Laporan Evaluasi Tengah Semester
Pembelajaran Mesin

Nama

Yuniar Ayu Rachmadini

NRP

2043201103

Dosen

Mukti Ratna Dewi S.Si, M.Sc

BAB I
LATAR BELAKANG

1.1 Latar Belakang

Amazon merupakan salah satu perusahaan multinasional terkemuka di dunia yang didirikan pada tahun 1994 oleh Jeff Bezos di Seattle, Amerika Serikat. Awalnya, Amazon adalah sebuah perusahaan yang berfokus pada penjualan buku secara daring (online), namun seiring berjalannya waktu, perusahaan ini mengalami pertumbuhan yang luar biasa dan bertransformasi menjadi salah satu entitas bisnis paling berpengaruh di berbagai sektor. Saat ini, Amazon merupakan salah satu pemain utama dalam industri e-Commerce global, yang menawarkan berbagai produk dan layanan dari sejumlah mitra dan penjual independen. Hingga saat ini, Amazon telah menjalani perubahan besar dalam upaya untuk memenuhi kebutuhan dan ekspektasi pelanggan. Penggunaan teknologi dan analisis data yang canggih telah menjadi salah satu pilar utama yang digunakan. Amazon mengelola sekitar 1.000.000.000 gigabyte data di lebih dari 1.400.000 server, dan inovasi terus-menerus dalam ilmu data dan big data telah memungkinkan Amazon untuk lebih memahami pelanggan, memberikan pengalaman belanja yang lebih baik, dan meningkatkan efisiensi operasional. Namun, sebagai e-Commerce Amazon memiliki risiko tinggi terhadap penipuan ritel. Sebagai tindakan pencegahan, perusahaan mengumpulkan data historis dan real-time untuk setiap pesanan. Dengan menggunakan algoritma pembelajaran mesin, perusahaan mampu untuk menemukan transaksi dengan kemungkinan penipuan (fraud) yang lebih tinggi. Tindakan ini telah membantu perusahaan untuk mengurangi jumlah pengembalian produk.

Algoritma pembelajaran mesin adalah cabang dari kecerdasan buatan (AI) yang berfokus pada pengembangan algoritma komputer yang memungkinkan sistem untuk belajar dari data, mengidentifikasi pola, dan membuat keputusan atau prediksi tanpa perlu diprogram secara eksplisit. Dalam machine learning, algoritma-algoritma ini menggunakan data pelatihan untuk meningkatkan kinerja mereka seiring berjalannya waktu, memungkinkan sistem untuk beradaptasi dengan perubahan dan menghasilkan hasil yang lebih akurat. Pembentukan model prediksi dapat menggunakan Regresi, Support Vector Regression, Random Forest, dan Neural Network. Sedangkan pembentukan model klasifikasi dapat menggunakan algoritma seperti Support Vector Machines, Random Forests, Neural Networks, dan Naïve Bayes untuk melatih model agar dapat mengidentifikasi pola yang mewakili perbedaan antara kelas yang berbeda.

Pada penelitian ini akan dilakukan analisis terhadap data laporan penjualan dan laporan jumlah penipuan transaksi yang terjadi di Amazon, peneliti akan melakukan pembentukan model prediksi menggunakan tiga metode yaitu regresi linear berganda, support vector regression, dan random forest. Serta pembentukan model klasifikasi menggunakan empat metode yaitu regresi logistik biner, support vector machine, decision tree, dan naïve bayes. Metode-metode tersebut akan dibandingkan sehingga dapat diketahui metode yang dapat membentuk model terbaik.

BAB II
METODOLOGI

Pada bab ini akan dijelaskan terkait pengertian dari metode pilihan yaitu Random Forest Regression dan Naïve Bayes Classifier, kemudian akan diuraikan langkah dalam melakukan analisis secara berurutan.

2.1 Random Forest Regression

Algoritma *random forest* pertama kali 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. Sedangkan *random forest regression* merupakan algoritma pembelajaran mesin dengan menggunakan kombinasi beberapa pohon keputusan acak yang masing-masing dilatih pada subset data. Metode ini merupakan sebuah *ensemble* (kumpulan) metode pembelajaran menggunakan *decision tree* sebagai *base classifier* yang dibangun dan dikombinasikan (Britanthia et al., 2020). Penggunaan banyak pohon memberikan stabilitas pada algoritma dan mengurangi varians atau dengan kata lain semakin banyak jumlah pohon maka akan menghasilkan akurasi yang lebih tinggi dan mencegah masalah overfitting. Algoritma *random forest regression* adalah model yang umum digunakan karena kemampuannya bekerja dengan baik untuk jenis data yang besar.

Random Forest bekerja dalam dua fase. Fase pertama yaitu menggabungkan sejumlah N decision tree untuk membuat Random Forest. Kemudian fase kedua adalah membuat prediksi untuk setiap tree yang dibuat pada fase pertama. Adapun tiga aspek penting dalam metode *random forest*, yaitu: (1) melakukan *bootstrap* sampling untuk membangun pohon prediksi; (2) masing-masing pohon keputusan memprediksi dengan prediktor acak; (3) lalu *random forest* melakukan prediksi dengan mengombinasikan hasil dari setiap pohon keputusan dengan cara *majority vote* untuk klasifikasi atau rata-rata untuk regresi.

2.2 Naïve Bayes Classifier

Pengklasifikasi bayes merupakan salah satu pengklasifikasi statistik, dimana pengklasifikasi ini dapat memprediksi probabilitas keanggotaan kelas suatu data yang akan masuk ke dalam kelas tertentu, sesuai dengan perhitungan probabilitas. Pengklasifikasi Bayes didasari oleh teorema bayes yang ditemukan oleh Thomas Bayes pada abad ke-18. Teorema Bayes adalah rumus matematika yang digunakan untuk menghitung probabilitas suatu peristiwa berdasarkan informasi yang berkaitan dengan peristiwa tersebut. Dalam konteks klasifikasi, kita ingin menghitung probabilitas bahwa suatu observasi (data) termasuk dalam suatu kelas tertentu berdasarkan atribut-atribut yang diamati (Susana & Suarna, 2022).

Dalam studi pembandingan algoritma klasifikasi telah ditemukan simple bayesian atau yang biasa dikenal dengan Naïve Bayes classifier. Asumsi "naif" dalam Naive Bayes adalah bahwa semua atribut dalam data adalah independen satu sama lain, yaitu, atribut Y dalam contoh di atas tidak memiliki ketergantungan. Naïve Bayes classifier menunjukkan akurasi dan kecepatan yang tinggi bila diterapkan pada database yang besar (Rachman & Handayani, 2021).

Metode ini sering digunakan dalam menyelesaikan masalah dalam bidang mesin pembelajaran karena metode ini dikena memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana.

2.3 Langkah Analisis

Langkah analisis dalam mencari model prediksi terbaik yang dilakukan dalam penelitian ini dijelaskan sebagai berikut.

1. Explorasi data dengan melakukan uji korelasi antara variabel respon dan masing-masing variabel prediktor
2. Mengeluarkan variabel yang tidak signifikan atau memiliki korelasi rendah
3. Membagi data menjadi data training dan data validasi
4. Membentuk model regresi berganda, support vector regression (SVR), dan random forest menggunakan data training
5. Melakukan prediksi dari model yang terbentuk menggunakan data validation
6. Menghitung RMSE dari masing – masing model prediksi dengan langkah sebagai berikut.
    a. Untuk setiap pengamatan dalam dataset, hitung selisih antara nilai sebenarnya (y) dan nilai prediksi (ŷ) dari model. Ini disebut residu (error atau kesalahan).
    b. Kuadrat setiap residu untuk menghilangkan nilai negatif dan menekankan kesalahan yang lebih besar
    c. Hitung rata-rata dari semua residu yang telah diubah menjadi kuadrat. Ini disebut Mean Squared Error (MSE)
    d. Akar kuadrat dari MSE untuk mendapatkan RMSE
7. Membandingkan nilai RMSE dari masing-masing model
8. Menguji model prediksi terbaik menggunakan data testing

Langkah analisis dalam mencari model klasifikasi dengan tingkat akurasi tertinggi yang dilakukan dalam penelitian ini dijelaskan sebagai berikut.

1. Explorasi data dengan melakukan uji korelasi dan pemeriksaan proporsi data
2. Mengeluarkan variabel yang tidak signifikan atau memiliki korelasi rendah
3. Menangani imbalance data
4. Membagi data menjadi data training dan data validasi
5. Membentuk model regresi logistik biner, support vector machine (SVM), decision tree, dan naïve bayes classifier menggunakan data training
6. Melakukan prediksi dari model yang terbentuk menggunakan data validasi
7. Menghitung *accuracy*, *sensitivity*, dan *specificity* dari masing – masing model klasifikasi menggunakan *confusion matrix.*
8. Membandingkan *accuracy* dari masing-masing model
9. Menguji model dengan *accuracy* tertinggi menggunakan data testing

BAB III
HASIL ANALISIS

Pada bab ini, akan diuraikan hasil analisis mulai dari eksplorasi data hingga penentuan model prediksi dan model klasifikasi terbaik untuk kemudian diuji dengan data testing.

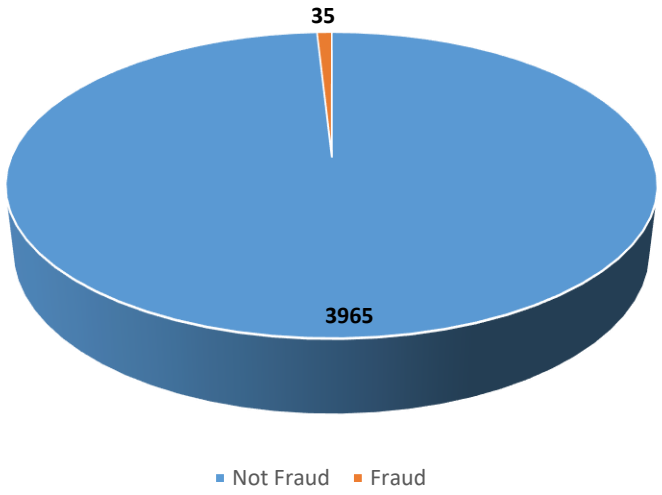
3.1 Eksplorasi Data

Uji korelasi pearson dilakukan untuk mengetahui hubungan antara variabel respon dengan masing – masing variabel prediktor. Hasil uji korelasi dituliskan dalam tabel sebagai berikut.

Variabel	**Pearson Correlation**	**T-test**	**P-*Value***
Y – V1	-0.199	-12.86	2.2e-16
Y – V2	-0.420	-29.31	2.2e-16
Y – V3	-0.097	-6.19	6.62e-10
Y – V4	0.111	7.04	2.235e-12
Y – V5	-0.262	-17.17	2.2e-16
Y – V6	0.206	13.35	2.2e-16
Y – V7	0.215	13.91	2.2e-16
Y – V8	-0.095	-6.01	1.981e-09
Y – V9	0.031	1.95	**0.051**
Y – V10	-0.082	-5.21	2e-07
Y – V11	-0.055	-3.46	0.00053
Y – V12	0.016	1.02	**0.3073**
Y – V13	-0.009	-0.59	**0.5522**
Y – V14	0.018	1.16	**0.2449**
Y – V15	-0.085	- 5.41	6.594e-08
Y – V16	0.003	0.17	**0.867**
Y – V17	0.024	1.51	**0.1311**
Y – V18	0.015	0.92	**0.3552**
Y – V19	-0.028	-1.79	**0.0721**
Y – V20	0.305	20.24	2.2e-16
Y – V21	0.004	0.25	**0.8049**
Y – V22	-0.007	-0.43	**0.6689**
Y – V23	-0.216	-14.02	2.2e-16
Y – V24	-0.007	-0.43	**0.6673**
Y – V25	-0.026	-1.63	**0.1041**
Y – V26	-0.076	-4.83	1.41e-06
Y – V27	0.068	4.32	1.627e-05
Y – V28	-0.008	-0.4995	**0.6175**

Dari tabel diatas diketahui bahwa ada 13 variabel yang memiliki nilai p-*value* > 0.05, artinya variabel-variabel tersebut tidak memiliki pengaruh yang signifikan terhadap variabel respon. Sehingga dalam proses analisis hanya digunakan 15 variabel yang signifikan.

Pemeriksaan proporsi atau jumlah antara data fraud dan tidak fraud penting untuk dilakukan, karena jumlah data yang tidak seimbang dapat mempengaruhi hasil akhir dari analisis. Berikut adalah visualisasi dari jumlah data yang fraud dan tidak fraud.



Gambar diatas menunjukkan bahwa terdapat indikasi *imbalance data* karena selisih antara data yang fraud dan tidak fraud sangat jauh yaitu 35:3965. Oleh karena itu, perlu dilakukan penanganan lebih lanjut untuk menyeimbangkan jumlah data sebelum dilakukan analisis khusus nya pada pembentukan model klasifikasi.

3.2 Analisis Model Prediksi

Pada sub ini, dilakukan analisis menggunakan tiga metode regresi yang berbeda yaitu regresi linear berganda, support vector regression (SVR), dan random forest regression. Tujuan dari analisis ini adalah untuk memahami kinerja, kelebihan, dan kekurangan masing-masing metode dalam konteks prediksi data numerik yang kemudian dibandingkan untuk menentukan model prediksi terbaik.

3.2.1 Membentuk dan Menentukan Model Prediksi Terbaik

Peneliti membentuk model prediksi menggunakan 3 metode yaitu regresi linear berganda, support vector regression (SVR), dan random forest dengan menggunakan data yang sudah dibagi menjadi data training dan data validation menggunakan perbandingan 75:25. Ketiga model tersebut akan dibandingkan berdasarkan nilai RMSE, dimana semakin kecil nilai RMSE maka model semakin baik.

Metode	RMSE
Regresi Linear Berganda	35,34%
Support Vector Regression	14,69%
Random Forest	21,74%

Tabel diatas menunjukkan bahwa model support vector regression menghasilkan nilai RMSE terkecil yaitu sebesar 14,69%, sehingga model tersebut dapat dikatakan sebagai model terbaik. Kemudian model terpilih akan diuji menggunakan data testing.

3.2.2 Menguji Model Prediksi Terbaik

Pengujian model prediksi terbaik support vector regression menggunakan data testing dilakukan untuk membandingkan data aktual dengan data hasil prediksi. Sehingga diperoleh hasil sebagai berikut.

Metode	RMSE
Support Vector Regression	65,12%

Tabel diatas menunjukkan nilai RMSE dari model SVR menggunakan data testing yaitu sebesar 65,12%, artinya model memiliki tingkat error yang cukup tinggi.

3.3 Analisis Model Klasifikasi

Pada sub ini, dilakukan analisis menggunakan empat metode klasifikasi yang berbeda yaitu regresi logistik biner, support vector machine (SVM), desicion tree, dan naïve bayes classifier. Tujuan dari analisis ini adalah untuk memahami kinerja, kelebihan, dan kekurangan masing-masing metode dalam konteks klasifikasi yang kemudian dibandingkan untuk menentukan model klasifikasi terbaik.

3.3.1 Penanganan Imbalance Data

Pada analisis model klasifikasi dilakukan penanganan *imbalance data* agar model yang terbentuk dapat menghasilkan nilai akurasi yang tinggi dan mengurangi bias yang mungkin muncul karena ada ketidakseimbangan data.

Histogram diatas menunjukkan variabel Class yang terdiri dari dua kategori, yaitu not fraud dan fraud ketika telah dilakukan penanganan imbalanced data. Sehingga diperoleh hasil akhir jumlah data not fraud sebanyak 3965 data, sedangkan jumlah data fraud sebanyak 3926 data. Sehingga dari hasil tersebut dapat disimpulkan bahwa data sudah *balance*. Data tersebut kemudian dibagi menjadi data training dan data validation dengan perbandingan 75:25.

3.3.2 Membentuk dan Menentukan Model Klasifikasi Terbaik

Peneliti membentuk model klasifikasi menggunakan 4 metode yaitu regresi logistik biner, support vector machine (SVM), decision tree, dan naïve bayes. Keempat model tersebut akan dibandingkan berdasarkan nilai *accuracy*, *sensitivity*, dan, *spesificity* dengan hasil sebagai berikut

Metode	Accuracy	Sensitivity	Spesificity
Regresi Logistik Biner	100%	100%	100%
Support Vector Machine	100%	100%	100%
Decision Tree	99,95%	99,90%	100%
Naïve Bayes	99,44%	100%	98,89%

Tabel diatas menunjukkan bahwa model regresi logistik biner dan support vector mechine menghasilkan nilai akurasi tertinggi yaitu sebesar 100%, sehingga kedua model tersebut dapat dikatakan sebagai model terbaik. Kemudian kedua model tersebut akan diuji menggunakan data testing.

3.3.3 Menguji Model Klasifikasi Terbaik

Pengujian model klasifikasi terbaik regresi logistik biner dan support vector mechine menggunakan data testing dilakukan untuk menguji tingkat akurasi. Sehingga diperoleh hasil sebagai berikut.

Metode	Accuracy	Sensitivity	Spesificity
Regresi Logistik Biner	99%	99,29%	70%
Support Vector Machine	98,7%	99,49%	20%

Tabel diatas menunjukkan bahwa nilai akurasi dari model regresi logistik biner lebih tinggi dibandingkan model SVM yaitu sebesar 99%, sehingga dapat disimpulkan bahwa metode regresi logistik biner dapat membentuk model klasifikasi terbaik dengan tingkat akurasi yang mendekati sempurna.

BAB IV
KESIMPULAN & SARAN

4.1 Kesimpulan

Kesimpulan yang diperoleh dari hasil analisis adalah sebagai berikut.

1. Explorasi data menunjukkan bahwa terdapat 13 variabel prediktor yang memiliki korelasi rendah terhadap variabel respon atau tidak berpengaruh signifikan terhadap variabel respon, serta diketahui bahwa data *imbalance*
2. Metode support vector regression (SVR) menghasilkan model prediksi terbaik dengan nilai RMSE terkecil
3. Metode regresi logistik biner menghasilkan model klasifikasi terbaik dengan nilai akurasi tertinggi

4.2 Saran

Saran untuk penelitian selanjutnya yaitu dapat melakukan eksplorasi data lebih dalam dan mencoba metode penanganan *imbalance data* yang lebih baik sehingga dapat dilakukan prediksi dan klasifikasi model yang lebih optimal dan akurat.

DAFTAR PUSTAKA

Britanthia, L., Tanujaya, C., Susanto, B., & Saragih, A. (2020). *Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Fitur Mode Audio Spotify*. *1*(3), 68–78.

Rachman, R., & Handayani, R. N. (2021). *Klasifikasi Algoritma Naive Bayes Dalam Memprediksi Tingkat Kelancaran Pembayaran Sewa Teras UMKM*. *8*(2), 111–122.

Susana, H., & Suarna, N. (2022). *PENERAPAN MODEL KLASIFIKASI METODE NAIVE BAYES*. *4*(1), 2–9.