

# Comparative Analysis of Local Outlier Factor (LOF) and Elliptic Envelope Algorithms in Outlier Detection of Customer Purchase Behavior Data

Viola<sup>a</sup>, Maya Sofhia<sup>b</sup>, Juliansyah Putra Tanjung<sup>c</sup>, Marciello Farrel<sup>d</sup>

<sup>a,b,c,d</sup>Universitas Prima Indonesia

Corresponding Author:

<sup>a</sup>chandraaviola@gmail.com

## ABSTRACT

Outlier detection is a crucial stage in customer data analysis, as the presence of anomalies can significantly affect analytical outcomes and decision-making processes. This study aims to compare the performance of the Local Outlier Factor (LOF) and Elliptic Envelope (EE) algorithms in detecting outliers in customer behavior data. The dataset consists of 2,240 customer records with eight key features representing demographic characteristics, purchasing behavior, and responses to marketing campaigns. The research stages include data preprocessing, feature selection, normalization, and the application of the LOF and EE algorithms. The results indicate that LOF detects a larger number of outliers compared to EE, while EE tends to be more selective and focuses on extreme global deviations. A number of overlapping data points were also detected by both algorithms, indicating the presence of anomalies with consistent characteristics. Feature characteristic analysis reveals that each algorithm has a distinct and complementary detection focus. These findings suggest that a comparative approach between LOF and EE provides a more comprehensive understanding of customer anomaly patterns and supports the selection of appropriate outlier detection methods aligned with specific analytical objectives.

**Keywords :** Outlier Detection, Local Outlier Factor, Elliptic Envelope, Customer Behavior Analysis, Data Mining

## INTRODUCTION

The development of Information and Communication Technology (ICT) has encouraged retail companies to leverage data as a strategic asset in data-driven decision-making. Customer purchasing behavior data obtained from Point of Sale (POS) systems, loyalty programs, and online shopping platforms contain valuable information related to consumption patterns, product preferences, and customer spending levels. The analysis of customer behavioral data has been shown to enhance the effectiveness of marketing strategies and strengthen corporate competitiveness (Ifmaini et al., 2025; Nugroho et al., 2023).

In practice, customer purchasing behavior data often contain outliers, defined as observations that deviate significantly from the majority of the data. The presence of outliers can reduce analytical quality and introduce model bias; however, outliers may also represent important phenomena such as extreme customer behavior, abnormal campaign responses, or indications of transactional anomalies (Riski et al., 2025; Amirillah, 2025). Therefore, outlier detection is

a crucial stage in the early analysis of retail data to ensure result validity and improve the understanding of customer behavior patterns (Kurniawan & Utomo, 2022).

The primary challenge in outlier detection for customer behavior data lies in its multidimensional, heterogeneous nature, which does not always follow a normal distribution. Conventional methods such as the z-score and Interquartile Range (IQR) generally rely on specific statistical assumptions and are less effective when applied to large-scale data with complex structures (Riski et al., 2025; Ramadhan & Sutoyo, 2023). Other approaches, such as One-Class Support Vector Machine and deep learning-based models (e.g., LSTM), are capable of capturing non-linear patterns; however, they often require complex parameter tuning and high computational resources, making them less efficient for the initial exploratory analysis of unlabeled data (Putri et al., 2025).

With the advancement of machine learning, unsupervised learning-based outlier detection approaches have gained increasing attention due to their ability to identify complex patterns without requiring labeled data or specific distributional assumptions (Pratama et al., 2023; Bajpai et al., 2024). In this context, the Local Outlier Factor (LOF) and Elliptic Envelope (EE) algorithms are two methods of interest for comparative analysis, as they represent distinct detection paradigms. LOF is a density-based method that identifies anomalies by comparing the local density of a data point with that of its nearest neighbors, making it effective for detecting local deviations (Ramadhan & Sutoyo, 2023; Bajpai et al., 2024). In contrast, Elliptic Envelope employs a statistical approach based on robust covariance to globally model the normal data distribution as a multidimensional ellipse, where observations lying outside the ellipse boundary are classified as outliers (Wahyuni et al., 2022).

Most previous studies have primarily focused on predictive modeling of customer behavior using supervised learning approaches, such as Random Forest, Support Vector Machine, and ensemble methods, with the aim of improving prediction accuracy (Hodge & Austin, 2018; Alfian et al., 2023; Deniz & Bülbül, 2024; Abu-Dalbouh et al., 2025). However, the pre-modeling stage—particularly unsupervised outlier detection—has received relatively limited attention, despite its critical role in improving data quality and facilitating early interpretation of customer behavior patterns.

Based on this research gap, the present study focuses on a comparative analysis of the Local Outlier Factor and Elliptic Envelope algorithms for detecting outliers in customer purchasing behavior data. The analysis compares the number of detected outliers, the level of result agreement (overlap), and the separation patterns between normal and anomalous data through Principal Component Analysis (PCA) visualization. This study is expected to contribute to a deeper understanding of the differences between local density-based and global statistical distribution-based outlier detection methods and to serve as a basis for selecting appropriate techniques for the early-stage analysis of retail data.

## **METHODS**

### **Research Design**

This study adopts a quantitative approach with a comparative analysis method, aiming to compare the performance of two unsupervised learning-based outlier detection algorithms, namely the Local Outlier Factor (LOF) and Elliptic Envelope (EE), on customer purchasing

behavior data. The quantitative approach was selected because it enables objective analysis of data patterns, the number of detected outliers, and differences in detection characteristics produced by each algorithm (Ghodang & Hantono, 2020).

### **Dataset**

The dataset used in this study was obtained from the Kaggle platform under the title Consumer Buying Behavior Analysis, consisting of 2,240 customer records. The dataset contains demographic information and retail purchasing behavior attributes, including income, recency, number of purchases across different channels, total spending, and responses to marketing campaigns. This dataset was selected because it represents typical retail data characteristics that are multidimensional, heterogeneous, and potentially contain behavioral anomalies.

### **Preprocessing and Feature Engineering**

The preprocessing stage was conducted to ensure data quality and readiness prior to analysis. Preprocessing steps included the removal of records containing missing values and the transformation of categorical variables into numerical form using label encoding. Subsequently, feature engineering was performed to construct aggregated variables that more comprehensively represent customer behavior, namely Total\_Spend (the accumulation of all expenditures), Total\_Accepted (the total number of campaign responses), and Total\_Children (the number of dependent children). This process aimed to simplify the data structure while enhancing the representativeness of customer behavior patterns.

### **Feature Selection and Normalization**

Feature selection was performed exploratively based on domain knowledge without employing formal statistical feature selection techniques. Eight numerical features were selected as the primary representation of customer purchasing behavior: Income, Recency, Total\_Spend, NumWebPurchases, NumStorePurchases, NumWebVisitsMonth, Total\_Accepted, and Total\_Children. All selected features were subsequently normalized using StandardScaler, resulting in variables with zero mean and unit variance. Normalization is essential because distance- and density-based outlier detection algorithms are highly sensitive to differences in data scale (Rachmawati et al., 2024).

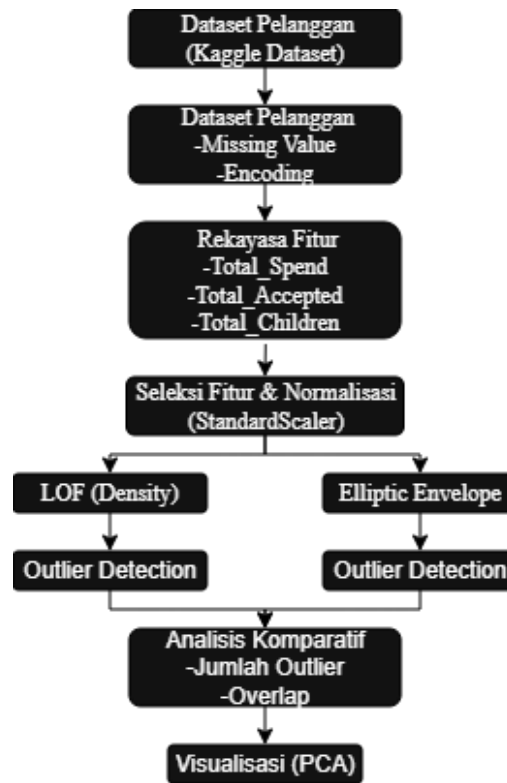


**Figure 1. Research workflow for outlier detection in customer purchasing behavior data using the Local Outlier Factor and Elliptic Envelope algorithms.**

The overall research workflow is illustrated in Figure 1, which depicts the sequential stages from data preprocessing to the comparative analysis of outlier detection results obtained using the LOF and Elliptic Envelope algorithms.

### Outlier Detection

Outlier detection was conducted using two unsupervised algorithms: Local Outlier Factor (LOF) and Elliptic Envelope (EE). The LOF algorithm identifies outliers by comparing the local density of a data point with that of its nearest neighbors, making it effective for detecting locally deviating observations. The `n_neighbors` parameter was set to 15, while the contamination parameter was fixed at 0.05, reflecting the assumption of a relatively small proportion of anomalies and aligning with commonly adopted values in the literature (Ramadhan & Sutoyo, 2023).

As a comparative method, Elliptic Envelope was applied to detect outliers using a global statistical approach by modeling the normal data distribution through robust covariance estimation. The contamination parameter for EE was also set to 0.05 to ensure a fair and consistent comparison with the LOF algorithm (Wahyuni et al., 2022). Both algorithms were applied to the normalized dataset, producing binary labels that distinguish normal observations from outliers.

## Visualization and Comparative Analysis

To visualize outlier detection results in high-dimensional data, Principal Component Analysis (PCA) was employed to reduce the data into two principal components. The visualization was presented using scatter plots to observe the separation patterns between normal data and outliers identified by each algorithm.

The comparative analysis focused on comparing the number of detected outliers, the level of result agreement (overlap) between LOF and EE, and the average feature characteristics of data points classified as outliers. This approach enables a descriptive and interpretative evaluation of differences between local density-based and global statistical distribution-based detection paradigms.

## Implementation Environment

All data processing and analysis were conducted using the Python programming language, supported by the Pandas, NumPy, Scikit-learn, and Matplotlib libraries. The implementation was carried out in the Google Colaboratory environment, ensuring that all research procedures are transparent and fully reproducible.

## RESULTS

### Data Characteristics After Preprocessing

The dataset used in this study consists of 2,240 customer records, as indicated during the dataset loading process. The initial condition of the data prior to preprocessing is presented to illustrate the distribution and baseline characteristics of the numerical features. After the preprocessing stage, the dataset exhibits a cleaner structure and is fully prepared for subsequent analysis.

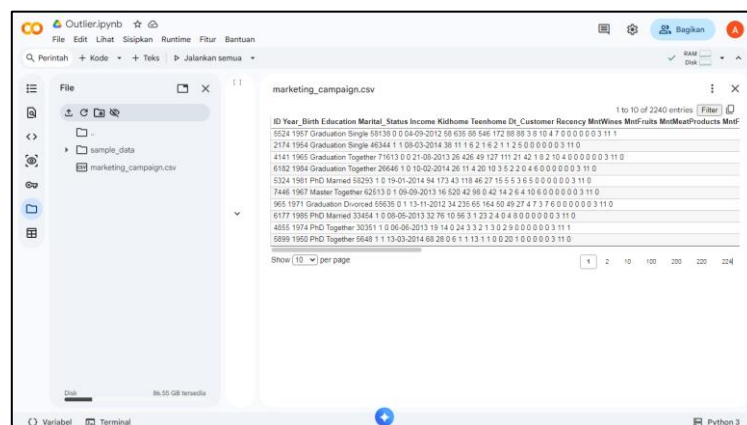


Figure 2. Dataset loading results

The dataset was successfully loaded into the programming environment, comprising 2,240 customer records, as shown in Figure 2. This display presents the initial data structure along with all attributes utilized in the subsequent analysis stages.

```

=== 5 DATA RAW (SEMUA KOLON) ===
ID Year_Birth Education Marital_Status Income Kidhome Teenhome \
0 5524 1957 Graduation Single 58138.0 0 0
1 2174 1954 Graduation Single 46344.0 1 1
2 4141 1965 Graduation Together 71613.0 0 0
3 6182 1984 Graduation Together 26646.0 1 0
4 5324 1981 PHD Married 58293.0 1 0

Dt_Customer Recency MntWines MntFruits MntMeatProducts MntFishProducts \
0 04-09-2012 58 635 88 546 172
1 08-02-2014 38 11 1 6 2
2 21-08-2013 26 426 49 127 111
3 18-02-2014 26 11 4 20 10
4 19-01-2014 94 173 43 118 46

MntSweetProducts MntGoldProds NumDealsPurchases NumWebPurchases \
0 88 88 3 8
1 1 6 2 1
2 21 42 1 8
3 3 5 2 2
4 27 15 5 5

NumCatalogPurchases NumStorePurchases NumWebVisitsMonth AcceptedCmp3 \
0 10 4 7 0
1 1 2 5 0
2 2 10 4 0
3 0 4 6 0
4 3 6 5 0

AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2 Complain \
0 0 0 0 0 0
1 0 0 0 0 0
2 0 0 0 0 0
3 0 0 0 0 0
4 0 0 0 0 0

Z_CostContact Z_Revenue Response
0 3 11 1
1 3 11 0
2 3 11 0
3 3 11 0
4 3 11 0

```

**Figure 3. Data before preprocessing.**

Figure 3 presents the first five records prior to preprocessing, illustrating the initial dataset structure, which consists of a combination of numerical and categorical variables and does not yet include aggregated features representing customer purchasing behavior.

```

=== 5 DATA SETELAH PREPROCESSING (SEMUA KOLON) ===
ID Year_Birth Education Marital_Status Income Kidhome Teenhome \
0 5524 1957 2 4 58138.0 0 0
1 2174 1954 2 4 46344.0 1 1
2 4141 1965 2 5 71613.0 0 0
3 6182 1984 2 5 26646.0 1 0
4 5324 1981 4 3 58293.0 1 0

Dt_Customer Recency MntWines MntFruits MntMeatProducts MntFishProducts \
0 04-09-2012 58 635 88 546 172
1 08-02-2014 38 11 1 6 2
2 21-08-2013 26 426 49 127 111
3 18-02-2014 26 11 4 20 10
4 19-01-2014 94 173 43 118 46

MntSweetProducts MntGoldProds NumDealsPurchases NumWebPurchases \
0 88 88 3 8
1 1 6 2 1
2 21 42 1 8
3 3 5 2 2
4 27 15 5 5

NumCatalogPurchases NumStorePurchases NumWebVisitsMonth AcceptedCmp3 \
0 10 4 7 0
1 1 2 5 0
2 2 10 4 0
3 0 4 6 0
4 3 6 5 0

AcceptedCmp4 AcceptedCmp5 AcceptedCmp1 AcceptedCmp2 Complain \
0 0 0 0 0 0
1 0 0 0 0 0
2 0 0 0 0 0
3 0 0 0 0 0
4 0 0 0 0 0

Z_CostContact Z_Revenue Response Total_Spend Total_Accepted \
0 3 11 1 1617
1 3 11 0 27
2 3 11 0 776
3 3 11 0 53
4 3 11 0 422

Total_Children
0 0
1 2
2 0
3 1
4 1

```

**Figure 4. Data after preprocessing**

Figure 4 illustrates the dataset after preprocessing, in which all attributes have been transformed into numerical form, and several new features resulting from feature engineering have been added to represent customer purchasing behavior.

### Results of Feature Selection and Data Normalization

The feature selection process was conducted to identify numerical variables that are most relevant for representing customer purchasing behavior. The results of feature selection indicate the primary features used in the outlier detection process. Subsequently, all selected

features were normalized to standardize the data scale prior to applying the outlier detection algorithms.

```

***
=== STATISTIK DESKRIPTIF FITUR TERPILIH ===
count      Income      Recency      Total_Spend      NumWebPurchases  \
mean      52247.251354    49.012635    607.075361      4.085289
std       25173.076661    28.948352    602.900476      2.740951
min        1730.000000     0.000000     5.000000      0.000000
25%       35303.000000    24.000000    69.000000      2.000000
50%       51381.500000    49.000000   396.500000      4.000000
75%       68522.000000    74.000000   1048.000000     6.000000
max       666666.000000   99.000000  2525.000000    27.000000

count      NumStorePurchases      NumWebVisitsMonth      Total_Accepted      Total_Children
mean         5.800993              5.319043              0.448556              0.947202
std         3.250785              2.425359              0.892440              0.749062
min         0.000000              0.000000              0.000000              0.000000
25%         3.000000              3.000000              0.000000              0.000000
50%         5.000000              6.000000              0.000000              1.000000
75%         8.000000              7.000000              1.000000              1.000000
max        13.000000             20.000000              5.000000              3.000000

```

Figure 5. Feature selection results.

Descriptive statistics indicate that the Income feature exhibits a very wide value range (1,730–666,666) with a high standard deviation, whereas the Recency feature ranges from 0 to 99 days with an average of 49 days. Shopping behavior features, such as Total\_Spend, NumWebPurchases, and NumStorePurchases, as well as Total\_Accepted and Total\_Children, demonstrate substantial variability and the presence of extreme values across customers.

```

=== 5 DATA SETELAH NORMALISASI ===
Income      Recency      Total_Spend      NumWebPurchases      NumStorePurchases  \
0  0.234063    0.310532      1.675488          1.428553            -0.554143
1 -0.234559 -0.380509     -0.962358         -1.125881           -1.169518
2  0.769478 -0.795134     0.280250          1.428553            1.291982
3 -1.017239 -0.795134     -0.919224         -0.760962           -0.554143
4  0.240221  1.554407     -0.307044          0.333796             0.061232

NumWebVisitsMonth      Total_Accepted      Total_Children
0      0.693232              0.618045          -1.264803
1     -0.131574             -0.502731           1.405806
2     -0.543978             -0.502731          -1.264803
3      0.280829             -0.502731           0.070501
4     -0.131574             -0.502731           0.070501

```

Figure 6. Data normalization results

Figure 6 presents the first five records after the normalization process, where all numerical features have been standardized with distributions centered around zero and are within comparable value ranges.

### Outlier Detection Results

Outlier detection was performed using the Local Outlier Factor and Elliptic Envelope algorithms with the same *contamination* parameter. The detection results indicate the presence of a number of data points identified as outliers, representing customer purchasing behaviors that deviate from the majority patterns.

```

=== HASIL DETEKSI OUTLIER ===
Outlier LOF      : 156
Outlier EE       : 67
Overlap          : 25
Overlap (%)      : 37.31%

```

Figure 7. Outlier detection results

The Local Outlier Factor (LOF) algorithm identified 156 customer records as outliers, whereas the Elliptic Envelope (EE) algorithm detected 67 outliers. Both methods shared 25 overlapping data points, resulting in an agreement level of 37.31% relative to the minimum number of detected outliers.

### Visualization of Outlier Detection Results

To clarify the separation between normal data and outliers, the detection results were visualized using two-dimensional plots. This visualization provides an overview of the outlier distributions produced by both algorithms and highlights the differences in the deviation patterns identified.

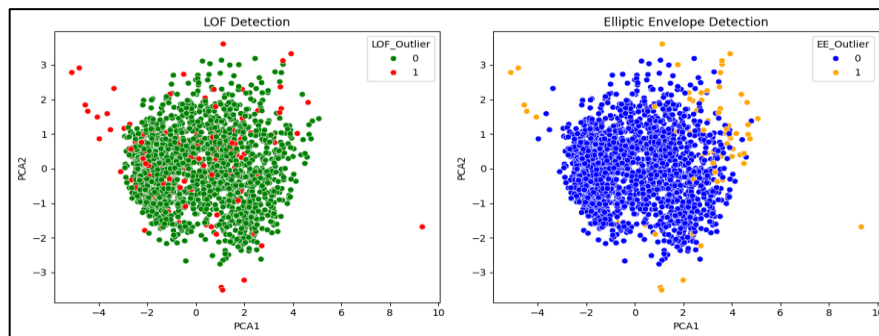


**Figure 8. Visualization of outlier detection results**

Figure 8 presents the PCA-based visualization of outlier detection results obtained using the Local Outlier Factor and Elliptic Envelope algorithms. Most data points are concentrated in the central region, while observations identified as outliers are located at the periphery of the distribution. Visually, the LOF results exhibit a larger number of outliers with a more dispersed pattern compared to those detected by the Elliptic Envelope method.

### Comparative Analysis of Outlier Detection Results

A comparative analysis was conducted to evaluate the outlier detection results produced by the Local Outlier Factor and Elliptic Envelope algorithms. The comparison was based on the feature characteristics of the detected outliers as well as the level of agreement between the two methods. The comparative results indicate differences in the characteristics of outliers identified by each algorithm.

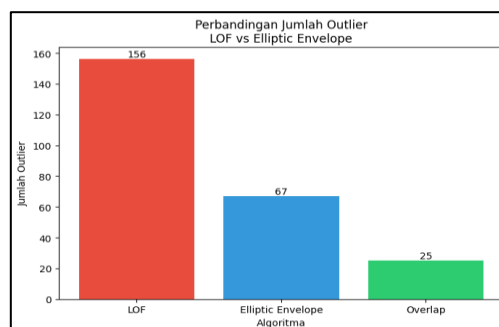


**Figure 9. Comparison of outlier detection results between algorithms**

Figure 9 illustrates the comparison of the number of outliers detected by the Local Outlier Factor (156 records) and Elliptic Envelope (67 records), as well as the 25 data points identified as overlapping outliers by both algorithms.

**Table 1. Comparative analysis of outlier feature characteristics.**

| Key Features      | Mean Value of LOF Outliers | Mean Value of EE Outliers | Differences in Characteristics                                                          |
|-------------------|----------------------------|---------------------------|-----------------------------------------------------------------------------------------|
| Income            | 55.333,33                  | 90.539,49                 | EE captures customers with extremely high income levels.                                |
| Recency           | 53,93                      | 48,75                     | Both methods show relatively similar patterns in recency (time since last transaction). |
| Total_Spend       | 583,1                      | 1.376,09                  | EE detects customers with extreme spending behavior.                                    |
| NumWebPurchases   | 4,03                       | 5,28                      | EE is more inclined toward identifying highly active web users.                         |
| NumStorePurchases | 5,08                       | 6,01                      | In-store purchasing activity appears relatively similar across both methods.            |
| NumWebVisitsMonth | 6,35                       | 4,45                      | LOF captures anomalies in web visit intensity.                                          |
| Total_Accepted    | 0,57                       | 3,09                      | EE is particularly effective in detecting massive campaign response behavior.           |
| Total_Children    | 1,25                       | 0,37                      | LOF identifies anomalies related to the number of dependents.                           |

Table 3.1 presents a comparison of the mean values of eight key features for outliers detected by the Local Outlier Factor (LOF) and Elliptic Envelope (EE) algorithms. The results indicate differences in average values between the two algorithms across purchasing behavior features, campaign response variables, and customer demographic characteristics.

### Summary of Results

Overall, the results indicate that the preprocessing, feature selection, and data normalization stages play a crucial role in the outlier detection process. The Local Outlier Factor and Elliptic Envelope algorithms produce different outlier detection patterns, as reflected in the visualizations and the comparative analysis of outlier feature characteristics.

### DISCUSSION

The results of this study reveal clear differences between the Local Outlier Factor (LOF) and Elliptic Envelope (EE) algorithms in identifying outliers within the customer dataset. These

differences are reflected in the number of detected outliers, the visual distribution patterns, and the feature characteristics of the outlier data, as illustrated in Figures 7 to 9 and Table 1. Based on the detection results, LOF identified a substantially larger number of outliers compared to EE. This finding indicates that LOF tends to capture anomalies based on local density, enabling it to detect customers whose behavioral patterns deviate relative to their immediate neighborhood. In contrast, EE detected only a smaller subset of data as outliers, reflecting its more selective nature due to its reliance on global data distribution modeling. This fundamental difference in detection approaches explains the imbalance in the number of outliers identified by the two algorithms.

The presence of overlapping data points detected by both algorithms suggests that a subset of customers exhibits consistent anomalous characteristics from both local density and global distribution perspectives. These overlapping outliers can be considered the most robust anomalies, as they are identifiable through two conceptually different detection mechanisms. This finding strengthens the quantitative comparison results and provides cross-validation for the outlier detection outcomes.

PCA-based visualizations further demonstrate that LOF-detected outliers are more widely dispersed in low-density regions, whereas EE-detected outliers tend to lie outside the elliptical boundary of the normal data distribution. These visual patterns are consistent with the underlying principles of each algorithm: LOF is sensitive to local density variations, while EE emphasizes distance from the multivariate distribution center. Consequently, visualization serves not only as a graphical representation but also as a means of clarifying the distinct detection characteristics of both methods.

The analysis of feature characteristics for outlier data presented in Table 1 shows that EE tends to detect customers with extreme values in economic and purchasing behavior features, such as Income, Total\_Spend, and Total\_Accepted. This indicates that EE is more effective in identifying customers with significant global deviations from the dominant data pattern. Conversely, LOF captures more anomalies related to activity intensity features, such as NumWebVisitsMonth and Total\_Children, suggesting the presence of unusual local patterns that are not necessarily extreme at the global level.

These differences highlight the complementary nature of LOF and EE in outlier detection. LOF is more suitable for identifying behavioral anomalies arising from local variations among customers, while EE is better suited for detecting customers with attribute values that deviate substantially from the overall data distribution. Therefore, the choice of an outlier detection algorithm should be aligned with the analytical objective, whether the focus is on local anomalies or global deviations.

Overall, the findings of this study demonstrate that employing multiple outlier detection algorithms provides a more comprehensive understanding of customer anomaly patterns. The comparative approach between LOF and EE not only enriches the analytical results but also facilitates the identification of customer groups with extreme or atypical characteristics, which may hold strategic value in customer behavior analysis and data-driven decision-making.

## CONCLUSION

This study aimed to analyze and compare the performance of the Local Outlier Factor (LOF) and Elliptic Envelope (EE) algorithms in detecting outliers in customer behavior data. Based on the experimental results, both algorithms exhibit significant differences in terms of the number of detected outliers, data distribution patterns, and the feature characteristics of the identified outliers.

The findings indicate that LOF is capable of detecting a larger number of outliers than EE, highlighting its effectiveness in capturing anomalies based on local density variations. In contrast, EE tends to be more selective and effective in identifying outliers with extreme global deviations from the overall data distribution. The presence of overlapping outliers detected by both algorithms suggests the existence of a consistent and robust group of anomalies that can be recognized from both local and global perspectives.

The analysis of feature characteristics further demonstrates that each algorithm has a distinct detection focus. EE predominantly identifies customers with extreme values in economic and purchasing behavior features, whereas LOF more frequently captures anomalies related to activity intensity and localized customer characteristics. These findings emphasize that the selection of an outlier detection algorithm should be aligned with the analytical objectives and the underlying characteristics of the data being examined.

Overall, this study concludes that adopting a comparative approach between LOF and EE provides a more comprehensive understanding of customer anomaly patterns than relying on a single algorithm. The results of this study are expected to serve as a foundation for selecting appropriate outlier detection methods and to open avenues for future research, such as integrating local and global detection approaches or applying alternative outlier detection algorithms to different data contexts.

## LIMITATION

This study has several limitations that should be considered when interpreting the obtained results. First, the outlier detection analysis was conducted using only two algorithms, namely the Local Outlier Factor (LOF) and Elliptic Envelope (EE). Although these methods represent local density-based and global distribution-based approaches, respectively, the findings do not reflect the performance of other outlier detection algorithms that may exhibit different characteristics and sensitivity levels.

Second, this study employed a single customer dataset with specific characteristics; therefore, the results are context-dependent and cannot be readily generalized to other data types or domains. Variations in data structure, feature distributions, and variable scales across different datasets may lead to distinct outlier patterns.

Third, parameter selection for each algorithm was based on default values or commonly adopted settings, without systematic parameter optimization. This condition may influence both the number and characteristics of the detected outliers, meaning that the results may not represent the optimal performance of each algorithm.

Fourth, the evaluation of outlier detection results was conducted using a comparative and descriptive approach, without the availability of ground truth or definitive anomaly labels. Consequently, this study was unable to quantitatively assess detection accuracy or

performance, and the interpretation of results relied primarily on feature analysis and data visualization.

These limitations provide opportunities for future research to develop more comprehensive analyses, such as incorporating additional outlier detection algorithms, performing systematic parameter optimization, utilizing more diverse datasets, and applying more objective evaluation methods.

## REFERENCES

- Abu-Dalbouh, H. M., Freihat, M. M., Jawarneh, R. I., Salim, M. A. M., & Alateyah, S. A. (2025). Analyzing consumer decision-making in digital environments using random forest algorithm and statistical methods. *International Journal of Advanced Computer Science and Applications*, 16(3), 1186–1200. <https://doi.org/10.14569/IJACSA.2025.01603114>
- Alfian, G., et al. (2023). Customer shopping behavior analysis using RFID and machine learning models. *MDPI*, 14.
- Amirillah, C. D. R. (2025). Detecting fraudulent transaction in banking sector using rule-based model and machine learning. *Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI)*, 14(2), 96–102.
- Bajpai, A., Khule, S., Sharma, V. P., et al. (2024). Real-time anomaly detection in big data streams: Machine learning approaches and performance evaluation. *International Journal of Intelligent Systems and Applications in Engineering*, 12(21s), 915–924.
- Deniz, E., & Bülbül, S. Ç. (2024). Predicting customer purchase behavior using machine learning models. *Information Technology, Economics and Business*, 1(1), 1–6.
- Hodge, V. J., & Austin, J. (2018). An evaluation of classification and outlier detection algorithms. *arXiv*.
- Ifmaini, A. F. C., Beng, J. T., & Perdana, N. J. (2025). Analisis deskriptif dan deteksi outlier pada data penjualan produk di marketplace. *INTECOMS: Journal of Information Technology and Computer Science*, 8(6).
- Kurniawan, R., & Utomo, D. S. (2022). Analisis perilaku konsumen berbasis data transaksi menggunakan teknik data mining. *Jurnal Sistem Informasi*, 18(2), 134–143.
- Nugroho, A., Wibowo, A., & Malik, R. F. (2023). Penerapan machine learning untuk analisis perilaku pelanggan ritel berbasis big data. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, 9(4), 689–698.
- Pratama, F. A., Nugroho, A. S., & Setiawan, H. (2023). Analisis data transaksi pelanggan menggunakan pendekatan unsupervised learning. *Jurnal Teknologi dan Sistem Komputer*, 11(2), 95–103.
- Putri, R. A., Surono, S., & Thobirin, A. (2025). Evaluating LSTM performance on multivariate time series with one-class SVM outlier detection. *JUITA: Jurnal Informatika*, 13(2), 127–134.
- Ramadhan, M. R., & Sutoyo, E. (2023). Deteksi anomali data transaksi menggunakan local outlier factor. *Jurnal Informatika Universitas Pamulang*, 8(3), 211–218.

- Riski, G., Hartama, D., & Solikhun. (2025). Evaluation and metrics in outlier detection for machine learning models. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 9(2).
- Wahyuni, S., Rahman, A., & Iqbal, M. (2022). Penerapan elliptic envelope untuk deteksi outlier pada data multivariat. *Jurnal Ilmu Komputer dan Informasi*, 14(1), 45–52.
- Ghodang, K. J., & Hantono, B. S. (2020). Analisis komparatif metode data mining dalam pengolahan data transaksi pelanggan. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 7(3), 455–462.
- Rachmawati, D., Prasetyo, E., & Widodo, A. (2024). Pengaruh normalisasi data terhadap kinerja algoritma berbasis jarak pada analisis data multivariat. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 8(1), 112–120.