

ETS PEMBELAJARAN MESIN

**PREDIKSI JUMLAH TRANSAKSI DAN KLASIFIKASI PENIPUAN KARTU KREDIT
PADA *E-COMMERCE* AMAZON MENGGUNAKAN *MACHINE LEARNING***



Disusun oleh
Aditya Agral Serhansyah
2043201082

Dosen Mata Kuliah
Mukti Ratna Dewi, S.Si., M.Sc.

**PROGRAM STUDI SARJANA TERAPAN STATISTIKA BISNIS
FAKULTAS VOKASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2023**

I. LATAR BELAKANG

E-commerce telah menjadi salah satu sektor terbesar dan paling dinamis dalam dunia perdagangan global. Amazon adalah sebuah perusahaan berbasis teknologi multinasional Amerika yang berbasis di Seattle, Amerika Serikat. Amazon, sebagai salah satu platform *e-commerce* terkemuka di dunia, telah menyediakan berbagai produk dan layanan kepada 300 juta pelanggan di seluruh dunia. Meskipun *e-commerce* telah memberikan kemudahan berbelanja bagi konsumen, platform seperti Amazon juga menghadapi berbagai tantangan, terutama dalam hal pengelolaan jumlah transaksi dan penanganan masalah penipuan.

Prediksi jumlah transaksi dan klasifikasi penipuan di platform *e-commerce* Amazon menjadi semakin penting dalam konteks ini. Seiring pertumbuhan *e-commerce*, jumlah transaksi yang dilakukan setiap hari di platform Amazon terus meningkat secara signifikan. Ini menciptakan kebutuhan untuk mengembangkan model prediksi yang dapat membantu Amazon dalam merencanakan, mengelola, dan mengoptimalkan proses logistik mereka. Selain itu, penipuan juga menjadi masalah serius yang mempengaruhi kepercayaan pelanggan dan keberlanjutan bisnis Amazon. Oleh karena itu, deteksi penipuan yang tepat waktu sangat penting untuk mengurangi kerugian finansial dan menjaga reputasi perusahaan.

Salah satu metode yang dapat digunakan untuk prediksi dan klasifikasi adalah *Machine Learning* (ML). Menurut (Patriya, 2020), *Machine Learning* adalah suatu metode yang memungkinkan program komputer memiliki kecerdasan dengan belajar dari data yang diberikan untuk memperoleh pengetahuan, yang kemudian dapat membantu manusia untuk mengambil suatu keputusan. Ciri khas dari *Machine Learning* yaitu terdapat proses pembelajaran dan pengujian. *Machine Learning* memerlukan data *training* untuk dipelajari dan data *testing* untuk diuji (Hania, 2017).

Model yang digunakan dalam melakukan prediksi jumlah transaksi adalah *Linear Regression*, *SVR*, dan *LASSO Regression*, dilakukan evaluasi model dengan RMSE, *R-squared*, MAE untuk menentukan model prediksi terbaik. Model yang digunakan dalam melakukan klasifikasi terjadi penipuan adalah *Logistic Regression*, *Support Vector Machines*, *Decision Tree Classifier*, dan *Naïve Bayes Classifier*, dilakukan evaluasi model dengan *confusion matrix* dan kurva ROC untuk menentukan model klasifikasi terbaik.

Penelitian ini bertujuan untuk mendapatkan model *machine learning* terbaik yang dapat memprediksi jumlah transaksi harian dan mengklasifikasikan transaksi sebagai penipuan atau bukan di platform *e-commerce* Amazon. Penelitian ini akan memberikan manfaat besar bagi Amazon dalam meningkatkan efisiensi operasional, mengurangi kerugian akibat penipuan, dan mempertahankan kepercayaan pelanggan.

II. METODOLOGI

Metode prediksi yang digunakan dalam penelitian ini selain metode yang ditetapkan yaitu *LASSO Regression*, sedangkan metode klasifikasinya adalah *Naïve Bayes Classifier*. Pada variabel 'Class' dimana menjadi variabel target untuk model klasifikasi terjadi data tidak seimbang yang kemudian ditangani dengan menggunakan ROSE.

2.1 LASSO Regression

Metode *Least Absolute Shrinkage and Selection Operator* (LASSO) diperkenalkan pertama kali oleh Tibshirani pada tahun 1996. LASSO menyusutkan koefisien regresi dari variabel prediktor yang memiliki korelasi tinggi dengan galat, menjadi tepat pada nol

atau mendekati nol (Tibshirani, 1996). Menurut Zhao dan Yu (2006), persamaan secara umum LASSO dinyatakan sebagai berikut:

Menurut Tibshirani (1996) estimasi koefisien LASSO menggunakan pemrograman kuadratik dengan kendala pertidaksamaan. Estimasi LASSO diperoleh dari persamaan berikut:

$$\mathbf{Y}^{**} = \mathbf{X}^{**} \boldsymbol{\beta} + \boldsymbol{\varepsilon}^{**}$$

Keterangan:

- \mathbf{Y}^{**} : vektor variabel respon berukuran $(n \times 1)$
- \mathbf{X}^{**} : matriks variabel prediktor berukuran $(n \times p)$
- $\boldsymbol{\beta}$: vektor dari koefisien LASSO berukuran $(k+1) \times 1$
- $\boldsymbol{\varepsilon}^{**}$: vektor galat berukuran $(n \times 1)$

Menurut Tibshirani (1996) estimasi koefisien LASSO menggunakan pemrograman kuadratik dengan kendala pertidaksamaan. Estimasi LASSO diperoleh dari persamaan berikut:

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg \min \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 \right\}$$

dengan syarat $\sum_{j=1}^k |\beta_j| \leq t$. Nilai t merupakan parameter tuning yang mengontrol penyusutan koefisien LASSO dengan $t \geq 0$.

Menurut Tibshirani (1996), jika $t < t_0$ dengan $t_0 = \sum_{j=1}^p |\hat{\beta}_j|$ maka akan menyebabkan koefisien menyusut mendekati nol atau tepat pada nol atau tepat pada nol, sehingga LASSO akan berperan sebagai seleksi variabel. Akan tetapi jika $t > t_0$ maka penduga koefisien LASSO memberikan hasil yang sama dengan penduga kuadrat terkecil.

Koefisien regresi LASSO ditentukan berdasarkan parameter tuning yang sudah dibakukan dengan $s = \frac{t}{\sum_{j=1}^k |\hat{\beta}_j^0|}$ dengan $t = \sum_{j=1}^p |\hat{\beta}_j| \hat{\beta}_j^0$, adalah penduga kuadrat terkecil untuk model penuh, nilai s optimal diperoleh melalui validasi silang (Dewi, 2010).

2.2 Random Over-Sampling Examples (ROSE)

ROSE (*Random Over-Sampling Examples*) adalah metode *oversampling* yang digunakan untuk menangani masalah ketidakseimbangan kelas pada dataset klasifikasi biner. Metode ini menghasilkan sampel buatan dari kelas minoritas dengan cara mengambil sampel acak dari kelas minoritas dan menempatkan Gaussian di atasnya. Kemudian, sampel baru dihasilkan dengan mensampling Gaussian tersebut. ROSE menggunakan aturan Silverman untuk menentukan *bandwidth* h sehingga *mean integrated squared error* diminimalkan. ROSE dapat digunakan bersamaan dengan teknik undersampling atau dengan teknik oversampling lain seperti SMOTE untuk menghasilkan dataset yang lebih seimbang. ROSE dapat membantu meningkatkan kinerja model klasifikasi pada dataset yang tidak seimbang dengan menghasilkan sampel buatan yang mewakili kelas minoritas.

2.3 Naïve Bayes Classifier (NBC)

Naïve Bayesian Classifier merupakan klasifikasi dengan model statistik untuk menghitung peluang dari suatu kelas yang memiliki masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal. Pada metode ini semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut yang sama penting dan setiap atribut saling bebas satu sama lain (Kusumadewi, 2009).

Dasar dari teorema *Naïve Bayes Classifier* yang dipakai dalam pemrograman adalah rumus Bayes sebagai berikut (Cahyanti et al., 2016):

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Dimana:

$P(H|X)$: probabilitas posterior H di dalam X

$P(X|H)$: probabilitas posterior X di dalam H

$P(H)$: probabilitas prior dari H

$P(X)$: probabilitas prior dari X

2.4 Langkah Analisis

Langkah analisis yang dilakukan dalam penelitian ini adalah sebagai berikut.

1. Mendeskripsikan karakteristik *dataset*. Berdasarkan hasil eksplorasi data terdapat data tidak seimbang pada variabel 'Class'.
2. Melakukan penanganan data tidak seimbang menggunakan metode ROSE.
3. Melakukan analisis prediksi jumlah transaksi dengan kartu kredit pada Amazon dengan tahapan sebagai berikut:
 - a. Membagi data menjadi data *train* sebesar 80% (3.200 data) untuk membentuk model dan data valid sebesar 20% (800 data) untuk memvalidasi model prediksi jumlah transaksi.
 - b. Menentukan variabel target yaitu 'Amount' atau jumlah transaksi dan *feature* seluruh kolom kecuali 'Amount', 'No.' dan 'Class'.
 - c. Membentuk model *Linear Regression*, *SVR*, dan *LASSO Regression* dengan menerapkan *k-cross validation* untuk menghindari terjadi *overfitting*.
 - d. Melakukan evaluasi model prediksi dengan RMSE, *R-squared*, dan MAE untuk memilih model terbaik yang akan digunakan dalam memprediksi data testing.
 - e. Mencari *feature important* dari model prediksi terbaik untuk mengetahui *feature* yang berpengaruh terhadap model.
 - f. Melakukan prediksi data *testing* dengan model prediksi terbaik.
4. Melakukan analisis klasifikasi penipuan dengan kartu kredit pada Amazon dengan tahapan sebagai berikut.
 - a. Mendeskripsikan karakteristik data seimbang setelah dilakukan penanganan dengan ROSE.
 - b. Menentukan variabel target yaitu 'Class' atau kelas (1: terjadi penipuan, 0: tidak terjadi penipuan) dan *feature* seluruh kolom kecuali 'Class' dan 'No.'.
 - c. Membentuk model *Logistic Regression*, *Support Vector Machines*, *Decision Tree Classifier*, dan *Naïve Bayes Classifier*.
 - d. Melakukan evaluasi model klasifikasi dengan *confusion matrix* dan kurva ROC untuk memilih model terbaik yang akan digunakan dalam memprediksi data *testing*.
 - e. Melakukan prediksi data *testing* dengan model prediksi terbaik.
5. Menarik kesimpulan dan saran

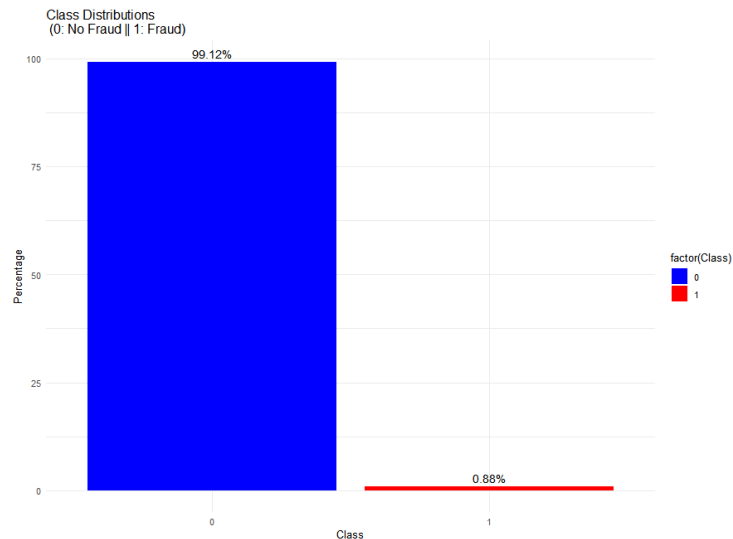
III. HASIL ANALISIS

3.1 Eksplorasi Data

Eksplorasi data dilakukan untuk memahami karakteristik data yang digunakan dalam penelitian.

1. Perbandingan Pelanggan yang Mengalami Penipuan dan Tidak Terjadi Penipuan

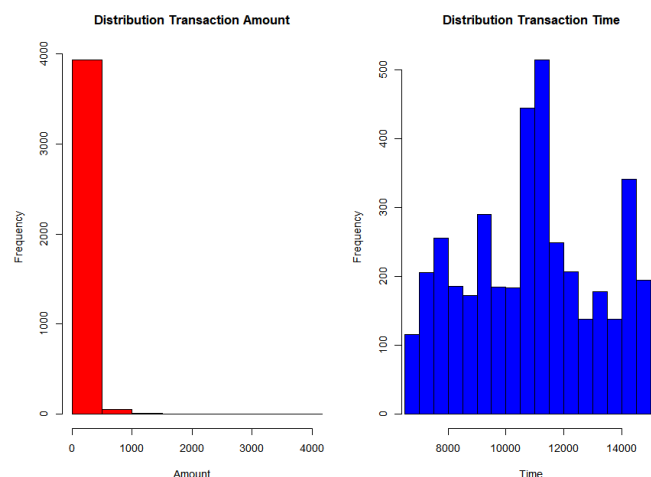
Perbandingan antara pelanggan yang mengalami penipuan dan tidak mengalami penipuan dilakukan pada variabel 'Class' untuk mengidentifikasi apakah distribusi kelas data sudah seimbang sebelum membangun model klasifikasi.



Gambar 3.1 Perbandingan Pelanggan yang Mengalami Penipuan dan Tidak Terjadi Penipuan

Gambar 3.1 menunjukkan bahwa hanya 35 transaksi atau 0,88% yang mengalami penipuan, sedangkan 3965 transaksi atau 99,12% lainnya tidak mengalami penipuan. Hal ini menunjukkan data sangat tidak seimbang pada variabel 'Class', sehingga perlu ditangani sebelum membangun model klasifikasi.

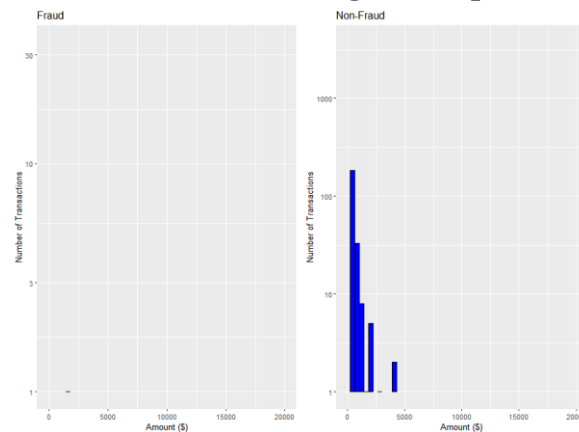
2. Pola Jumlah Transaksi dan Waktu Transaksi



Gambar 3.2 Histogram Jumlah Transaksi dan Waktu Transaksi

Histogram jumlah transaksi menunjukkan sebagai besar pelanggan melakukan transaksi dibawah \$500 transaksi dengan total pelanggan lebih dari 3.800 pelanggan, serta terdapat data *outlier* atas. Rata-rata waktu transaksi terjadi pada \$10.000 sampai \$12.000.

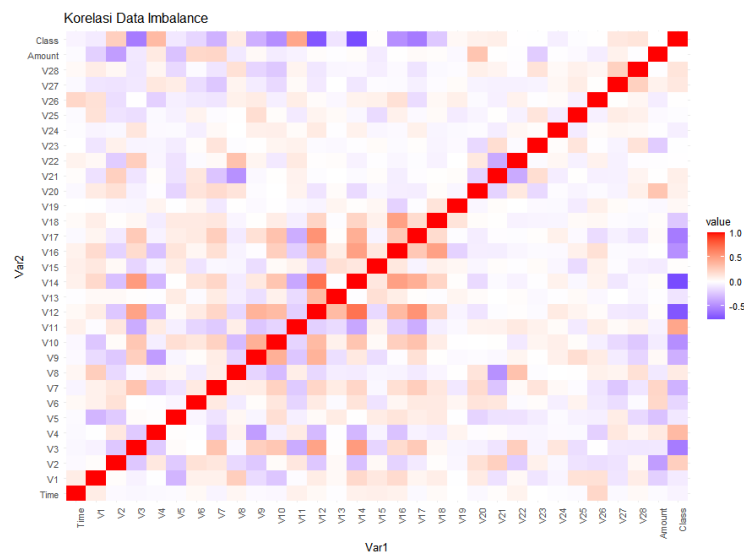
3. Pola Jumlah Transaksi berdasarkan Kategori Penipuan dan Tidak Penipuan



Gambar 3.3 Histogram Jumlah Transaksi Berdasarkan Penipuan dan Tidak Penipuan

Histogram jumlah transaksi pada kelas Fraud atau terjadi penipuan tidak dapat diidentifikasi dengan jelas karena memiliki jumlah data yang sedikit dibandingkan dengan Non-Fraud, pada kelas Fraud memiliki rata-rata jumlah transaksi sebesar \$60,25 serta memiliki *outlier* atas. Histogram jumlah transaksi pada kelas Non-Fraud atau tidak terjadi penipuan memiliki rata-rata sebesar \$59.76 serta memiliki *outlier* atas dan bawah.

4. Korelasi antar Variabel



Gambar 3.4 Korelasi antar Variabel pada Data Tidak Seimbang

Korelasi antar variabel pada data tidak seimbang tidak memiliki hubungan yang kuat antar variabelnya. Variabel yang memiliki hubungan yang kuat dimana nilai korelasinya <-0.6 dan >0.6 dengan variabel 'Class' hanya variabel V12 dan V14, sedangkan pada variabel 'Amount' atau jumlah transaksi tidak ada.

3.2 Prediksi Jumlah Transaksi

Pada dataset yang diberikan sebesar 4.000 pengamatan, peneliti membagi menjadi 80% data train dan 20% data validation. Prediksi jumlah transaksi dibangun dengan 3 model yaitu *Linear Regression*, *Support Vector Regression*, dan *LASSO Regression*.

1. Perbandingan Model Prediksi

Pada setiap model yang dibangun diterapkan *k-cross validation* untuk menghindari adanya *overfitting* pada model, dengan evaluasi sebagai berikut.

Tabel 3.1 Perbandingan Model Prediksi

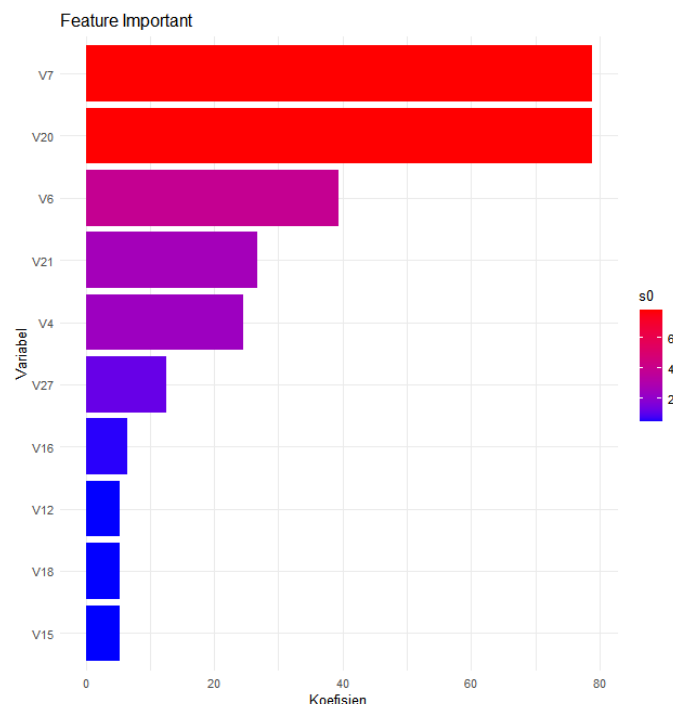
Model	Data	RMSE	R-squared	MAE
Linear Regression	Train	31,678	96,18%	12,022
	Valid	44,823	95,13%	12,517
Support Vector Regression – Linear	Train	35,739	95,36%	11,201
	Valid	25,376	98,47%	10,859
LASSO Regression	Train	30,449	96,49%	10,840
	Valid	46,742	94,70%	11,924

Tabel 3.1 menunjukkan bahwa pada data train model yang memiliki evaluasi model terbaik yang diukur berdasarkan nilai RMSE terendah, R-square tertinggi, dan MAE terendah yaitu model *LASSO Regression*. Jika data validation pada model SVR-Linear, namun dapat dilihat model SVR-Linear kurang konsisten sehingga mungkin terjadi *overfitting* jika diterapkan pada data test. Sehingga model terbaik yang dipilih adalah model yang konsisten yaitu *LASSO Regression*.

2. Model Prediksi Terbaik

Model prediksi terbaik berdasarkan sub bab sebelumnya dengan menggunakan *LASSO Regression*, sehingga dilakukan analisis lebih lanjut terkait model tersebut. Pada model diperoleh variabel terpenting atau *features important*. *Features important* digunakan dalam konteks *machine learning* untuk mengukur sejauh mana setiap fitur (variabel atau atribut) dalam data berkontribusi terhadap kemampuan model untuk membuat prediksi yang akurat.

Berikut adalah *features important* dari model *LASSO Regression* yang ditampilkan dalam bentuk *bar chart*.

**Gambar 3.5** Feature Important Model *LASSO Regression*

Gambar 3.5 menunjukkan bahwa variabel yang memiliki kontribusi atau pengaruh paling besar terhadap jumlah transaksi sebagai variabel respon adalah V7 dan V20, diikuti dengan variabel lainnya seperti pada *chart* yang tertera diatas.

Model prediksi *LASSO Regression* dilakukan prediksi pada dataset yang berjumlah 1.000 pengamatan baru, menghasilkan evaluasi model sebagai berikut.

Tabel 3.2 Evaluasi Model *LASSO Regression* Data Test

Model	RMSE	R-squared	MAE
<i>LASSO Regression</i>	55,056	87,969%	14,635

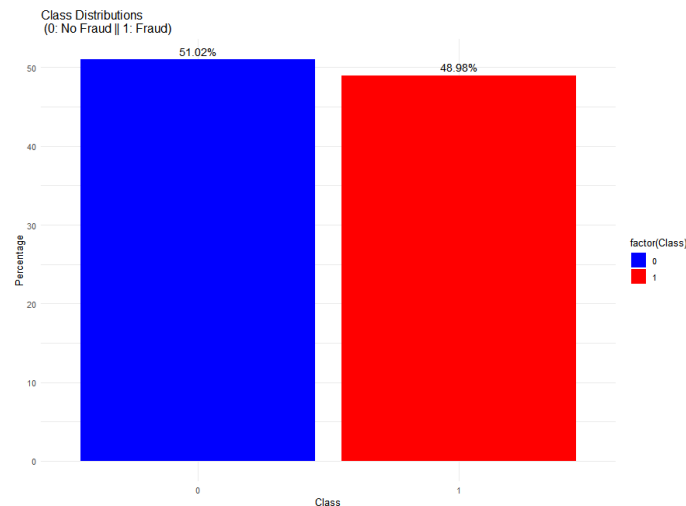
Tabel 3.2 menunjukkan bahwa nilai RMSE sebesar 55,056 berarti secara rata-rata, perbedaan antara nilai prediksi dan nilai sebenarnya adalah sekitar 55,056. Semakin kecil nilai RMSE, semakin baik model dalam memprediksi data. Nilai R-squared sebesar 0,8797 berarti model mampu menjelaskan sekitar 87,97% variasi jumlah transaksi (variabel respon). Nilai MAE sebesar 14,635 berarti, secara rata-rata, perbedaan *absolut* antara nilai prediksi dan nilai sebenarnya adalah sekitar 14,635. Hal tersebut menunjukkan bahwa model *LASSO Regression* memiliki performa yang baik, dengan RMSE dan MAE yang relatif rendah, serta nilai R-squared yang cukup tinggi, yang menunjukkan bahwa model mampu menjelaskan sebagian besar variasi dalam data target dan tidak terjadi *overfitting*.

3.3 Klasifikasi Penipuan

Klasifikasi penipuan pada kartu kredit dilakukan dengan membangun model *machine learning* seperti *Logistic Regression*, *Support Vector Machines*, *Decision Tree Classifier*, dan *Naïve Bayes Classifier*. Pada sub bab 3.1 (1) variabel ‘Class’ sebagai variabel respon dengan kategori 0 (*Non-Fraud*) dan 1 (*Fraud*) memiliki data yang tidak seimbang, sehingga perlu dilakukan penanganan untuk agar tidak menghasilkan model yang bias. Penanganan data tidak seimbang dilakukan dengan metode *Random Over-Sampling Examples* (ROSE).

1. Karakteristik Data Setelah Penanganan Data Tidak Seimbang

Berikut perbandingan kategori pada ‘Class’ setelah dilakukan penanganan data tidak seimbang dengan ROSE.

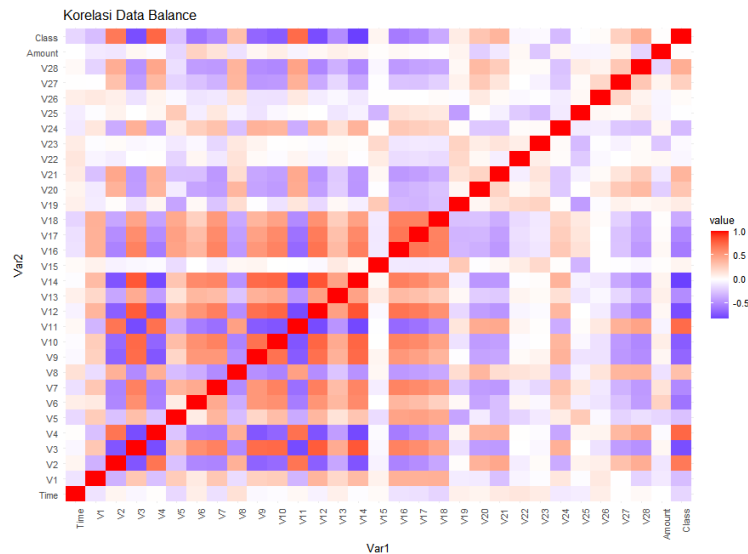


Gambar 3.6 Perbandingan Pelanggan yang Mengalami Penipuan dan Tidak Terjadi Penipuan Setelah Penanganan Data Tidak Seimbang

Gambar 3.6 menunjukkan bahwa 1.959 transaksi atau 48,98% yang mengalami penipuan, sedangkan 2.041 transaksi atau 51,02% lainnya tidak mengalami penipuan. Hal ini menunjukkan data seimbang pada variabel ‘Class’, sehingga dapat dilanjutkan untuk membangun model klasifikasi. Perlu diketahui dari matriks korelasi apakah setelah dilakukan

penanganan data tidak seimbang, variabel yang memiliki hubungan kuat dengan variabel ‘Class’ meningkat.

Berikut matriks korelasi setelah dilakukan penanganan data tidak seimbang dengan ROSE.



Gambar 3.7 Korelasi antar Variabel pada Data Seimbang

Korelasi antar variabel pada data seimbang mengalami peningkatan nilai korelasi atau meningkatkan hubungan yang kuat antar variabelnya. Variabel yang memiliki hubungan yang kuat dimana nilai korelasinya $<-0,6$ dan $>0,6$ dengan variabel ‘Class’ variabel V2, V3, V4, V6, V9, V10, V11, V12 dan V14.

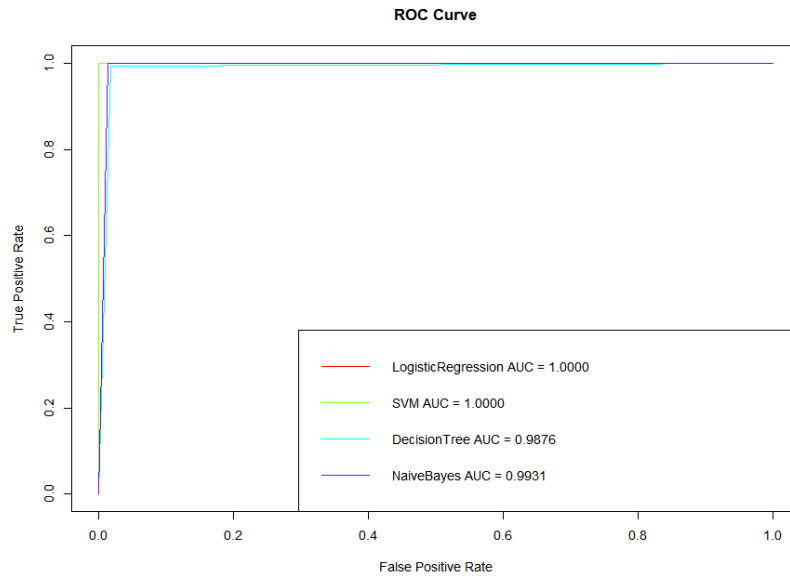
2. Perbandingan Model Klasifikasi

Tabel 3.3 Perbandingan Model Klasifikasi

Model	Accuracy	Sensitivity	Specificity	F1-score	Run-Time
<i>Logistic Regression</i>	1,000	1,000	1,000	1,000	0,5 s
<i>Support Vector Machine – Linear</i>	0,9985	0,999	0,998	0,9985	1,44 s
<i>Support Vector Machine – RBF</i>	1,000	1,000	1,000	1,000	5,45 s
<i>Decision Tree</i>	0,9875	0,9823	0,9926	0,9873	0,5 s
<i>Naïve Bayes Classifier</i>	0,993	0,9859	1,000	0,9929	7,46 s

Berdasarkan model yang sudah dibangun dengan 5 model di atas dapat diketahui berdasarkan evaluasi model yaitu *accuracy*, *sensitivity*, *specificity*, *F1-score* yang memiliki nilai tertinggi yaitu model dengan *Logistic Regression* dan *SVM-RBF* dengan nilai masing-masing 1,000 menunjukkan bahwa model memiliki kinerja yang sangat baik dalam melakukan klasifikasi. Namun jika dilihat dari *run-time* atau waktu dalam membangun model yang memiliki waktu tercepat yaitu *Logistic Regression* dengan waktu 0,5 detik. Waktu eksekusi yang cepat (0,5 detik) menunjukkan efisiensi dalam membangun model.

Kurva ROC digunakan untuk mengukur sejauh mana model klasifikasi mampu membedakan antara dua kelas target dengan benar, berikut adalah kurva ROC dari setiap model klasifikasi.

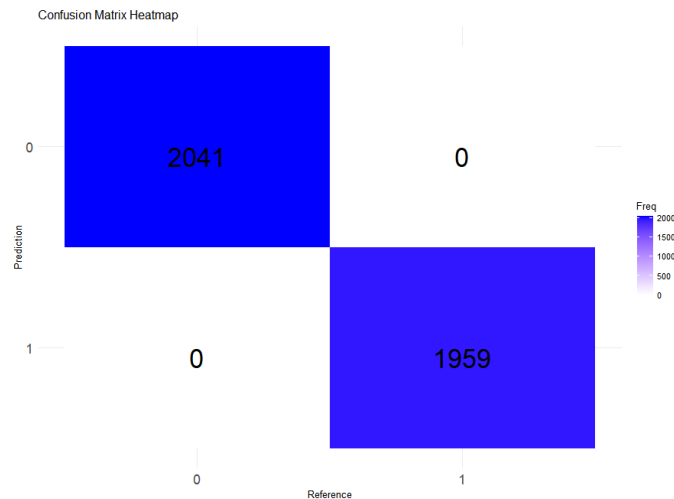


Gambar 3.8 Kurva ROC Setiap Model Klasifikasi

Kurva ROC pada Gambar 3.8 memiliki empat garis, masing-masing mewakili model yang berbeda: *Logistic Regression*, *SVM*, *Decision Tree*, dan *Naïve Bayes Classifier*. Model *Logistic Regression* dan *SVM* memiliki AUC (*Area Under the Curve*) sebesar 1.0000, model *Decision Tree* memiliki AUC sebesar 0.9876, dan model *Naïve Bayes Classifier* memiliki AUC sebesar 0.9931. Garis untuk model *Logistic Regression* dan *SVM* hampir vertikal, menunjukkan tingkat *True Positive* yang tinggi dan tingkat *False Positive* yang rendah. Garis untuk model *Decision Tree* dan *Naïve Bayes Classifier* sedikit kurang vertikal, menunjukkan tingkat *True Positive* yang sedikit lebih rendah dan tingkat *False Positive* yang sedikit lebih tinggi. *Logistic Regression* adalah model yang paling baik dalam memprediksi kelas positif. *SVM* juga memiliki kinerja yang baik, tetapi tidak sebaik *Logistic Regression*. Sehingga model klasifikasi terbaik pada penelitian ini adalah *Logistic Regression*.

3. Model Klasifikasi Terbaik

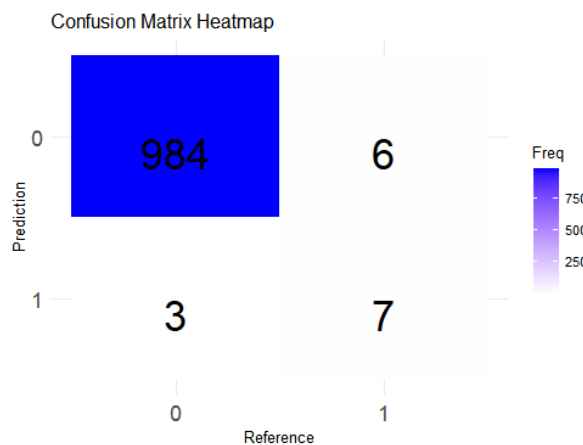
Model klasifikasi terbaik berdasarkan sub bab sebelumnya dengan menggunakan *Logistic Regression*, sehingga dilakukan analisis lebih lanjut terkait model tersebut. *Confusion matrix* digunakan untuk mengevaluasi kinerja model klasifikasi. *Confusion matrix* menunjukkan jumlah prediksi yang benar dan salah yang dibuat oleh model dibandingkan dengan klasifikasi aktual dalam set pengujian. *Confusion matrix* terdiri dari empat istilah yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Berikut adalah *Confusion matrix* model *Logistic Regression*.



Gambar 3.9 *Confusion Matrix Logistic Regression*

Heatmap confusion matrix Logistic Regression menunjukkan bahwa *True Positive* (TP) atau terdapat 1.959 sampel yang diprediksi sebagai 1 dan sebenarnya juga termasuk dalam kelas 1. *True Negative* (TN) atau terdapat 2.041 sampel yang diprediksi sebagai 0 dan sebenarnya juga termasuk dalam kelas 0. Pada model ini tidak ada *false positives* (FP = 0) dan tidak ada *false negatives* (FN = 0). Sehingga model klasifikasi memiliki kinerja yang sangat baik dalam membedakan antara kelas 0 dan kelas 1. Ini menunjukkan bahwa tidak ada kesalahan prediksi positif (*false positives*) atau negatif (*false negatives*) yang dibuat oleh model, dan model ini memprediksi dengan benar semua sampel kelas 0 dan kelas 1.

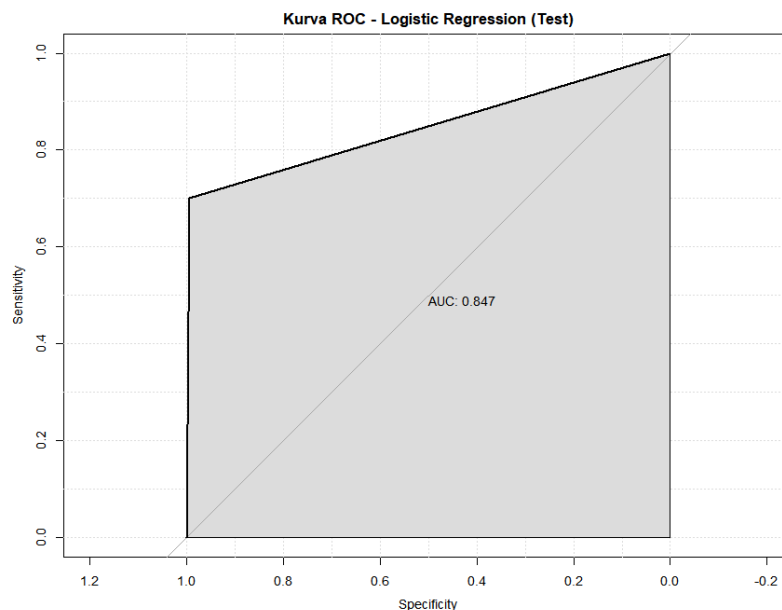
Model klasifikasi *Logistic Regression* dilakukan prediksi pada dataset yang berjumlah 1.000 pengamatan baru, menghasilkan *confusion matrix* sebagai berikut.



Gambar 3.10 *Confusion Matrix Logistic Regression*

Heatmap confusion matrix Logistic Regression pada 1.000 pengamatan pada data test menunjukkan bahwa model memprediksi 984 data kelas 0 dengan benar dan 7 data kelas 1 dengan benar, serta model memprediksi 5 data kelas 0 sebagai kelas 1 dan 8 data kelas 1 sebagai kelas 0. Model memiliki kinerja yang baik dalam memprediksi kelas 0. Model juga dapat memprediksi kelas 1 dengan cukup akurat, tetapi terdapat beberapa data kelas 1 yang salah diprediksi sebagai kelas 0. Akurasi model adalah 0,991, yang berarti model ini dapat memprediksi kelas dengan benar sebesar 99,1%. *Sensitivity* model sebesar 0,5385, yang berarti model ini dapat memprediksi 53,85% dari data kelas positif dengan benar. *Specificity* model sebesar 0,997, yang berarti model ini dapat memprediksi 99,7% dari data kelas negatif dengan

benar. Model memiliki akurasi yang tinggi dan spesifisitas yang baik, tetapi sensitivitasnya relatif rendah, yang berarti model mungkin kurang dalam mengidentifikasi kelas positif. *F1-score* yang cukup baik menunjukkan bahwa model mampu mencapai keseimbangan antara presisi dan recall.



Gambar 3.11 Kurva ROC *Logistic Regression*

Model ini memiliki kurva ROC yang lebih dekat dengan sudut kanan atas. Ini berarti model ini memiliki kinerja yang lebih baik dalam memprediksi kelas positif. Ditunjukkan dengan skor AUC sebesar 0,847 bahwa model memiliki kemampuan yang baik dalam memprediksi kelas dengan benar.

IV. KESIMPULAN DAN SARAN

1. Kesimpulan

Terjadi data tidak seimbang sehingga menyebabkan model cenderung memprediksi kelas mayoritas, yang akhirnya menghasilkan model yang tidak baik dalam mengidentifikasi kelas minoritas, pada penelitian ini dilakukan penanganan dengan ROSE. Model klasifikasi terjadinya penipuan transaksi dengan kartu kredit terbaik adalah *Logistic Regression* dengan akurasi sebesar 99,1%. Model prediksi jumlah transaksi di *e-commerce* Amazon adalah *LASSO Regression* dengan RMSE sebesar 55,056, *R-squared* sebesar 87,969%, dan MAE sebesar 14,635. Jika model dilihat berdasarkan perbandingan MAE pada data train, valid, dan test tidak terjadi *overfitting*.

2. Saran

Diharapkan penelitian selanjutnya dapat menerapkan metode data tidak seimbang selain ROSE atau SMOTE sehingga dapat menghasilkan model yang baik dalam mengidentifikasi kelas minoritas pada model klasifikasi.

DAFTAR PUSTAKA

- Cahyanti, A. F., Saptono, R., & Sihwi, S. W. (2016). Penentuan Model Terbaik pada Metode Naive Bayes Classifier dalam Menentukan Status Gizi Balita dengan Mempertimbangkan Independensi Parameter. *Jurnal Teknologi & Informasi ITSmart*, 4(1), 28. <https://doi.org/10.20961/its.v4i1.1754>
- Dewi, Y. S. (2010). OLS, LASSO dan PLS Pada data Mengandung Multikolinearitas. *Jurnal ILMU DASAR*, 11(1), 83–91.
- Hania, A. A. (2017). Mengenal Artificial Intelligence, Machine Learning, dan Deep Learning. *Jurnal Teknologi Indonesia*, 1(June), 1–6. <https://amt-it.com/mengenal-perbedaan-artificial-intelligence-machine-learning-deep-learning/>
- Kusumadewi, S. (2009). Klasifikasi Status Gizi Menggunakan Naive Bayesian Classification. *Communication and Information Technology Journal*, 3(1), 6–11.
- Patriya, E. (2020). Implementasi Support Vector Machine Pada Prediksi Harga Saham Gabungan (Ihsg). *Jurnal Ilmiah Teknologi Dan Rekayasa*, 25(1), 24–38. <https://doi.org/10.35760/tr.2020.v25i1.2571>
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288. [jstor.org/stable/2346178](https://www.jstor.org/stable/2346178)