Isolation Forest dengan Exploratory Data Analysis pada Anomaly Detection untuk Data Transaksi

I Made Sudarsana Taksa Wibawa[a1], Anak Agung Istri Ngurah Eka Karyawati [a2]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Udayana
Jalan Raya Kampus Udayana, Bukit Jimbaran, Kuta Selatan, Badung, Bali Indonesia
[1]taksawibawa@gmail.com
[2]eka.karyawati@unud.ac.id

Abstract

Managing value of data is one of the key aspects of presenting analysis for decision making support in various cases. One of such method is by managing detecting anomaly in the data. This research focuses on implementing Isolation Forest result of anomaly detection. This method is used on transaction dataset from Kaggle with about more than 500.000 records. The result this research shows that Isolation Forest used in the dataset have 0.899 in accuracy, 0.00649 in precision, 0.504 in recall, and 0.013 in F1 score.

Keywords: *Isolation Forest, iForest, Anomaly Detection*

1. Pendahuluan

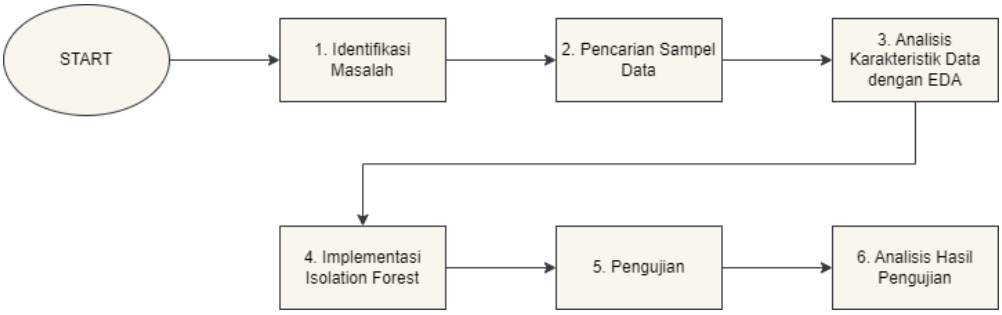
Dalam era Big Data, organisasi menghadapi tantangan baru dalam mengelola dan menganalisis volume data yang besar dan beragam. Kunci dari nilai yang terkandung dalam Big Data adalah kemampuan untuk mengubah data tersebut menjadi wawasan yang bernilai bagi organisasi. Wawasan ini menjadi landasan untuk pengambilan keputusan yang lebih baik, strategi yang lebih efektif, dan keunggulan kompetitif. Karakteristik Big Data sering disebut sebagai 5 V's of Big Data, yaitu volume, variety, velocity, viability, dan value. Dari kelima karakteristik tersebut, value memiliki peran utama dalam menentukan valuasi dari data yang digunakan untuk pengambilan keputusan suatu organisasi. Salah satu tantangan dalam menjaga valuasi dari suatu data adalah adanya anomaly terhadap data tersebut. Anomali data merujuk pada kejadian atau pola yang tidak biasa, tidak terduga, atau tidak sesuai dengan harapan yang ada dalam data. Anomali dapat muncul dalam berbagai bentuk, seperti outlier yang mencolok, kesalahan pengukuran, atau perubahan drastis dalam pola data. Keberadaan anomali dapat menyebabkan bias, ketidaktepatan, atau kesalahan dalam analisis dan pengambilan keputusan. Dalam pengembangan sistem, terdapat beberapa metode yang dapat dilakukan untuk melakukan deteksi anomali data. Salah satu metode yang digunakan adalah machine learning.

Dalam deteksi anomali data, machine learning telah menjadi salah satu metode yang populer dan efektif. Metode ini melibatkan penggunaan algoritma dan teknik pembelajaran mesin untuk mengidentifikasi pola atau perilaku yang tidak biasa dalam data. Salah satu pendekatan dari machine learning yang populer digunakan untuk deteksi anomali adalah unsupervised learning approach. Menurut penelitian yang dilakukan oleh Bakumenko dan rekan mengenai perbandingan deteksi anomali data menggunakan beberapa model machine learning, di dapatkan bahwa pada ranah pendekatan secara unsupervised, Isolation Forest memiliki nilai akurasi tertinggi [2]. Isolation Forest merupakan algoritma unsupervised learning yang didasarkan atas Algoritma pohon keputusan atau decision tree. Secara garis besar cara kerja algoritma ini akan memisahkan anomali dari dataset yang digunakan dengan cara membagi suatu data dengan pohon keputusan sampai anomali benar – benar terpisah [1].

Pada penelitian ini, penulis akan menguji performa dari Isolation Forest untuk deteksi anomali data sehingga dapat digunakan untuk decision making yang lebih baik ke depannya.

2. Metode Penelitian

2.1 Alur Penelitian



Gambar 1. Alur Penelitian

Alur kerangka penelitian di atas memiliki penjelasan sebagai berikut:

a. Identifikasi Masalah: tahapan awal dari penelitian, di mana peneliti mengidentifikasi konsep dari deteksi anomali data melalui studi literatur.

b. Pencarian Sampel Data: pada tahapan ini, data dicari dengan menggunakan data sekunder yang berasal dari Kaggle.com. Data yang digunakan merupakan data transaksi.

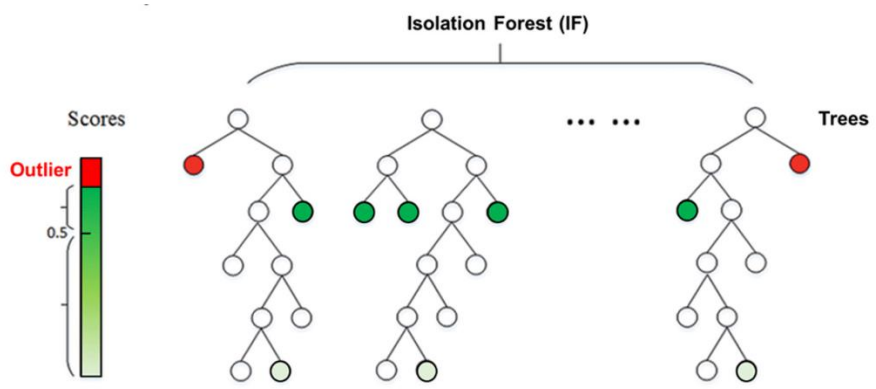
c. Implementasi Isolation Forest: pada tahapan ini, peneliti melakukan pemodelan terhadap data yang telah dianalisis dan dibersihkan dengan algoritma Isolation Forest. Pada tahapan ini akan memanfaatkan bahasa pemrograman Python dan Visual Studio Code untuk implementasinya.

d. Pengujian: pada tahapan ini, peneliti akan melakukan training dan testing data terhadap model yang telah dibuat dengan pembagian data yaitu 80% untuk training dan 20% untuk testing model.

e. Analisis Hasil Pengujian: pada tahapan ini, peneliti akan menganalisis performa dari Isolation Forest untuk deteksi anomali data. Tolak ukur pengujian yang digunakan meliputi Accuracy, Precision, Recall, F1 Score, dan Average Anomaly Score.

2.2 Isolation Forest

Isolation Forest atau iForest merupakan algoritma unsupervised learning approach, yang didasarkan pada konsep pemisahan atau isolasi data anomali dari data normal. Isolation Forest menggunakan pendekatan berbasis pohon (tree-based approach) untuk membagi data menjadi subgrup yang semakin kecil, dengan tujuan memisahkan data anomali dengan cepat. Pada saat membangun pohon dalam Isolation Forest, data dipecah secara acak dan dipisahkan dengan menggunakan fitur-fitur yang ada. Anomali cenderung membutuhkan jumlah pemisahan yang lebih sedikit dibandingkan dengan data normal. Dengan demikian, ketika suatu data ditempatkan pada posisi yang lebih awal dalam pohon (memiliki jarak lebih pendek dari root), kemungkinan besar data tersebut merupakan anomali. Proses isolasi dan pemisahan berulang-ulang pada pohon-pohon yang dibangun membentuk model deteksi anomali. Kemudian, dengan menggunakan model ini, data baru dapat diuji untuk melihat apakah data tersebut termasuk anomali atau tidak.



Gambar 2. Ilustrasi Isolation Forest Model. Sumber: datrics.ai

3. Hasil dan Diskusi

Berdasarkan alur penelitian dan juga studi literatur telah dijelaskan, maka didapatkan hasil penelitian sebagai berikut.

3.1. Pencarian Sampel Data

Penelitian ini menggunakan data transaksi dari Kaggle.com yang di-publish oleh Edgar Lopez-Rojaz [4]. Format data tersebut dapat dilihat pada **Tabel 1** di bawah.

Tabel 1. Format Dataset

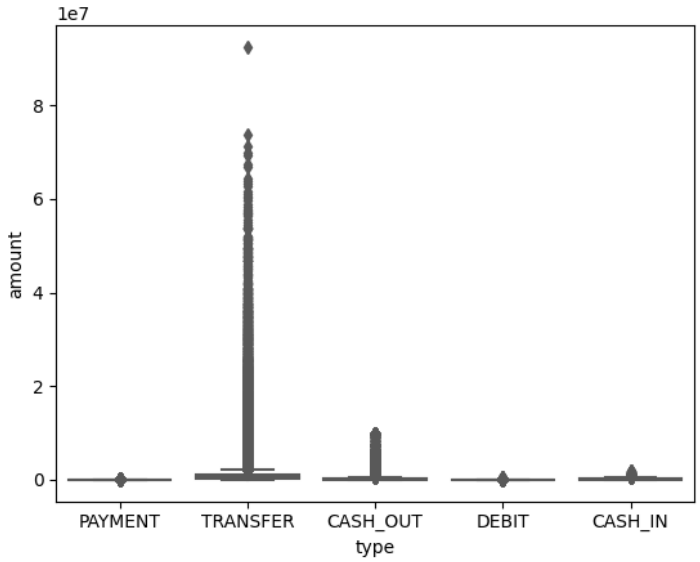
Key	Description
step	Waktu transaksi itu dilakukan, 1 step bernilai 1 jam
type	Tipe transaksi yang dilakukan
amount	Nominal dari transaksi yang dilakukan (CASH_IN, CASH_OUT, DEBIT, TRANSFER, PAYMENT)
nameOrig	Pelanggan yang memulai transaksi
oldbalanceOrg	Saldo awal pelanggan sebelum transaksi
newbalanceOrig	Saldo baru pelanggan setelah transaksi
nameDest	Pelanggan Tujuan transaksi
oldbalanceDest	Saldo awal pelanggan tujuan
newbalanceDest	Saldo akhir pelanggan tujuan
isFraud	Indikator transaksi normal atau tidak
isFlaggedFraud	Indikator normal atau tidak normal jika nominal suatu transaksi melebihi 200.000

3.2. Analisis Karakteristrik Data dengan EDA

Proses analisis karakteristik pada dataset yang digunakan dengan EDA akan dilakukan dalam beberapa tahap yaitu sebagai berikut

a. Melihat Distribusi Transaksi Berdasarkan Tipe Transaksi

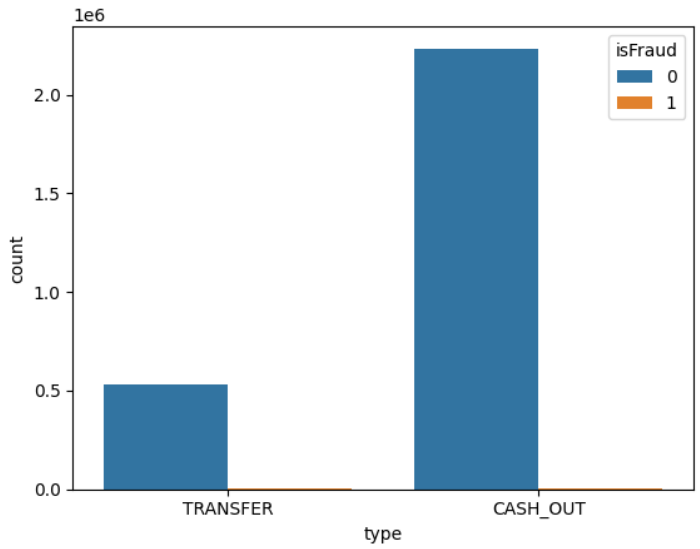
Yang pertama dilakukan adalah melihat persebaran dari data transaksi pada dataset tersebut. Dapat dilihat bahwa ternyata ada beberapa transaksi TRANSFER dan CASH_OUT yang terlihat janggal berdasarkan visualisasi ini.



Gambar 3. Distribusi Transaksi Berdasarkan Tipe Transaksi

b. Melihat Persebaran Anomali pada TRANSFER dan CASH_OUT

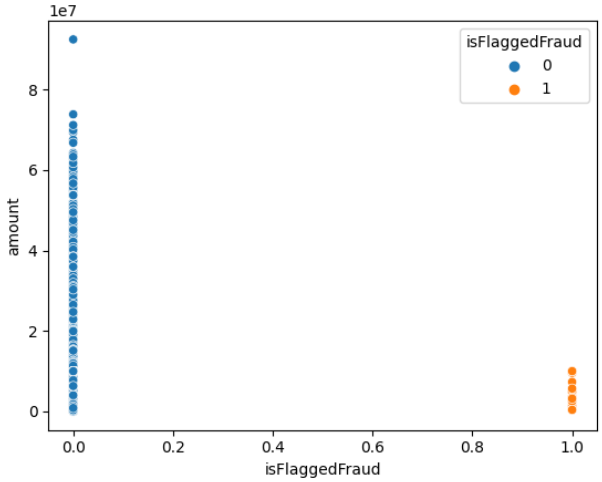
Kemudian dilakukan pengecekan jumlah transaksi yang anomali pada TRANSFER dan CASH_OUT. Ditemukan bahwa sebanyak 4116 CASH_OUT dan 4097 TRANSFER yang dinyatakan anomali.



Gambar 4. Visualisasi Data Normal dan Anomali

c. Analisis Pengaruh Fitur isFlaggedFraud pada Dataset

Setelah mengetahui bahwa persebaran anomali ada pada tipe TRANSFER dan CASH_OUT, perlu dilakukan pengecekan apakah anomali – anomali tersebut memiliki nominal transaksi melebihi 200.000 atau tidak. Setelah dilakukan pengecekan, dinyatakan bahwa fitur tersebut hanya dimiliki oleh transaksi TRANSFER, sehingga perlu dilakukan analisis seberapa pengaruhnya fitur tersebut. Pertama dicek terlebih dahulu apakah seluruh transaksi yang memiliki fitur ini memiliki transaksi lebih dari 200.000 atau tidak.



Gambar 5. Distribusi Transaksi Berdasarkan Fitur isFlaggedFraud

Berdasarkan visualisasi tersebut, dapat dikatakan bahwa ternyata fitur ini tidak berpengaruh pada keanomalian data. Hal ini dikarenakan transaksi yang dilabeli tidak memiliki fitur isFlaggedFraud juga mempunyai nominal transaksi yang melebihi 200.000, sehingga fitur ini dapat dihilangkan dari groundtruth dan hanya isFraud yang akan digunakan.

d. Pembersihan Data (Data Cleaning)

Langkah selanjutnya adalah membersihkan dataset tersebut agar dapat digunakan untuk perhitungan. Mengingat Isolation Forest akan bekerja lebih baik jika menerima data berupa numerik, maka kita perlu menghilangkan beberapa fitur yang tidak relevan. Pada kasus ini, kita akan menghilangkan fitur nameOrig, nameDest, dan isFlaggedFraud karena dianggap tidak berpengaruh signifikan pada perhitungan berdasarkan analisis yang telah dilakukan. Kemudian, kita akan mengubah nilai TRANSFER dan CASH_OUT tadi menjadi numerik yaitu 0 dan 1 untuk mempermudah perhitungan. Selanjutnya, dilakukan pelabelan dari nilai – nilai fitur yang berelasi sama - sama bernilai kosong atau 0. Pada kasus ini yang memiliki nilai 0 adalah fitur oldbalanceDest dengan newbalanceDest dan oldbalanceOrg, dengan newbalanceOrg. Sehingga hasil dari pembersihan tadi akan seperti berikut.

	step	type	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
2	1	0	181.00	181.0	0.0	-1.0	-1.00	1
3	1	1	181.00	181.0	0.0	21182.0	0.00	1
15	1	1	229133.94	15325.0	0.0	5083.0	51513.44	0
19	1	0	215310.30	705.0	0.0	22425.0	0.00	0
24	1	0	311685.89	10835.0	0.0	6267.0	2719172.89	0

Gambar 6. Hasil Data Cleaning Terhadap Dataset

e. Pembentukan Ground Truth

Untuk ground truth di sini akan menggunakan fitur isFraud sebagai label untuk nanti digunakan sebagai salah satu parameter untuk melakukan proses prediksi.

	isFraud
0	0
1	0
2	1
3	1
4	0

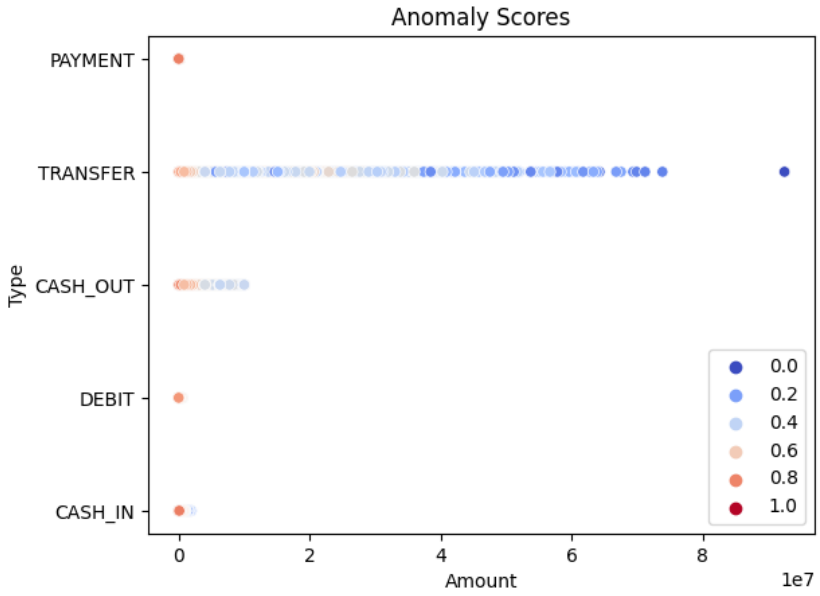
Gambar 7. Ground Truth

3.3. Implementasi Isolation Forest

Pada tahapan ini, peneliti akan mengimplementasikan Isolation Forest untuk deteksi anomali pada data. Hal pertama yang perlu dilakukan adalah menentukan outlier fraction yang akan digunakan. Nilai ini akan menjadi nilai parameter contamination pada model atau nilai prediksi dari persebaran anomali. Ini dapat dilakukan dengan membagi jumlah transaksi yang dinyatakan dianggap anomali dengan jumlah transaksi normal. Didapatkan nilai contamination sebesar 0.1. Kemudian kita akan membagi dataset tersebut menjadi dua bagian, 80% untuk training dan 20% untuk testing. Kemudian data training tersebut akan dimasukan pada model Isolation Forest. Waktu training ini memakan waktu sebesar 3 menit 52 detik. Berdasarkan pengujian, diperoleh bahwa nilai akurasi dari model yang dibuat mencapai akurasi sebesar 0.899, precision sebesar 0.00649, recall sebesar 0.503, dan F1 score sebesar 0.0128.

```
Jumlah data yang terdeteksi sebagai outlier: 127815
Accuracy: 0.8999233020359537
Precision: 0.006496433007595660 5
Recall: 0.5037136247412638
F1 Score: 0.012827429677035275
```

Gambar 8. Hasil Pengujian Model



Gambar 9. Visualisasi Distribusi Skor Anomali

4. Kesimpulan

Berdasarkan penelitian, ditemukan bahwa Isolation Forest memiliki performa yang baik. EDA sangat membantu dalam menganalisis dataset yang digunakan dan mengeliminasi fitur – fitur yang tidak berpengaruh secara signifikan sehingga memudahkan perhitungan pada model. Isolation Forest memiliki performa yang baik dengan total anomali yang ditemukan sebanyak 1646. Algoritma ini juga mampu mencapai anomaly score sebesar 0.47%.

Daftar Pustaka

[1] P. Patil, “What is exploratory data analysis?,” Medium, https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15 (accessed Jun. 11, 2023).
[2] E. S. Nugroho, “Catatan Belajar Anomaly Detection Menggunakan Algoritma Isolation Forest,” Medium, https://edistywn.medium.com/catatan-belajar-anomaly-detection-menggunakan-algoritma-isolation-forest-4e4d13e61c3d (accessed Jun. 11, 2023).
[3] A. Bakumenko and A. Elragal, “Detecting anomalies in financial data using machine learning algorithms,” Systems, vol. 10, no. 5, p. 130, 2022. doi:10.3390/systems10050130.
[4] E. A. Lopez-Rojas, A. Elmir, and S. Axelsson. "PaySim: A financial mobile money simulator for fraud detection". In: The 28th European Modeling and Simulation Symposium-EMSS, Larnaca, Cyprus. 2016
[5] S. V. Daele and G. Janssenswillen, “Identifying the steps in an exploratory data analysis: A process-oriented approach,” Lecture Notes in Business Information Processing, pp. 526–538, 2023. doi:10.1007/978-3-031-27815-0_38

Halaman ini sengaja dibiarkan kosong