang diakibatkan penipuan transaksi.

1.2 Rumusan Masalah

Bagaimana membangun model deteksi penipuan yang akurat dengan menggunakan regresi logistik.

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah membangun model untuk mendeteksi penipuan dalam transaksi yang akurat dengan menggunakan regresi logistik.

2. PEMBAHASAN

2.1 Data

Model *Logistic Regression* yang dibuat memanfaatkan *dataset* Synthetic Finacial Datasets For Fraud Detection (sumber: https://kaggle.com) yang terdiri 6362620 baris dan 11 fitur (hanya diambil 50000 baris pertama untuk membangun model). Berikut ini penjelasan fitur-fitur yang tersedia pada *dataset* tersebut:

- **step**: memetakan satuan waktu di dunia nyata, 1 step adalah 1 jam waktu (total langkah adalah 744 dengan waktu simulasi selama 30 hari).
- **type**: tipe transaksi (contoh: CASH_IN, CASH_OUT, DEBIT, PAYMENT, dan TRANSFER)
- **amount**: besar transaksi
- **nameOrig**: asal transaksi
- **oldbalanceOrg**: saldo awal sebelum transaksi
- **newbalanceOrig**: saldo terbaru setelah transaksi
- **nameDest**: penerima transaksi
- **oldbalanceDest**: saldo awal penerima sebelum transaksi

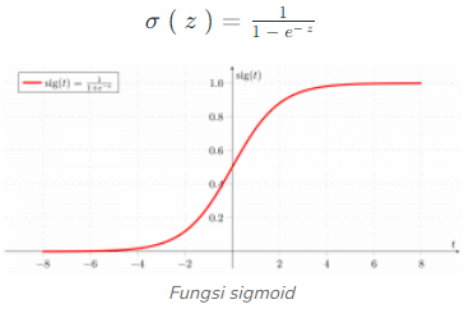
- **newbalanceDest**: saldo terbaru penerima setelah transaksi
- **isFraud**: fitur biner yang mengindikasikan bahwa transaksi adalah penipuan atau bukan (*fraudulent* (1) atau *not fraudulent* (0))
- **isFlaggedFraud**: fitur biner yang mungkin mengindikasikan apabila transaksi ditandai (*flagged*) sebagai penipuan dengan suatu sistem (1) atau tidak (0)

Tahapan yang kami gunakan dalam pemodelan ini adalah persiapan data (*pre-processing*), eksplorasi data, analisis, dan pengaplikasian SMOTE (*Synthetic Minority Oversampling Technique*). Pemodelan dengan regresi logistik ini menggunakan variabel target atau variabel yang akan diprediksi adalah 'isFraud', sebuah kolom dengan tipe data biner. Setiap langkah dan metode yang kami lakukan dilaksanakan menggunakan bantuan Google Collaboratory (*notebook* dapat diakses melalui tautan berikut: ristek.link/ProjekSainsData-4). Proses pemodelan ini terdiri dari beberapa tahapan, yaitu:

1. *Import* data ke dalam Google Collaboratory
2. Persiapan data (misalnya deteksi *missing value*, duplikasi, *encoding*, dan lain-lain)
3. Eksplorasi data
4. Pembangunan model Regresi Logistik dengan module scikit-learn
5. Aplikasi SMOTE untuk membuat data menjadi seimbang (*balanced*)
6. Analisis akurasi model

2.2 Regresi Logistik

Regresi logistik digunakan untuk klasifikasi biner dengan menggunakan fungsi sigmoid, yang mengambil input berupa variabel independen dan output berupa variabel dependen dan menghasilkan nilai probabilitas antara 0 dan 1.



Gambar 1. Fungsi Sigmoid

Logarithm of the odds, $log(\frac{p}{1-p})$, adalah linear dalam prediktor, X dan $log(\frac{p}{1-p})$ disebut fungsi logit. Logit dapat diekspresikan dengan fungsi linear dari prediktor X, untuk kasus yang lebih umum, yang melibatkan beberapa variabel independen x, berikut ini adalah persamaannya:

$$logit(p) = log(\frac{p}{1-p}) = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

$$p = \frac{e^{logit}}{1+e^{logit}}$$

2.3 Implementasi

2.3.1 Pre-processing

- Setelah memeriksa *missing value* dan duplikasi pada dataset ini, dapat disimpulkan bahwa dataset ini bersih dari nilai-nilai yang kosong ataupun duplikasi.
- Terdapat juga variabel ‘isFlaggedFraud’ yang merupakan sistem yang dibawa oleh dataset untuk menandai apakah sebuah transaksi adalah penipuan atau tidak. Namun, saat diperiksa, dari 50000 baris data pertama, tidak ada yang ditandai sebagai *fraud*, meskipun sebenarnya ada transaksi *fraud*. Maka, dapat disimpulkan bahwa sistem belum bisa mengidentifikasi dengan baik. Oleh karena itu, variabel ‘isFlaggedFraud’ akan dihapus.
- Hapus juga fitur yang tidak digunakan, yaitu ‘step’, ‘nameOrig’, ‘nameDest’. Sehingga, dataset sekarang hanya tersisa 7 fitur, dengan 1 fitur kategorik dan 6 fitur numerik.
- Karena module Linear Regression pada scikit-learn hanya memproses data numerik, maka diperlukan *encoding* untuk mengubah fitur kategorik menjadi numerik. Oleh karena itu, dilakukan One Hot Encoder pada kolom ‘type’ untuk mengubahnya menjadi kolom-kolom dengan tipe data numerik. Pada *Gambar 4*, terdapat kolom-kolom baru yang merupakan representasi dari kolom ‘type’. Setelah itu, kolom ‘type’ yang lama dihapus.

type_CASH_IN	type_CASH_OUT	type_DEBIT	type_PAYMENT	type_TRANSFER
0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	1.0
0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0

Gambar 2. Kolom-kolom baru hasil *encoding* (5 data pertama)

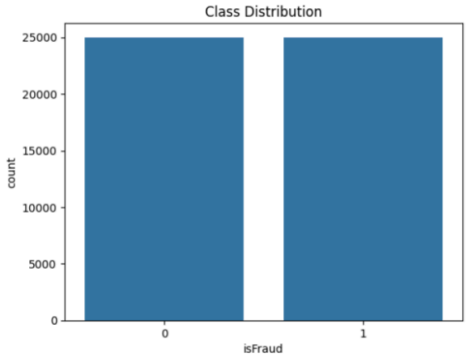
2.3.2 Eksplorasi Data

Setelah dataset diperiksa keseimbangannya, ditemukan bahwa dataset ini tidak seimbang. Pada variabel isFraud, kasus *fraudulent* (1) sangat mendominasi dibandingkan kasus *not-fraudulent* (0), yaitu *fraud* (0): 49900 dan *not-fraud* (1): 100. Jadi, dapat disimpulkan bahwa dataset ini adalah data *imbalanced*. Namun, pada tahap ini, kita akan tetap mengerjakan model dengan data yang *imbalanced* ini.

2.3.3 Pembentukan Model Regresi Logistik

- Pemodelan Regresi Logistik ini akan dilakukan dengan menggunakan variabel ‘isFraud’ sebagai variabel target dan lainnya adalah prediktor. Sebelum masuk ke dalam model, dataset terlebih dahulu dilakukan *scalling* agar berada pada skala yang sama menggunakan StandardScaler.
- Pemecahan data menjadi data *training* dan data *testing* dilakukan dengan train_test_split. Perbandingan yang digunakan adalah 80% menjadi data *training* dan 20% menjadi data *testing*.

- Model berjalan dengan baik, tetapi pada *classification report*, recall hanya bernilai 0.05 pada *frauduluent* (1), yang artinya dari semua transaksi penipuan, model hanya menebak 5% yang benar-benar adalah transaksi penipuan.
- Nilai recall yang kecil dapat terjadi karena data yang *imbalanced*, maka dari itu dilakukan SMOTE (Synthetic Minority Oversampling Technique) untuk membuat data menjadi *balanced*. SMOTE ini didapatkan dari module imblearn.



Gambar 3. Identifikasi keseimbangan data setelah pengaplikasian SMOTE

- Setelah dilakukan SMOTE, terdapat dataset yang telah *balanced* dan kembali dimasukkan dalam model. Nilai recall meningkat menjadi 0.98 pada *frauduluent* (1), yang artinya dari semua transaksi penipuan, model menebak 98% yang benar-benar adalah transaksi penipuan.

2.4 Analisis Akurasi Model

Pada bagian ini, akan dibahas mengenai akurasi model. Namun, karena model dibuat sebanyak 2 kali, yaitu pada dataset yang *imbalanced* dan dataset *balanced* (setelah diaplikasikan SMOTE). Analisis akurasi ini juga akan menggunakan *confusion matrix* seperti berikut:

		Predicted	
		No	**Yes**
Actual	**No**	**True Negative** *Diprediksi not-fraud dan Benar not-fraud*	**False Positive** *Diprediksi fraud, tetapi not-fraud*
	Yes	**False Negative** *Diprediksi not-fraud, tetapi fraud*	**True Positive** *Diprediksi fraud dan Benar fraud*

Gambar 4. Confusion matrix

Selain itu, model juga akan dievaluasi dengan beberapa metrik sebagai berikut:

- *Accuracy*: Menunjukkan proporsi dari kasus yang dimasukkan ke dalam kelas dengan benar.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total}$$

- *Precision:* Fokus kepada akurasi dari prediksi positif.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- *Recall* (*Sensitivity or True Positive Rate*): Mengukur proporsi dari kasus yang diprediksi positif secara benar terhadap semua kasus yang sebenarnya adalah positif.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- *F1 Score*: *Harmonic mean* dari *precision* dan *recall*.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

2.4.1 Analisis Akurasi Model (Sebelum SMOTE)

Dataset sebelum dilakukan SMOTE merupakan dataset yang *imbalanced*. Kasus *not-fraud* merupakan mayoritas, sehingga mendominasi yang lain. Jika dilihat dari nilai akurasi, maka model ini sangat baik, yaitu 0.9981. Namun, nilai *recall* pada kasus *fraud* (1) hanya sebesar 0.05. Hal ini terjadi karena data yang tidak seimbang antara kasus *not-fraud* (0) dan *fraud* (1).

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      9980
           1       1.00      0.05      0.10        20

    accuracy                           1.00     10000
   macro avg       1.00      0.53      0.55     10000
weighted avg       1.00      1.00      1.00     10000


Confusion Matrix:
[[9980    0]
 [  19    1]]

Model accuracy score: 0.9981
```

Gambar 5. Laporan akurasi model (sebelum dataset diaplikasikan SMOTE)

Didapatkan juga dan koefisien regresi dan persamaan dari model regresi logistik tersebut, yaitu:

```
# Lihat koefisien
print("Intercept:", lr.intercept_)
print("Koefisien:", lr.coef_)
```

```
Intercept: [-8.92952592]
Koefisien: [[-0.0450357   3.71438042 -3.8255013  -1.04745205 -0.99899604 -0.643429
   1.22396814 -0.21621456 -1.21864968  1.12902483]]
```

Gambar 6. Koefisien regresi

$$logit(p) = -8.9295 - 0.0450X_1 + 3.7144X_2 - 3.8255X_3 - 1.0475X_4 - 0.9990X_5 - 0.6434X_6 + 1.2240X_7 - 0.2162X_8 - 1.2186X_9 + 1.1290X_{10}$$

2.4.2 Analisis Akurasi Model (Setelah SMOTE)

Untuk membuat model ini lebih optimal, dataset yang *imbalanced* perlu diubah menjadi data yang *balanced*. Salah satu cara untuk menyeimbangkan dataset dengan kelas mayoritas adalah dengan mensintesis datum baru melalui SMOTE (Synthetic Minority Oversampling Technique). SMOTE memilih contoh-contoh yang dekat dalam ruang fitur, menarik garis antara contoh-contoh tersebut dalam ruang fitur, dan menggambar sampel baru pada titik di sepanjang garis tersebut.

Dataset setelah dilakukan SMOTE merupakan dataset yang *balanced*. Kasus *not-fraud* dan *fraud* kini telah seimbang dan tidak ada yang mendominasi. Jika dilihat dari nilai akurasi, model ini masing sangat baik, yaitu 0.9225. Meskipun nilai akurasi menurun sedikit, tetapi nilai *recall* pada kasus *fraud* (1) meningkat menjadi 0.98. Pada *classification report*, terlihat jelas bahwa nilai *precision*, *recall*, dan *F1-score* sangat baik, berkisar antara 0.80 - 0.97.

```
Classification Report (Setelah SMOTE):
              precision    recall  f1-score   support

           0       0.97      0.87      0.92      4972
           1       0.88      0.98      0.93      5028

    accuracy                           0.92     10000
   macro avg       0.93      0.92      0.92     10000
weighted avg       0.93      0.92      0.92     10000


Confusion Matrix (Setelah SMOTE):
[[4322  650]
 [ 125 4903]]

Model accuracy score: 0.9225
```

Gambar 7. Laporan akurasi model (setelah dataset diaplikasikan SMOTE)

Didapatkan juga dan koefisien regresi dan persamaan dari model regresi logistik tersebut, yaitu:

```
[29] # Lihat koefisien
     print("Intercept:", lr1.intercept_)
     print("Koefisien:", lr1.coef_)

     Intercept: [-3.37424799]
     Koefisien: [[-6.89054020e+00  2.09363408e+01 -2.72168249e+01 -1.20474397e+00
       -5.28318600e-03 -1.09747401e+00  1.29576396e+00 -5.13070414e-01
       -2.80787995e+00  2.11582168e+00]]
```

Gambar -. Koefisien regresi

$$logit(p) = -3.3742 - 6.8905X_1 + 20.9263X_2 - 27.2168X_3 - 1.2047X_4 - 0.0053X_5 - 1.0975X_6 + 1.2958X_7 - 0.5131X_8 - 2.8079X_9 + 2.1158X_{10}$$

3. PENUTUP

3.1 Kesimpulan

Berdasarkan model regresi logistik yang telah kami buat, diperoleh nilai akurasi yang sangat tinggi walaupun terjadi penurunan nilai akurasi yang tidak signifikan setelah menggunakan *balanced* dataset. Hal ini menunjukkan bahwa regresi logistik terbukti menjadi metode yang efektif dan efisien dalam mendeteksi penipuan dalam transaksi. Model yang dibangun mampu memberikan probabilitas transaksi menjadi penipuan, sehingga memungkinkan klasifikasi yang akurat.

3.2 Saran

Agar mendapatkan hasil yang lebih akurat pada Regresi Logistik, diperlukan *encoding* data yang tepat dan memastikan keseimbangan dataset, terkhusus pada variabel target.

REFERENSI

[1] Kotu, V., & Deshpande, B. (2018). Data Science: Concepts and Practice (2nd ed.). Morgan Kaufmann.

[2] A. Mahajan, V. S. Baghel and R. Jayaraman, "Credit Card Fraud Detection using Logistic Regression with Imbalanced Dataset," 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2023, pp. 339-342.

[3] M. Devika, S. R. Kishan, L. S. Manohar and N. Vijaya, "Credit Card Fraud Detection Using Logistic Regression," 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/ICATIECE56365.2022.10046976.

[4] GeeksforGeeks. (2024). Logistic Regression in Machine Learning. https://www.geeksforgeeks.org/understanding-logistic-regression/

[5] Machine Learning Mastery. (2021). SMOTE for Imbalanced Classification with Python. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/