


Combining XGBoost and hybrid filtering algorithm in e-commerce recommendation system

Vincentius Loanka Sinaga, Antoni Wibowo

Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info	ABSTRACT
<p>Article history:</p> <p>Received Nov 12, 2024 Revised Apr 7, 2025 Accepted Apr 23, 2025</p> <p>Keywords:</p> <p>Collaborative filtering Content-based filtering E-commerce Recommendation systems XGBoost</p>	<p>This study proposes a hybrid filtering algorithm (HFA) that combines extreme gradient boosting (XGBoost), content-based filtering (CBF), and collaborative filtering (CF) to improve recommendation accuracy in electronic commerce (e-commerce). XGBoost first leverages demographic data (e.g., age, gender, and location) to address cold start conditions, producing an initial product prediction; CBF refines this prediction by measuring product similarities through term frequency-inverse document frequency (TF-IDF) and cosine similarity, while CF (implemented via singular value decomposition) further incorporates user interaction patterns to enhance recommendations. Experimental results across multiple datasets demonstrate that HFA consistently outperforms standalone XGBoost in key metrics, including precision, F1-score, and hit ratio (HR). HFA's precision often exceeds 90%, indicating fewer irrelevant recommendations. Although recall levels remain modest, HFA exhibits stronger adaptability under cold start scenarios due to its reliance on demographic features and user-item interactions. These findings highlight the efficacy of combining advanced machine learning with hybrid filtering techniques, offering a more robust and context-aware solution for e-commerce recommendation systems.</p> <p><i>This is an open access article under the CC BY-SA license.</i></p> <div></div>
<p>Corresponding Author:</p> <p>Antoni Wibowo Graduate Program, Master of Computer Science, Bina Nusantara University Kebon Jeruk Highway, Kemanggisian, Palmerah, Jakarta Barat 11480, Indonesia Email: anwibowo@binus.edu</p>	

1. INTRODUCTION

Electronic commerce (e-commerce) refers to the purchasing and selling of information, goods, and services via the internet [1], [2]. With the advancement of the internet and smart devices, e-commerce has become an integral part of daily life, offering a wide range of products with significant variation, which can make it challenging for users to choose items that match their preferences. To address this, recommendation systems have been developed, mimicking natural social behaviors such as word-of-mouth suggestions to guide user decisions by predicting future preferences based on past evaluations [3], [4]. Two widely used approaches in recommendation systems are content-based filtering (CBF) and collaborative filtering (CF). CBF suggests items with similar attributes to those a user has previously shown interest in [5], [6], while CF recommends products by analyzing similarities between users based on their interaction history [7], [8].

However, both methods have limitations. CBF can suffer from overspecialization, offering overly similar suggestions, while CF is susceptible to data sparsity and cold start problems, particularly when dealing with new users lacking sufficient historical data [9]. To overcome these issues, hybrid systems that combine CF and CBF have been proposed. Sharma *et al.* [10] integrated CF and CBF, achieving a significantly lower mean absolute error (MAE) than either method alone, while Li *et al.* [11] and

Bahl *et al.* [12] further confirmed that hybrid CF-CBF systems outperform standalone approaches. Nevertheless, these models still depend on existing user interaction data and do not fully resolve the cold start problem. Advancements in CF, such as integrating singular value decomposition (SVD), have improved accuracy and reduced sparsity [13], [14], yet cold start scenarios remain a challenge due to insufficient user interaction data.

Parallel improvements in CBF have been made through enhanced similarity measures and integration with machine learning. Abdurrafi and Ningsih [15] used cosine similarity with CBF to achieve high precision, while Shahbazi *et al.* [16] enhanced CBF by integrating it with extreme gradient boosting (XGBoost), achieving superior accuracy over various machine learning models. Similarly, Malek *et al.* [17] demonstrated XGBoost's strength in handling imbalanced data, reinforcing its value as an initial predictor. Despite promising results from hybrid CF-CBF systems and XGBoost-enhanced CBF models, gaps remain—particularly in incorporating demographic features (e.g., age, gender, and location) and fully integrating XGBoost with both CBF and CF to balance content and user behavior data.

Based on these insights, this study proposes a novel hybrid filtering algorithm (HFA) that integrates XGBoost, CBF, and CF-SVD to address the cold start problem and enhance recommendation accuracy. XGBoost functions as the initial predictor, utilizing demographic data, followed by refinement through CBF's content similarity and CF-SVD's interaction-based recommendations. The final recommendation output is generated through a weighted scoring mechanism that balances relevance and diversity. Compared to standalone XGBoost, this integrated framework consistently achieves higher precision, F1-scores, and hit ratios (HRs) across various datasets and scenarios. The remainder of this paper is structured as follows: section 2 reviews the literature, section 3 presents the methodology, section 4 discusses experimental results, and section 5 concludes with future research directions.

2. RESEARCH METHOD

This study has introduced a HFA that combines XGBoost, CBF, and CF to improve recommendation accuracy in e-commerce, as illustrated in Figure 1. Initially, XGBoost utilizes demographic data, such as age, gender, and location, to make preliminary product recommendations for new users. Subsequently, CBF refines these recommendations by assessing the similarity between products using term frequency-inverse document frequency (TF-IDF) and cosine similarity. Finally, CF employs singular value decomposition to incorporate user interactions with products, further improving the overall recommendations.

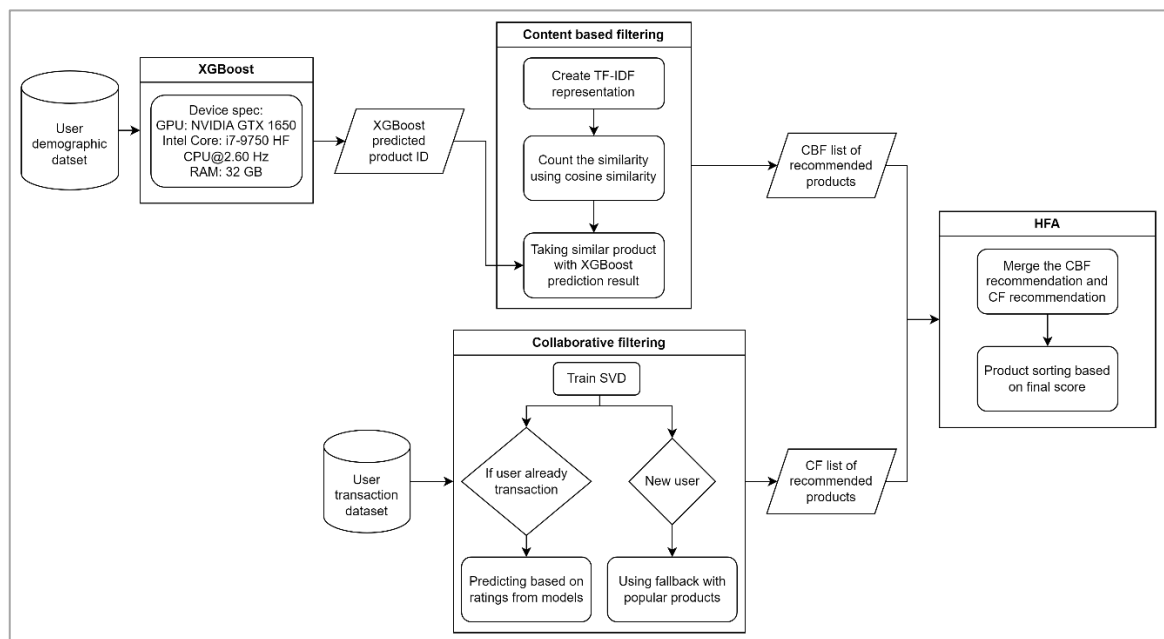


Figure 1. The proposed method

2.1. Dataset

The first dataset was obtained from Kaggle and is titled “E-commerce sales data 2023-24” [18]. This dataset consists of transaction data from 2024 (E-commerce sales data 2024), product details (product_detail), and user-related data (customer_detail). The second dataset used is titled “Brazilian_ecommerce_analyses_v2” [19]. The transaction dataset is a combination of olist_order_dataset, olist_order_items, and olist_order_reviews_dataset. However, product details are sourced from olist_products_dataset, while user details are extracted from olist_customer_dataset. The third dataset is titled “Ecommerce_bigQuery” [20]. This dataset has a structure like the “E-commerce sales data 2023-24” dataset but with different naming conventions. For instance, the transaction dataset is labeled “order_item,” the product dataset is named “product_old,” and the user dataset is referred to as “user_old.”

Although all three datasets contain similar e-commerce elements, focusing on transactions, products, and customers, they differ in data organization and naming conventions. “Brazilian_ecommerce_analyses_v2” has the most comprehensive structure, including review data and integrated transaction information. However, since the datasets used are not ideal compared to those utilized by real-world e-commerce platforms, the authors have applied several modifications to enhance their suitability for this study.

2.2. Data pre-processing

The process begins with raw data, which is first preprocessed to ensure quality and relevance. This pre-processing stage typically includes cleaning data, handling missing values, encoding categorical values, and selecting features. Once pre-processed, the data is ready for the model training and evaluation phase. For the process, the authors will use the first dataset for simulation.

2.2.1. Handling missing value

We utilize a two-step process for addressing missing values specifically within the product dataframe. The first step involves eliminating columns that contain NaN (missing) or null values, as detailed in Table 1. The second step consists of removing columns that are not relevant for input into the model, as outlined in Table 2.

Table 1. List of products dataframe columns that mostly contain NaN or null values

Brand Name	Asin	List Price
Quantity	SKU	Stock
Product details	Dimension	Color
Ingredients	Direction to use	Size quantity variant
	Product description	

Table 2. List of products dataframe columns that have no relevance to the model

Variants	Product url	Image
	Is Amazon seller	

2.2.2. Create main transaction dataframe and split process

The process of unification is critical as it necessitates a reference dataframe that includes the complete transaction history of the system. Developing a primary transaction dataframe and managing the split process within a workflow for data merging is a fundamental technique in data pre-processing. This approach is particularly relevant when dealing with datasets that contain transactional information. Figure 2 illustrates the unification process visually.

After constructing the complete transaction dataset in the full_transaction_df dataframe, the next step is to derive three specialized dataframes: rating, product, and user, which are essential for preparing the data for model training. These dataframes are created by selectively extracting and transforming relevant subsets of information from the comprehensive transactional data. The rating dataframe typically includes user-product interaction data, such as ratings or purchase frequency, serving as a critical input for recommendation algorithms. The product dataframe captures details about individual products, including product IDs, names, categories, and other descriptive attributes. Similarly, the user dataframe aggregates user-related information, such as user IDs and demographic or behavioral features when available. By organizing the data into these distinct but interrelated structures, the modeling process becomes more streamlined and interpretable. The composition and structure of each of these dataframes are outlined in Table 3, offering a clear overview of the attributes included and their respective roles in the recommendation system pipeline.

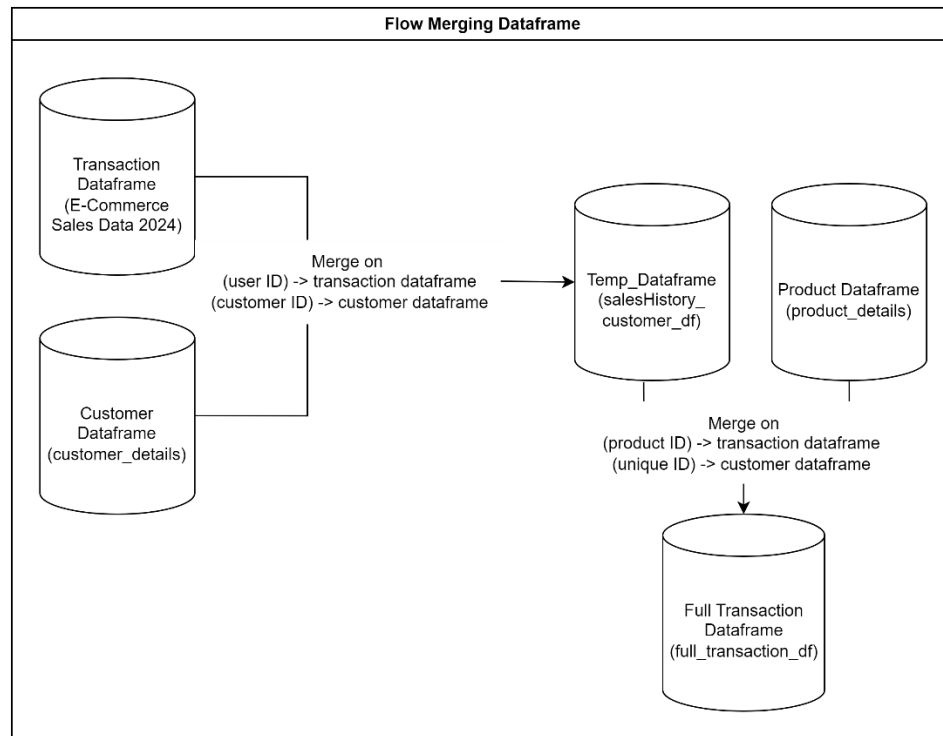


Figure 2. Flow merging dataframe

Table 3. Composition of rating, product, and user dataframe formation

Dataframe	Component	Data Type (dataframe)
Df_rating	User ID	Int32
	Product ID	Int32
	Review rating	Float64
Df_user	User ID	Int32
	Age	Int64
	Gender	Int32
	Location	Int32
Df_Product	Product ID	Int32
	Product name	Object

2.3. XGBoost

XGBoost is a scalable end-to-end tree [21]. In this context, the XGBoost model predicts a single product (product ID) deemed most relevant for each user based on demographic features. This prediction serves as the "seed" or starting point in the hybrid recommendation system, allowing the results to be further refined using the CBF approach. The output of this model consists of a list of predicted product IDs, which serve as input for the following recommendation stage. The role of XGBoost in this hybrid model is as follows, i) Pre-prediction of products: before generating further recommendations (e.g., identifying similar products or estimating ratings using CF), the system aims to determine an initial product that is likely to be preferred by the user and ii) Utilization of demographic features: features such as age, gender, and location often serve as initial indicators of product preferences. The XGBoost model processes this data and maps it to a specific product.

2.4. Content-based filtering with TF-IDF and cosine similarity

In this research, CBF will employ TF-IDF and cosine similarity. TF-IDF converts product text data into numerical vectors by calculating the product of term frequency and inverse document frequency [22]. Cosine similarity then quantifies the angular distance between these vectors, reflecting how closely related two products are [23]. The initial step involves utilizing TfidfVectorizer to transform text data into numerical vector representations. After obtaining the TF-IDF representations for each product, we compute cosine similarity to assess product similarity. After establishing the similarity matrix, we proceed to identify products that are similar to a reference product, which is the product that XGBoost predicts.

2.5. Collaborative filtering using SVD (surprise)

CF in this study is implemented via SVD, a matrix factorization technique that decomposes the user-item rating matrix into latent factors [24]. Once the SVD model is trained, it predicts ratings for each (user, product) pair. If a user is known (present in the training data), the system applies the predicted ratings; for cold start scenarios (new users), it returns to recommending popular products.

2.6. Hybrid model

HFA is a recommendation system that integrates two or more approaches (CBF and CF in this study) to leverage the strengths of each method while mitigating the limitations of individual techniques. Based on Figure 1, the proposed hybrid system consists of three main components, i) XGBoost as pre-prediction of products: the XGBoost model generates an initial prediction, identifying the most relevant product based on user features such as age, gender, and location. This component produces an initial recommendation, acting as a “seed” for finding similar items; ii) CBF: once XGBoost has identified a product, CBF is used to find similar products to the predicted item. This approach employs TF-IDF to convert product descriptions or names into numerical vector representations and cosine similarity to measure the degree of similarity between products. The CBF component returns a list of products that share content-based similarities with the initial product and their respective similarity scores; and iii) CF: the CF component utilizes the SVD model from the surprise library to predict user ratings for each product. If a user has an interaction history (e.g., ratings or past purchases), CF estimates the likelihood of the user preferring certain products. If the user is new and does not exist in the training data (cold start scenario), the system applies a fallback strategy, recommending popular products. This ensures the generation of predicted ratings (preference scores) for each product, from which the highest-scoring products are selected.

After getting the output from each of the above components, the next step is to combine the scores from CBF and CF. This is done using the weighted combination formula as in (1).

$$HFA_FinalScore = \frac{(\alpha * 0.5) + (\beta * 0.5)}{2} \quad (1)$$

Where α is CF score and β is CBF score.

2.7. Evaluation

One of the evaluation methods used in this study is the HR. HR is a metric used to evaluate the performance of a recommendation system by measuring how often the model predicts products that are relevant to the user [25]. The calculation formula for HR can be seen in (2).

$$HR = \frac{t}{n} \quad (2)$$

Where HR is the hit ratio, t is the number of correct predictions, and n is the total number of user interactions.

In addition to using the HR evaluation method, the study will use the accuracy, precision, recall, and F1-score evaluation methods. This method functions to evaluate the performance of the recommendation system by assessing how accurately the model predicts products that interact with users [26]. The calculation formula for evaluating accuracy, as in (3), precision, as in (4), recall, as in (5), and F1-score, as in (6).

$$AC = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$PCS = \frac{TP}{TP+FP} \quad (4)$$

$$RC = \frac{TN}{TN+FP} \quad (5)$$

$$F1 - score = 2 \times \frac{RC \times PCS}{RC + PCS} \quad (6)$$

Where AC is accuracy, PCS is precision, RC is recall, F1-score is F1-score, TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

3. RESULTS AND DISCUSSION

3.1. Evaluation based on cold start scenario

To evaluate the model's capability in handling cold start cases, particularly the user cold start scenario, the authors conducted an experiment in which the model was required to generate recommendations

for new users. This case was achieved by isolating a subset of users and all their associated transactions. The results of the model evaluation are as follows.

Reviewing Table 4, in handling user cold start cases, the HFA model outperforms in all evaluation metrics. Regarding accuracy, HFA demonstrates a 9% improvement (40% for XGBoost vs. 49% for HFA), indicating that HFA is more accurate in predicting recommendations for new users. Regarding precision, HFA achieves 100%, as the HFA method tends to be highly selective in cold start scenarios, emphasizing CBF and/or CF to minimize noise. From a recall perspective, HFA is more effective in identifying relevant products. Despite the limited interaction data for new users, HFA benefits from CBF and/or CF, which increases the likelihood of finding relevant products. Unlike XGBoost, which relies solely on demographic data, the XGBoost-CBF and/or CF approach enhances the probability of identifying relevant items.

Table 4. XGBoost and HFA comparison by accuracy, precision, recall, F1-score, and HR based on cold start problem

	XGBoost (%)	HFA (%)
Accuracy	40	49
Precision	45	100
Recall	40	49
F1-score	42	65
HR	25.25	29.25

The F1-score analysis indicates a better balance between precision (100%) and recall (49%) in HFA, compared to XGBoost, which has precision (47%) and recall (42%). However, it is important to note that while high precision (100%) ensures highly confident recommendations, the low recall (49%) suggests that the model is overly conservative, focusing only on the most certain recommendations while overlooking some relevant ones. The HR also aligns with the overall performance improvement, as recommendations generated by HFA more frequently match the actual user preferences, even with limited user data.

3.2. Evaluation based on model adaptability

A detailed comparison of the XGBoost and HFA methods is shown in Tables 5 and 6, highlighting how well they perform on three different e-commerce datasets: e-commerce sales data 2023–24, Brazil e-commerce analysis v2, and Ecommerce_bigQuery. These datasets were chosen to represent various market contexts and data characteristics, facilitating a thorough evaluation of the model's effectiveness. The comparison emphasizes several key performance metrics commonly utilized in classification and recommendation tasks, including accuracy, precision, recall, F1-score, and average HR. Accuracy measures the overall correctness of the model's predictions, while precision and recall offer information about the model's ability to accurately identify relevant instances and retrieve all potential relevant results, respectively. The F1-score, which is the harmonic mean of precision and recall, provides a balanced perspective on both metrics. Lastly, the average HR assesses the success rate of the recommendation component, indicating how frequently the recommended items align with actual user interactions. By looking at these metrics together, the tables provide a straightforward way to compare the strengths and weaknesses of the XGBoost and HFA methods in different e-commerce data situations.

The HFA method consistently does better than XGBoost in all three datasets e-commerce sales data 2023–24, Brazil e-commerce analysis v2, and Ecommerce_bigQuery especially in terms of accuracy, precision, and average HR. Among these metrics, precision demonstrates the most significant and consistent enhancement, with HFA surpassing XGBoost in every dataset. Although the recall and accuracy may see slight increases, remain stable, or experience marginal declines depending on the dataset, these fluctuations are minor compared to the consistently elevated precision achieved by HFA. Consequently, the F1-score is generally higher for HFA due to its strong precision component. The average hit ratio also sees improvements with HFA, although the extent of this improvement varies by dataset. In the Ecommerce_bigQuery dataset, HFA shows a slightly lower accuracy of 21% compared to XGBoost's 22%.

Across all datasets, HFA achieves an impressive precision range of 92–98%, signifying its capability to generate more accurate and targeted recommendations while minimizing false positives. This performance indicates that HFA, which combines XGBoost with content-based and collaborative filtering techniques, implements a more stringent selection process when recommending items. While this approach decreases the number of incorrect predictions, it also creates a noticeable disparity between precision and recall, suggesting that some relevant items may not be included. This trade-off underscores HFA's focus on delivering high-quality recommendations over maximizing coverage, making it particularly effective in scenarios where relevance and user satisfaction are paramount.

Table 5. XGBoost test results with different datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Average HR (%)
E-commerce sales data 2023-24	32	47	32	38	27.5
Brazilian_ecommerce_analyses_v2	26	74	23	31	24.6
Ecommerce_bigQuery dataset	22	78	22	29	24.2

Table 6. HFA test results with different datasets

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Average HR (%)
E-commerce sales data 2023-24	33	92	33	46	34.4
Brazilian_ecommerce_analyses_v2	26	94	25	36	32.4
Ecommerce_bigQuery dataset	21	98	21	32	26.5

4. CONCLUSION

The HFA-integrating XGBoost, CBF, and CF consistently outperform standalone XGBoost across various evaluation metrics. By combining content similarity with user preference data, HFA achieves higher precision, a better F1-score, and an improved HR in both general and cold start scenarios. This approach adapts well to different datasets, providing more stable and relevant recommendations, although there remains a need to enhance recall and scalability for large datasets further. Overall, HFA's performance advantages make it a more reliable choice for complex recommendation systems than XGBoost, which relies heavily on static features and exhibits less stable performance. With additional refinements aimed at improving recall and optimizing cold start handling, HFA has the potential to evolve into a robust solution that effectively capitalizes on the strengths of both content-based and CF. Next, researchers can explore transitioning from traditional CF to neural collaborative filtering (NCF), which uses specialized embeddings for user-product interactions and neural networks for more accurate predictions. CBF can also benefit from advanced language models like bidirectional encoder representations from transformers (BERT) or other transformer-based approaches to capture deeper contextual and semantic nuances in product descriptions.

ACKNOWLEDGEMENTS

Authors would like to thank Bina Nusantara University for providing support for this research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Vincentius Loanka Sinaga	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Antoni Wibowo	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board or equivalent committee.

DATA AVAILABILITY




All data used in this study are public datasets.

REFERENCES




- [1] M. Kütz, *Introduction to E-commerce: combining business and information technology*, Bookboon, 2016.
- [2] V. Jain, B. Malviya, and S. Arya, "An overview of electronic commerce (e-commerce)," *Journal of Contemporary Issues in Business and Government*, vol. 27, no. 3, 2021, doi: 10.47750/cibg.2021.27.03.090.
- [3] F. T. A. Hussien, A. M. S. Rahma, and H. B. A. Wahab, "Recommendation systems for e-commerce systems an overview," *Journal of Physics: Conference Series*, vol. 1897, no. 1, 2021, doi: 10.1088/1742-6596/1897/1/012024.
- [4] L. Lü, M. Medo, C. H. Yeung, Y. C. Zhang, Z. K. Zhang, and T. Zhou, "Recommender systems," *Physics Reports*, vol. 519, no. 1, pp. 1–49, 2012, doi: 10.1016/j.physrep.2012.02.006.
- [5] A. Nurcahya and S. Supriyanto, "Content-based recommender system architecture for similar e-commerce products," *Jurnal Informatika*, vol. 14, no. 3, p. 90, 2020, doi: 10.26555/jifo.v14i3.a18511.
- [6] D. Wang, Y. Liang, D. Xu, X. Feng, and R. Guan, "A content-based recommender system for computer science publications," *Knowledge-Based Systems*, vol. 157, pp. 1–9, 2018, doi: 10.1016/j.knosys.2018.05.001.
- [7] M. Anbazhagan and M. Arock, "A study and analysis of collaborative filtering algorithms for recommender systems," *International Journal of Control Theory and Applications*, vol. 9, no. 27, pp. 127–136, 2016.
- [8] K. Y. Jung, D. H. Park, and J. H. Lee, "Hybrid collaborative filtering and content-based filtering for improved recommender system," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3036, pp. 295–302, 2004, doi: 10.1007/978-3-540-24685-5_37.
- [9] S. Eliyas and P. Ranjana, "Recommendation systems: content-based filtering vs collaborative filtering," *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering, ICACITE 2022*, pp. 1360–1365, 2022, doi: 10.1109/ICACITE53722.2022.9823730.
- [10] S. Sharma, V. Rana, and M. Malhotra, "Automatic recommendation system based on hybrid filtering algorithm," *Education and Information Technologies*, vol. 27, no. 2, pp. 1523–1538, 2022, doi: 10.1007/s10639-021-10643-8.
- [11] L. Li, Z. Zhang, and S. Zhang, "Hybrid algorithm based on content and collaborative filtering in recommendation system optimization and simulation," *Scientific Programming*, vol. 2021, 2021, doi: 10.1155/2021/7427409.
- [12] D. Bahl, V. Kain, A. Sharma, and M. Sharma, "A novel hybrid approach towards movie recommender systems," *Journal of Statistics and Management Systems*, vol. 23, no. 6, pp. 1049–1058, 2020, doi: 10.1080/09720510.2020.1799503.
- [13] W. Hong-Xia, "An improved collaborative filtering recommendation algorithm," *2019 4th IEEE International Conference on Big Data Analytics, ICBDA 2019*, pp. 431–435, 2019, doi: 10.1109/ICBDA.2019.8713205.
- [14] R. M. Sallam, M. Hussein, and H. M. Mousa, "An enhanced collaborative filtering-based approach for recommender systems," *International Journal of Computer Applications*, vol. 176, no. 41, pp. 9–15, 2020, doi: 10.5120/ijca2020920531.
- [15] M. F. Abdurrafi and D. H. U. Ningsih, "Content-based filtering using cosine similarity algorithm for alternative selection on training programs," *Journal of Soft Computing Exploration*, vol. 4, no. 4, pp. 204–212, 2023, doi: 10.52465/josce.v4i4.232.
- [16] Z. Shahbazi, Y. Byun, and Y.-C. Byun, "Product recommendation based on content-based filtering using XGBoost classifier," *International Journal of Advanced Science and Technology*, vol. 29, no. 04, pp. 6979–6988, 2020, [Online]. Available: <https://www.researchgate.net/publication/342864588>.
- [17] N. H. A. Malek, W. F. W. Yaacob, Y. B. Wah, S. A. Md Nasir, N. Shaadan, and S. W. Indratno, "Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 1, pp. 598–608, 2023, doi: 10.11591/ijeecs.v29.i1.pp598-608.
- [18] A. Ali, "E-commerce sales data 2023-24." Kaggle. Accessed Oct. 30, 2024. [Online]. Available: <https://www.kaggle.com/datasets/ahmedaliraja/e-commerce-sales-data-2023-24>
- [19] A. Sionek, "Brazilian_e-commerce_analyses_v2." Kaggle. Accessed Feb. 19, 2025. [Online]. Available: <https://www.kaggle.com/datasets/burayamail/brazilian-e-commerce-analyses-v2>
- [20] C. Givan, "Ecommerce_bigQuery." Kaggle. Accessed Feb. 19, 2025. [Online]. Available: <https://www.kaggle.com/datasets/chiraggivan82/e-commerce-bigquery>
- [21] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
- [22] A. Aizawa, "An information-theoretic perspective of TF-IDF measures," *Information Processing and Management*, vol. 39, no. 1, pp. 45–65, 2003, doi: 10.1016/S0306-4573(02)00021-3.
- [23] A. Huang, "Similarity measures for text document clustering," in *New Zealand Computer Science Research Student Conference, NZCSRSC 2008 - Proceedings*, 2008, pp. 49–56.
- [24] D. Kalman, "A singularly valuable decomposition: the SVD of a matrix," *The College Mathematics Journal*, vol. 27, no. 1, pp. 2–23, 1996, doi: 10.1080/07468342.1996.11973744.
- [25] C. Channarong, C. Paosirikul, S. Maneeroj, and A. Takasu, "HybridBERT4Rec: a hybrid (content-based filtering and collaborative filtering) recommender system based on BERT," *IEEE Access*, vol. 10, pp. 56193–56206, 2022, doi: 10.1109/ACCESS.2022.3177610.
- [26] Z. Fayyaz, M. Ebrahimi, D. Nawara, A. Ibrahim, and R. Kashef, "Recommendation systems: algorithms, challenges, metrics, and business opportunities," *Applied Sciences (Switzerland)*, vol. 10, no. 21, pp. 1–20, 2020, doi: 10.3390/app10217748.

BIOGRAPHIES OF AUTHORS



Vincentius Loanka Sinaga    is a computer science student and researcher at Bina Nusantara University (BINUS), Indonesia. He completed his undergraduate degree at BINUS University in 2024 and was subsequently admitted to the Master of Computer Science program at the BINUS Graduate Program in 2023. During his studies, he participated in the 2022 International Conference on Electrical and Information Technology (IEIT), where he presented a paper titled “Graph attention network on extracting features from simplified molecular-input line-entry system for HIV classification.” His research interests include optimizing e-commerce recommendation systems using multiple ensemble classifiers. He can be reached at vincentius.sinaga@binus.ac.id.



Dr. Eng. Antoni Wibowo, S.Si., M.Kom., M.Eng.    received his first degree in Applied Mathematics in 1995 and master's degree in Computer Science in 2000. In 2003, He was awarded a Japanese Government Scholarship (Monbukagakusho) to attend Master's and PhD programs in Systems and Information Engineering at the University of Tsukuba-Japan. He completed his second master's degree in 2006 and PhD degree in 2009, respectively. His PhD research focused on machine learning, operations research, multivariate statistical analysis, and mathematical programming, especially in developing nonlinear robust regressions using statistical learning theory. He worked from 1997 to 2010 as a researcher in the Agency for the Assessment and Application of Technology-Indonesia. From April 2010-September 2014, he worked as a senior lecturer in the Department of Computer Science-Faculty of Computing, and a researcher in the Operation Business Intelligence (OBI) Research Group, Universiti Teknologi Malaysia (UTM), Malaysia. From October 2014 to October 2016, he was an Associate Professor at the Department of Decision Sciences, School of Quantitative Sciences at Universiti Utara Malaysia (UUM). Dr. Eng. Wibowo is currently working at Binus Graduate Program (Master in Computer Science) in Bina Nusantara University, Indonesia as a Specialist Lecturer and continues his research activities in machine learning, optimization, operations research, multivariate data analysis, data mining, computational intelligence, and artificial intelligence. He can be contacted at email: anwibowo@binus.edu.