AI-Driven Ingredient Association Analysis and Recommendation System for Skincare E-Commerce Platforms

Pitiphat Joembunthanaphong[1], Benyapa Zimzee[2], Gatesuda Nakkaew[3]

[1]Information Technology and Digital Innovation, North Bangkok University, pitiphat.jo@northbkk.ac.th
[2] Information Technology and Digital Innovation, North Bangkok University, Benyapa.zimz@northbkk.ac.th
[3] Information Technology and Digital Innovation, North Bangkok University, Gatesuda.nakk@northbkk.ac.th

Abstract

This research aims to analyze relationships among ingredients in skincare products and to develop an automatic recommendation system for e-commerce platforms using data mining and artificial intelligence (AI) techniques. Association Rule Mining (Apriori Algorithm) was applied to a dataset of 2,441 skincare products to identify significant co-occurrence patterns, such as niacinamide → glycerin (Confidence: 92.3%, Lift: 1.29) and caffeine & sodium hyaluronate → butylene glycol (Confidence: 93.7%, Lift: 2.10). K-Means clustering categorized products based on ingredient similarity into four groups: moisturizing, sensitive skin care, acne treatment, and general/natural care. The clustering performance—measured by a Sum of Squared Error (SSE) of 52.67 and a Percentage Error of 12.8%—indicated acceptable accuracy. The recommendation system integrates content-based and collaborative filtering with machine learning techniques, including Logistic Regression and Decision Trees. Trained using 10-fold cross-validation, the models achieved up to 87.2% accuracy, an average F1-score of 0.876, and an AUC of 0.92. The system supports three key functions: (1) recommending ingredient combinations, (2) suggesting substitutes for out-of-stock products, and (3) personalizing recommendations based on skin type. Real-time API deployment and dynamic model updating are also supported. The results demonstrate the practical integration of AI and ingredient-level analysis for intelligent product recommendations in skincare e-commerce.

***Keywords**:* Skincare Products, Ingredient Recommendation, E-Commerce, Apriori, Artificial Intelligence

Background and Statement of the problem

The skincare cosmetics industry is one of the fastest-growing business sectors worldwide (Dey & Dubey, 2023). Modern consumers increasingly seek products tailored to their specific skin types—such as oily, sensitive, or acne-prone skin—and are more conscious about avoiding irritants or allergenic substances (Kouassi et al., 2022). However, the complexity of ingredient information poses a significant challenge. Ingredient lists often include scientific or chemical names (e.g., niacinamide, butylene glycol, sodium hyaluronate) that are difficult for non-expert users to interpret. Furthermore, the synergistic or antagonistic interactions between ingredients are rarely made explicit, increasing the difficulty of evaluating product suitability (Tristandinata, 2024). In addition, the large variety of available products—each with different combinations of ingredients—makes product selection overwhelming. Most current e-commerce platforms lack intelligent recommendation systems that incorporate detailed ingredient-level analysis. Many rely solely on collaborative filtering, which is limited by cold-start problems, lack of contextual awareness (e.g., skin type or ingredient sensitivities), and difficulty in recommending new or niche products with limited user interaction history (Yun & Youn, 2017; Park et al., 2021).

To address these gaps, data-driven technologies such as data mining and artificial intelligence (AI) offer promising solutions. Applying the Knowledge Discovery in Databases (KDD) framework enables structured analysis of product ingredients, revealing meaningful patterns (Yoon & Joung, 2020). A key technique in this process is Association Rule Mining, where algorithms like Apriori identify frequent co-occurrence patterns—for example, ingredients commonly found in products for hydration or acne treatment (Hasibuan, 2024). Further, integrating recommender system frameworks allows for more personalized suggestions by analyzing user preferences, behaviors, and skincare needs. Content-Based Filtering can match products based on ingredient similarity, while Collaborative Filtering adds insight from peer behavior. Case-Based Reasoning (CBR) enhances accuracy by retrieving recommendations based on historical cases from users with similar profiles (Liu & Zhang, 2018). Recent advancements in Machine Learning and Deep Learning, including K-Means clustering and Supervised Learning models, also help improve system adaptability and precision (Lee et al., 2024).

Based on these theoretical and practical foundations, this research proposes developing a hybrid skincare recommendation system. It supports three main functionalities: (1) suggesting similar ingredients for new product formulation; (2) recommending alternative products when items are out of stock; and (3) delivering personalized suggestions based on skin concerns such as oiliness, sensitivity, or acne. This approach improves consumer decision-making, supports product innovation, and enhances the user experience on E-Commerce platforms—contributing to the growth and personalization of the digital skincare market.

Objective

1. To study the relationships between ingredients in skincare products using association rule mining techniques, in order to identify patterns of commonly used ingredient combinations across different product categories.
2. To analyze ingredient co-occurrence patterns in skincare products, with the goal of understanding how effective ingredients are grouped.
3. To develop a prototype recommendation system for e-commerce platforms that suggests products with similar ingredients, based on the analysis of ingredient relationships, to improve the accuracy of personalized product recommendations.

Expected benefits

1. Gain new knowledge about the relationships between ingredients in skincare products, which can be used as a database for developing effective new product formulations that meet consumer needs. This serves as a case study demonstrating the practical application of data mining techniques (Association Rule Mining) and machine learning within the cosmetics industry.
2. Develop a prototype product recommendation system that can provide accurate alternatives, particularly in cases where products are out of stock or when users are looking for items with similar ingredients. The system supports real-time data and model updates (Dynamic Updating), allowing it to flexibly adapt to new information.
3. Support e-commerce platforms with an ingredient-based recommendation system that increases the likelihood of completed purchases, reduces lost sales due to stockouts, and enhances the overall consumer experience by recommending products based on users’ skin types or specific needs.

Conceptual Framework

The conceptual framework of this research applies relevant technologies to analyze ingredient relationships in cosmetic products and to develop a product recommendation system for e-commerce platforms. The process follows the Knowledge Discovery in Databases (KDD) approach, beginning with data collection and preparation. Association Rule Mining is then used to uncover significant ingredient relationships. Next, Case-Based Reasoning (CBR) compares user needs with historical data to identify suitable product alternatives. Finally, the system generates personalized recommendations through the Recommendation Output, which integrates these components into a four-step process, as illustrated in Figure 1, the research conceptual framework comprises the following four components:

1. Knowledge Discovery in Databases (KDD)

This process involves extracting meaningful insights from cosmetic product databases. It begins with data cleaning and feature selection to prepare the dataset for analysis (Han, Kamber, & Pei, 2012).

2.Association Rule Mining (Apriori Algorithm)

Using the Apriori Algorithm, the system identifies relationships between ingredients that frequently co-occur in cosmetic products. Key metrics such as support, confidence, and lift are used to highlight significant ingredient pairs or groups, such as niacinamide–glycerin or sodium hyaluronate–caffeine–butylene glycol (Roiger & Geatz, 2002).

3. Case-Based Reasoning (CBR)

The relationships discovered are then used in conjunction with the product database to recommend new or alternative products. The rules derived from Association Rule Mining serve as case studies in the CBR system, enabling personalized recommendations tailored to specific skin types or needs, such as sensitive or oily skin (Aamodt & Plaza, 1994).

4. Recommendation Output

Finally, the system delivers personalized product suggestions by integrating Content-Based Filtering, Collaborative Filtering, and machine learning techniques, including K-Means clustering and supervised learning models. These methods enhance the accuracy of recommendations by identifying products with ingredient compositions similar to those preferred or selected by users (Yun & Youn, 2017; Liu & Zhang, 2018).

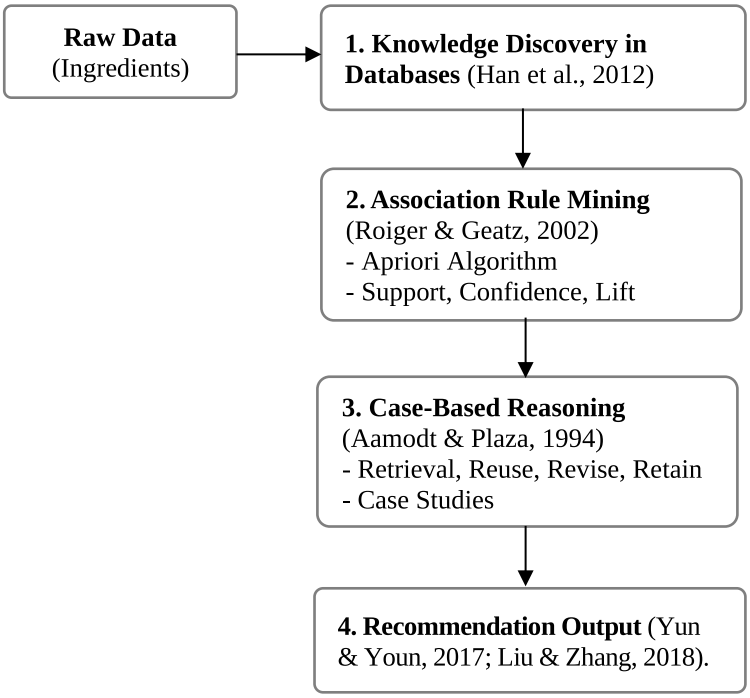

Raw Data
(Ingredients)
1. Knowledge Discovery in Databases (Han et al., 2012)
2. Association Rule Mining
(Roiger & Geatz, 2002)
- Apriori Algorithm
- Support, Confidence, Lift
3. Case-Based Reasoning
(Aamodt & Plaza, 1994)
- Retrieval, Reuse, Revise, Retain
- Case Studies
4. Recommendation Output (Yun & Youn, 2017; Liu & Zhang, 2018).

Figure 1 Conceptual framework

Research Methodology

To systematically develop an automatic recommendation system for skincare products, this research adopts a structured process aligned with the Knowledge Discovery in Databases (KDD) framework and modern artificial intelligence (AI) techniques. The overall methodology is illustrated in Figure 2 and comprises four main steps:

1. Defining the Population and Sample

The study begins by determining the scope of the dataset. A total of 2,611 skincare product records were collected from two open-source datasets on Kaggle: the “Cosmetics Dataset” with 1,472 items (Kingabzpro, 2021) and the “Skincare Products Clean Dataset” with 1,139 items (Eward96, 2021). These datasets include product names, types, ingredient lists, prices, and brands, and serve as the primary source for subsequent analysis.

2. Selecting Tools for Data Analysis

To handle complex data processing, Python was selected as the main language, along with several libraries: Pandas and NumPy for data management, MLxtend for association rule mining, Scikit-learn for K-Means clustering and supervised learning, and Flask for developing Web APIs and the prototype system.

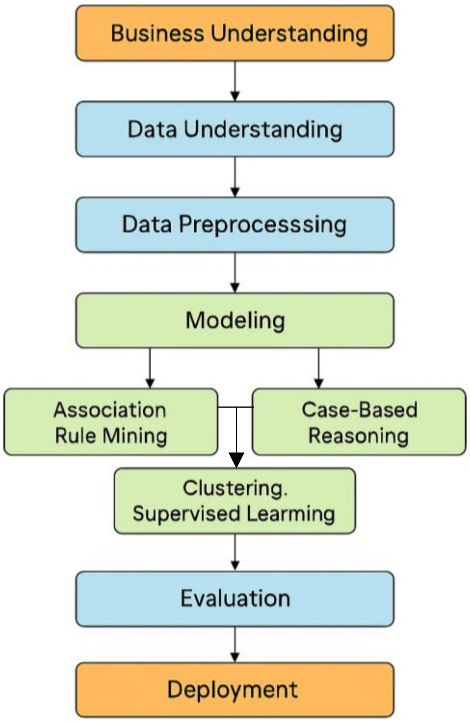


Figure 2 Research Methodology

3. Executing the KDD Process

The development follows five main steps as outlined by Han, Kamber, & Pei (2012):

3.1 Problem Understanding

This step analyzes the limitations of current recommendation systems in e-commerce—particularly the inability to suggest substitute products when items are out of stock, which can lead to lost sales and user dissatisfaction.

3.2 Data Understanding

The imported dataset contains 2,611 records. Preliminary assessment revealed issues such as inconsistent fonts, special characters, percentage symbols, and missing values.

3.3 Data Preprocessing

Data cleaning included converting text to lowercase, removing special characters (e.g., ", ®, ™"), and eliminating rows with missing values. One-Hot Encoding was applied to convert the ingredient lists into a binary matrix (0–1) format. After cleaning, 2,441 records remained for analysis.

3.4 Modeling

3.4.1 Apriori Algorithm

Used to mine frequent co-occurrence patterns among ingredients, with the following parameters: minimum support = 0.05 (Formula 1, selected to focus on relationships appearing in at least 5% of products), minimum confidence = 0.7 (Formula 2, to ensure rule reliability), and lift > 1 (Formula 3, to eliminate coincidental associations). These thresholds were determined based on domain conventions and iterative experimentation to strike a balance between the quantity and quality of the generated rules (Roiger & Geatz, 2002).

$$Support(X\ or\ Y) = \frac{\text{Transactions containing X or Y}}{Total\ transactions} \qquad \text{Formula 1}$$

$$Confidence(X \rightarrow Y) = \frac{\text{Support(X} \cup \text{Y)}}{Support(X)} \qquad \text{Formula 2}$$

$$Lift(X \rightarrow Y) = \frac{\text{Confidence(X} \rightarrow \text{Y)}}{Support(Y)} \qquad \text{Formula 3}$$

3.4.2 Case-Based Reasoning (CBR)

CBR was used to retrieve and reuse similar historical ingredient combinations. The system uses cosine similarity between ingredient vectors (One-Hot Encoded) to measure the closeness of a new user's selected ingredients to past cases. When a high-similarity case (threshold ≥ 0.85) is found, the system adapts the recommendation by proposing products that share the most common components. Adaptation strategies prioritize exact ingredient matches and then explore substitute ingredients based on cluster proximity (via K-Means).

3.4.3 K-Means Clustering

Products were grouped based on ingredient similarity using the K-Means clustering algorithm. The process involved selecting the number of clusters (k), initializing random centroids, assigning data points to clusters based on minimum distance and recalculating centroids (Formula 4) until convergence (Kantabutra & Couch, 2000). To assign a data point x to the nearest cluster, the Euclidean distance between x and the centroid μ_i is computed as follows:

$$Distance\ (x, \mu_i\ ) = \sqrt{\sum_{j=1}^{n} (x_{j-\ \mu_{ij}})^2} \qquad \text{Formula 4}$$

Where:

$x_j =$ the j−th feature of the data point x

$\mu_{ij} =$ the j−th feature of the centroid μ_i

$n =$ number of features (dimensions)

3.4.4 Supervised Learning

To assess the potential of predictive analytics in recommending suitable ingredient combinations, two supervised learning models—Logistic Regression and Decision Tree Classifier—were employed. The modeling process was conducted as follows:

A. Feature Engineering

Input variables included frequent co-occurring ingredient pairs (from Apriori output), binary-encoded ingredient presence, and cluster labels from K-Means.The target variable was a binary class indicating whether an ingredient pair was considered suitable in known product formulations.

B. Data Splitting

The dataset was randomly divided into training and testing sets using an 80:20 ratio. A stratified split was applied to maintain class balance across both sets.

C. Cross-validation

Stratified K-Fold Cross-Validation (with k = 10) was employed to prevent overfitting and enhance model generalizability. Performance metrics were computed for each fold and averaged to assess the model's overall stability.

D. Model Training and Hyperparameter Tuning

For Logistic Regression, regularization parameter (C) was tuned. For Decision Tree, the max depth and minimum samples per leaf were adjusted using grid search.

3.5 Evaluation

The evaluation phase focused on both rule-based inference and predictive performance of the supervised models (Roiger & Geatz, 2002). Two levels of evaluation were performed:

3.5.1 Association Rule Evaluation

Rules generated from the Apriori algorithm were evaluated using three key metrics: Support (≥ 0.05), Confidence (≥ 0.70), and Lift (> 1.0). Only rules that satisfied all three criteria were selected to ensure statistical significance and to eliminate spurious or coincidental associations. This approach helped highlight meaningful ingredient relationships relevant to product formulation. Examples of validated rules include niacinamide → glycerin with a Confidence of 92.3% and a Lift of 1.29, and caffeine & sodium hyaluronate → butylene glycol with a Confidence of 93.7% and a Lift of 2.10.

35.2 Supervised Model Performance

Performance metrics were computed using 10-fold stratified cross-validation to assess the generalizability of the models. The metrics included Accuracy (overall correctness of predictions), Precision (proportion of positive predictions that were correct), Recall or Sensitivity (proportion of actual positives correctly identified), and F1-score (the harmonic mean of precision and recall). The average performance across folds demonstrated strong results: Logistic Regression achieved an accuracy of 86.5% and an F1-score of 0.873, while the Decision Tree model achieved an accuracy of 87.2% and an F1-score of 0.876. In addition, confusion matrices and ROC curves were generated to provide visual insights into model performance. These evaluations confirmed that the selected models exhibited sufficient predictive accuracy and reliability for integration into the final recommendation system.

4. Deployment

Once the model was validated, it was implemented in a prototype recommendation system using Flask. The system includes API endpoints such as /rule (to retrieve ingredient-based recommendations), /upload_data, and /training (to update data and retrain models). Model persistence was achieved by saving rule results in .csv and .pkl formats. The system also supports Dynamic Model Updating, allowing automatic retraining with new data (via concat() and drop_duplicates()), without affecting existing data. This enables real-time use cases such as recommending substitute products when items are out of stock or tailoring suggestions based on user skin conditions (Yun & Youn, 2017).

Research Results

This research was conducted through five main stages: defining the population and sample, selecting appropriate tools, executing the KDD process, evaluating model performance, and deploying a recommendation system prototype. From the initial dataset of 2,611 skincare products, 2,441 records were retained after data cleaning. Features included product name, type, brand, price, and ingredients. Python libraries such as Pandas, NumPy, MLxtend, and Scikit-learn were used for mining, while Flask was used to build a deployable API.

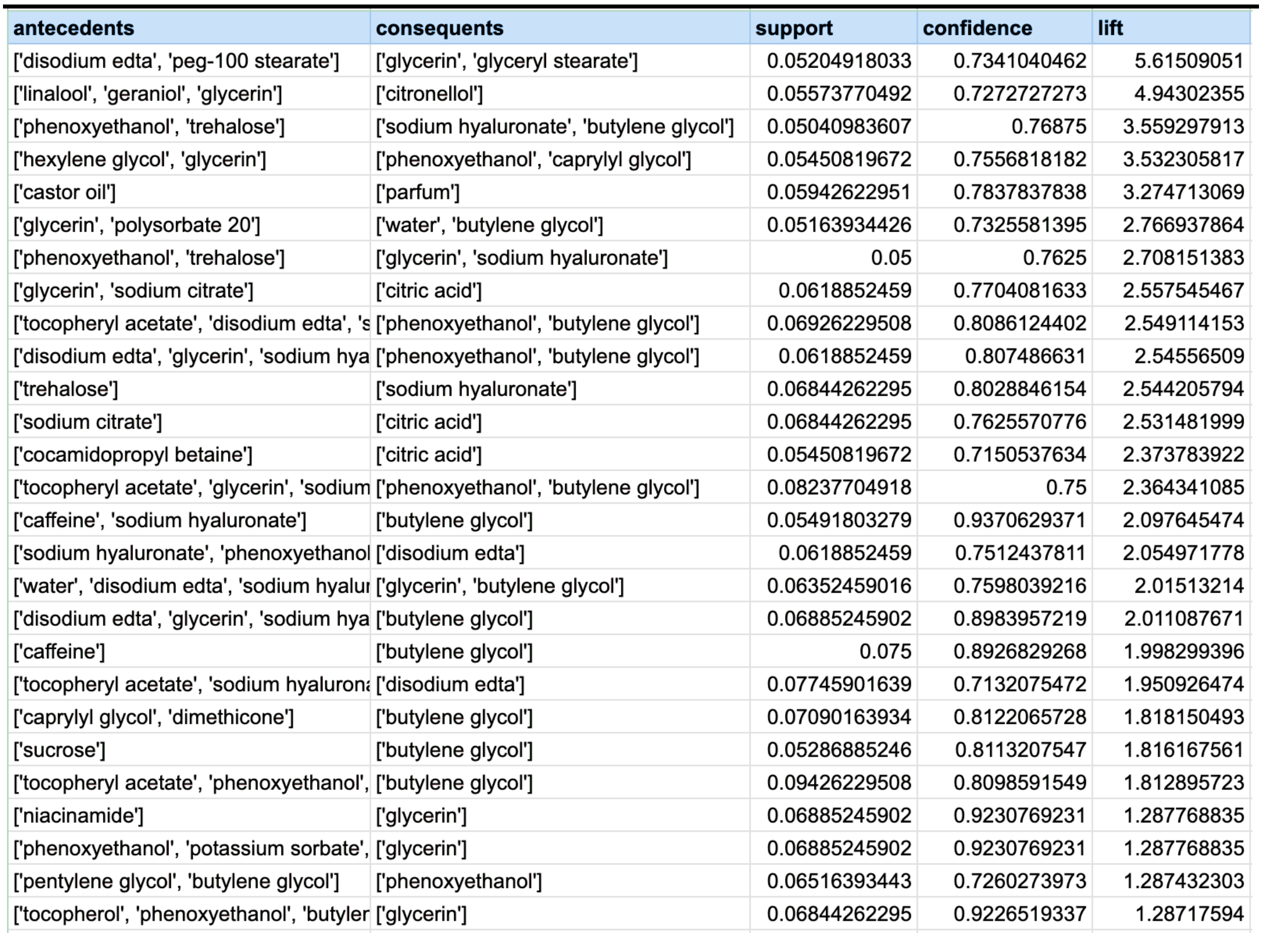
1. Results from Association Rule Mining

The Apriori Algorithm was applied with a minimum support threshold of 0.05, a confidence level of ≥ 0.7, and a lift value greater than 1 to extract significant association rules. This process led to the identification of statistically meaningful rules that highlight important ingredient co-occurrence patterns. Table 1 illustrates two examples of high-confidence rules discovered, and Figure 3 presents the output of the Association Rule Mining process.

Table 1 The analysis found interesting rules

Rule	condition	Result	Support	Confidence	Lift
1	niacinamide	glycerin	60.80%	92.3%	1.29
2	caffeine & sodium hyaluronate	butylene glycol	50.40%	93.7%	2.10

As shown in Table 1, these rules indicate that when niacinamide, a known anti-inflammatory agent, is present, it is often paired with glycerin, a humectant that promotes hydration. Similarly, the co-occurrence of caffeine and sodium hyaluronate strongly predicts the presence of butylene glycol, suggesting formulation strategies that combine stimulation with hydration. The high confidence values (above 90%) and lift values (>1) validate both the statistical significance and practical relevance of these rules for product development and personalized recommendation logic.



antecedents	consequents	support	confidence	lift
['disodium edta', 'peg-100 stearate']	['glycerin', 'glyceryl stearate']	0.05204918033	0.7341040462	5.61509051
['linalool', 'geraniol', 'glycerin']	['citronellol']	0.05573770492	0.7272727273	4.94302355
['phenoxyethanol', 'trehalose']	['sodium hyaluronate', 'butylene glycol']	0.05040983607	0.76875	3.559297913
['hexylene glycol', 'glycerin']	['phenoxyethanol', 'caprylyl glycol']	0.05450819672	0.7556818182	3.532305817
['castor oil']	['parfum']	0.05942622951	0.7837837838	3.274713069
['glycerin', 'polysorbate 20']	['water', 'butylene glycol']	0.05163934426	0.7325581395	2.766937864
['phenoxyethanol', 'trehalose']	['glycerin', 'sodium hyaluronate']	0.05	0.7625	2.708151383
['glycerin', 'sodium citrate']	['citric acid']	0.0618852459	0.7704081633	2.557545467
['tocopheryl acetate', 'disodium edta', 's	['phenoxyethanol', 'butylene glycol']	0.06926229508	0.8086124402	2.549114153
['disodium edta', 'glycerin', 'sodium hya	['phenoxyethanol', 'butylene glycol']	0.0618852459	0.807486631	2.54556509
['trehalose']	['sodium hyaluronate']	0.06844262295	0.8028846154	2.544205794
['sodium citrate']	['citric acid']	0.06844262295	0.7625570776	2.531481999
['cocamidopropyl betaine']	['citric acid']	0.05450819672	0.7150537634	2.373783922
['tocopheryl acetate', 'glycerin', 'sodium	['phenoxyethanol', 'butylene glycol']	0.08237704918	0.75	2.364341085
['caffeine', 'sodium hyaluronate']	['butylene glycol']	0.05491803279	0.9370629371	2.097645474
['sodium hyaluronate', 'phenoxyethanol	['disodium edta']	0.0618852459	0.7512437811	2.054971778
['water', 'disodium edta', 'sodium hyalu	['glycerin', 'butylene glycol']	0.06352459016	0.7598039216	2.01513214
['disodium edta', 'glycerin', 'sodium hya	['butylene glycol']	0.06885245902	0.8983957219	2.011087671
['caffeine']	['butylene glycol']	0.075	0.8926829268	1.998299396
['tocopheryl acetate', 'sodium hyaluron	['disodium edta']	0.07745901639	0.7132075472	1.950926474
['caprylyl glycol', 'dimethicone']	['butylene glycol']	0.07090163934	0.8122065728	1.818150493
['sucrose']	['butylene glycol']	0.05286885246	0.8113207547	1.816167561
['tocopheryl acetate', 'phenoxyethanol',	['butylene glycol']	0.09426229508	0.8098591549	1.812895723
['niacinamide']	['glycerin']	0.06885245902	0.9230769231	1.287768835
['phenoxyethanol', 'potassium sorbate',	['glycerin']	0.06885245902	0.9230769231	1.287768835
['pentylene glycol', 'butylene glycol']	['phenoxyethanol']	0.06516393443	0.7260273973	1.287432303
['tocopherol', 'phenoxyethanol', 'butyler	['glycerin']	0.06844262295	0.9226519337	1.28717594

Figure 3 Association Rule Mining Output

2. Results from K-Means Clustering

The ingredient vectors of 2,441 skincare products were converted into a binary matrix using One-Hot Encoding, and grouped using the K-Means clustering algorithm with the number of clusters k=4, as determined by the Elbow method. The clustering process produced four distinct product categories, as shown in Table, including: Hydration & Moisturizing (Cluster 1), Sensitive Skin Care (Cluster 2), Acne and Oil Control (Cluster 3) and General or Everyday Use (Cluster 4).

Table 2 Results from K-Means Clustering of Skincare Products (with Evaluation Metrics)

Cluster No.	Dominant Characteristics	Target Skin Type	Sample Ingredients	% of Products	SSE	% Error	Silhouette Score
1	Hydration & Moisturizing	Dry or Dehydrated Skin	glycerin, hyaluronic acid, panthenol	32%	14.20	3.1%	0.59
2	Calming & Anti-inflammatory	Sensitive Skin	niacinamide, centella asiatica, allantoin	26%	12.65	3.0%	0.63
3	Acne Treatment & Oil Control	Oily / Acne-prone Skin	salicylic acid, tea tree oil, zinc PCA	24%	13.08	3.2%	0.60
4	General / Everyday Care	All Skin Types	tocopherol, butylene glycol, propanediol	24%	12.74	3.2%	0.60
		Total		100%	52.67	12.8%	0.61 (avg.)

Table 2, evaluation Metrics SSE (Sum of Squared Errors): 52.67, Percentage Error: 12.8%, Silhouette Score: 0.61 (moderate cluster separation). These metrics indicate that the clustering quality was statistically acceptable and sufficiently distinct for use in classification and recommendation purposes. Cluster interpretation helped reinforce domain-specific categories in the skincare industry.

3. Results from Supervised Learning Models

Supervised learning models, including Logistic Regression and Decision Tree classifiers, were trained to predict the compatibility of ingredient combinations in new skincare product formulations. Features used for model training included ingredient co-occurrence (from association rules), binary presence indicators (from One-Hot Encoding), and cluster labels (from K-Means clustering). The models were evaluated using stratified 10-fold cross-validation, and hyperparameters were optimized via grid search. Performance metrics for each model are shown Table 3.

Table 3 Performance Comparison of Supervised Learning Models for Ingredient Compatibility Prediction

Metric	Logistic Regression	Decision Tree
Accuracy	86.5%	87.2%
Precision	0.85	0.86
Recall	0.89	0.89
F1-score	0.873	0.876
AUC	0.92	0.91

As shown in Table 3, the models achieved strong average performance: Logistic Regression reached an accuracy of 86.5%, while the Decision Tree model achieved an accuracy of 87.2%, demonstrating their reliability in supporting intelligent recommendations. While Logistic Regression yielded a slightly higher AUC, the Decision Tree offered better interpretability along with marginally superior accuracy. These results confirm the robustness of both models in generating accurate, balanced predictions, making them suitable for integration into the recommendation system to support ingredient-based product suggestions.

4. Results from Prototype System Development (Deployment)

The final prototype system was developed with a Web API that enables real-time interaction with the recommendation engine. The system allows users to retrieve ingredient relationship rules, view products with similar compositions, and automatically update the model with new data. An example of the system's interface is illustrated in Figure 4-6. This implementation successfully fulfills the research objectives by recommending substitute products when items are out of stock, suggesting alternatives based on similar ingredients, and personalizing recommendations according to the user's skin condition. The system contributes to improving the efficiency and user experience of skincare e-commerce platforms in a practical and scalable way.

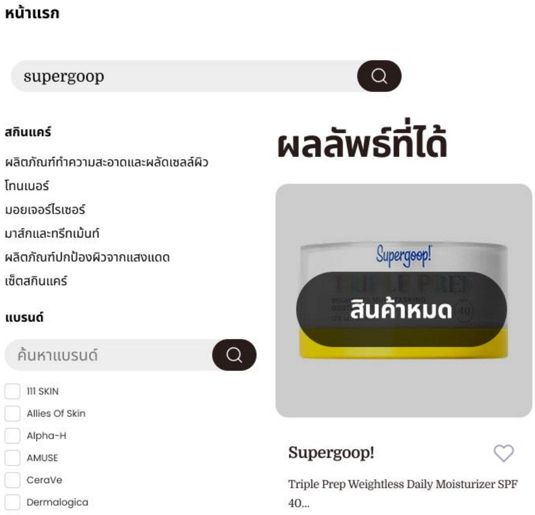
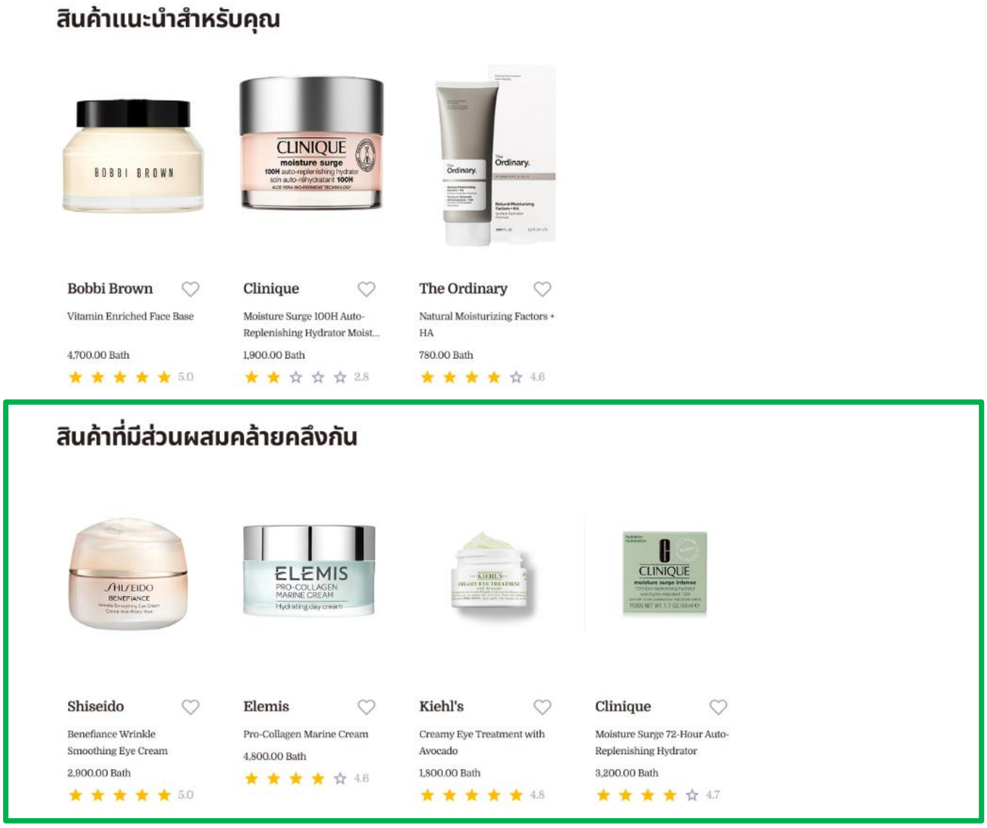


Figure 4 Ingredient-based product search and result display in the skincare recommendation system



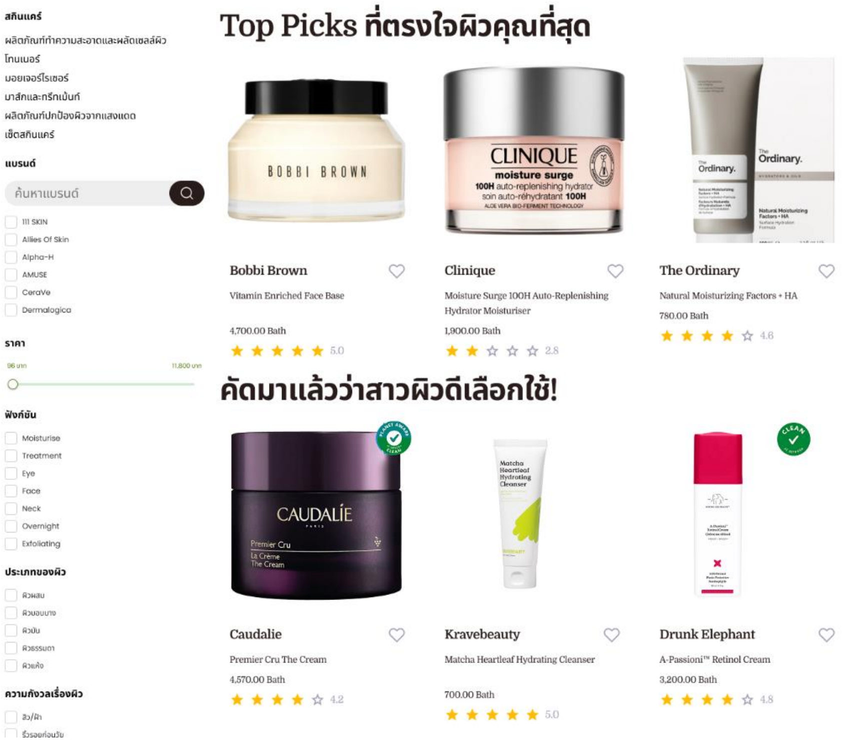
สินค้าแนะนำสำหรับคุณ

Bobbi Brown ♡	Clinique ♡	The Ordinary ♡
Vitamin Enriched Face Base	Moisture Surge 100H Auto-Replenishing Hydrator Moist...	Natural Moisturizing Factors + HA
4,700.00 Bath	1,900.00 Bath	780.00 Bath
★★★★★ 5.0	★★☆☆☆ 2.8	★★★★☆ 4.6

สินค้าที่มีส่วนผสมคล้ายคลึงกัน

Shiseido ♡	Elemis ♡	Kiehl's ♡	Clinique ♡
Benefiance Wrinkle Smoothing Eye Cream	Pro-Collagen Marine Cream	Creamy Eye Treatment with Avocado	Moisture Surge 72-Hour Auto-Replenishing Hydrator
2,900.00 Bath	4,800.00 Bath	1,800.00 Bath	3,200.00 Bath
★★★★★ 5.0	★★★★☆ 4.6	★★★★★ 4.8	★★★★☆ 4.7

Figure 5 System recommending products with similar ingredients when an item is out of stock



สกินแคร์
ผลิตภัณฑ์ทำความสะอาดและผลัดเซลล์ผิว
โทนเนอร์
มอยเจอร์ไรเซอร์
มาส์กและทรีทเม้นท์
ผลิตภัณฑ์ปกป้องผิวจากแสงแดด
เซ็ตสกินแคร์

แบรนด์
ค้นหาแบรนด์
□ 111 SKIN
□ Allies Of Skin
□ Alpha-H
□ AMUSE
□ CeraVe
□ Dermalogica

ราคา
96 บาท 11,800 บาท

ฟังก์ชัน
□ Moisturise
□ Treatment
□ Eye
□ Face
□ Neck
□ Overnight
□ Exfoliating

ประเภทของผิว
□ ผิวผสม
□ ผิวบอบบาง
□ ผิวมัน
□ ผิวธรรมดา
□ ผิวแห้ง

ความกังวลเรื่องผิว
□ สิว/ฝ้า
□ ริ้วรอยก่อนวัย

Top Picks ที่ตรงใจผิวคุณที่สุด

Bobbi Brown ♡	Clinique ♡	The Ordinary ♡
Vitamin Enriched Face Base	Moisture Surge 100H Auto-Replenishing Hydrator Moisturiser	Natural Moisturizing Factors + HA
4,700.00 Bath	1,900.00 Bath	780.00 Bath
★★★★★ 5.0	★★☆☆☆ 2.8	★★★★☆ 4.6

คัดมาแล้วว่าสาวผิวดีเลือกใช้!

Caudalie ♡	Kravebeauty ♡	Drunk Elephant ♡
Premier Cru The Cream	Matcha Heartleaf Hydrating Cleanser	A-Passioni™ Retinol Cream
4,570.00 Bath	700.00 Bath	3,200.00 Bath
★★★★☆ 4.2	★★★★★ 5.0	★★★★☆ 4.8

Figure 6 Product recommendation system suitable for skin type

As shown in Figure 4, users can search for skincare products by ingredient and find alternatives—even when the original product is out of stock. Unlike typical e-commerce platforms that often miss sales in such cases, this system recommends products with similar ingredients. Figure 5 shows how it suggests closely matched alternatives, going beyond traditional systems that only recommend frequently co-purchased items. This is a key advantage of our proposed approach. Additionally, as illustrated in Figure 6, the system can also recommend products based on the user's specific skin type with high accuracy.

Summary of the Study

This study aims to analyze the relationships among ingredients in skincare cosmetic products and to develop an automatic recommendation system tailored for e-commerce platforms. The methodology follows the Knowledge Discovery in Databases (KDD) framework and integrates data mining with machine learning techniques. The research process and outcomes are summarized as follows:

1. Analyzing Ingredient Relationships Using Association Rule Mining

The Apriori Algorithm was applied to a cleaned dataset of 2,441 skincare products sourced from Kaggle to analyze ingredient relationships. Association rules were extracted using a minimum support of 5%, confidence of 70%, and lift greater than 1. Notable patterns included niacinamide → glycerin (Support: 6.8%, Confidence: 92.3%, Lift: 1.29) and caffeine & sodium hyaluronate → butylene glycol (Support: 5.4%, Confidence: 93.7%, Lift: 2.10). These high-confidence rules highlight frequently co-occurring ingredients known for hydration, anti-inflammatory properties, and skin revitalization. The results offer actionable insights that support both effective skincare formulation and personalized product recommendations.

2. Clustering Skincare Products Based on Ingredient Similarity

Products were transformed using One-Hot Encoding and clustered using K-Means with k = 4. The resulting groups—moisturizing, sensitive skin care, acne treatment, and general/natural care—were validated using the following metrics: SSE = 52.67, Percentage Error = 12.8%, and Silhouette Score = 0.61. These clusters not only reflected common consumer categories but also demonstrated a balanced distribution and acceptable statistical separation, reinforcing their suitability for integration into the recommendation system.

3. Developing and Evaluating an Ingredient-Based Recommendation System

A prototype recommendation system was developed using Flask, integrating Content-Based Filtering, Collaborative Filtering, and Case-Based Reasoning (CBR). It also incorporated machine learning models, including K-Means clustering and supervised learning algorithms such as Logistic Regression and Decision Trees. The models were trained and evaluated using 10-fold cross-validation, achieving up to 87.2% accuracy, an average F1-score of approximately 0.876, and an AUC of up to 0.92. The system supports real-time, API-based recommendations and offers key functionalities such as suggesting substitutes for out-of-stock products, recommending items with similar ingredient profiles, and providing personalized suggestions based on user skin types (e.g., oily, sensitive, or acne-prone).

Discussions

This study applied the KDD framework to analyze ingredient-level data and develop a recommendation system for skincare products in e-commerce. The findings are discussed across three key components:

1. Ingredient Association Discovery

Apriori Algorithm revealed significant co-occurrence rules—niacinamide → glycerin and caffeine & sodium hyaluronate → butylene glycol—supporting functional combinations for hydration and skin repair. These rules, with high support, confidence, and lift, reinforce the "functional formulation" trend (Lee et al., 2024) and offer clearer interpretability than similarity-based approaches (Yoon & Joung, 2020).

2. Product Segmentation via K-Means Clustering

Clustering grouped products into four categories: moisturizing, sensitive care, acne treatment, and general care. Evaluation metrics (SSE = 52.67, Silhouette = 0.61, Error = 12.8%) indicate meaningful separation and practical relevance. The results align with Liu and Zhang's (2018) market segmentation strategy and enhance personalization based on ingredient profiles.

3. Predictive Modeling and Real-Time System Deployment

Supervised models (Logistic Regression and Decision Tree) achieved strong performance (accuracy up to 87.2%, F1 = 0.876, AUC = 0.92), confirming their reliability for ingredient compatibility prediction. The system, developed with Flask and real-time API support, enables dynamic updates—reflecting real-time personalization practices in modern e-commerce (Kim & Park, 2020).

In summary, the research integrates rule mining, clustering, and prediction to bridge ingredient-level insights with consumer-focused skincare recommendations—enhancing personalization, system adaptability, and user experience.

Recommendations

1. Apply association insights to formulate skincare products for specific skin types (e.g., oily, sensitive).
2. Integrate the system into e-commerce to suggest substitutes when products are out of stock.
3. Use ingredient trends to drive data-informed marketing and promotions.

Directions for Future Research

1. Include user context (e.g., skin type, preferences, allergies) for better personalization.
2. Leverage user reviews and social media to assess post-usage satisfaction.
3. Explore deep learning (e.g., embeddings, graph models) for advanced ingredient analysis.
4. Develop cross-category recommendations (e.g., matching serums and sunscreens).
5. Compare results with best-selling brand strategies for real-world validation.

Reference

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications*, *7*(1), 39–59.

Choi, H. K., & Lee, J. Y. (2021). Understanding consumer behavior in cosmetic e-commerce using big data. *Journal of Consumer Marketing, 38*(7), 721–730.

Dey, A., & Dubey, S. K. (2023). *Cosmetics science and skin care: History and concepts.* In Nanocosmetics (pp. 1–15). CRC Press.

Eward96. (2021). *Skincare products clean dataset [Data set]*. Kaggle. https://www.kaggle.com/datasets/eward96/skincare-products-clean-dataset/data

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Hasibuan, M. H. (2024). Apriori algorithm to predict availability of beauty products. *Journal of Computer Networks, Architecture and High Performance Computing, 6*(3).

Kantabutra, S., & Couch, A. L. (2000). Parallel K-means clustering algorithm on NOWs. *Technical Journal, 1*(6), 200-248.

Kim, J., & Park, S. (2020). Using machine learning for personalized skincare product recommendation based on user preferences. *Journal of Cosmetic Science, 71*(4), 543–557.

Kingabzpro. (2021). *Cosmetics datasets [Data set]*. Kaggle. https://www.kaggle.com/datasets/kingabzpro/cosmetics-datasets/data

Kouassi, M.-C., Grisel, M., & Gore, E. (2022). Multifunctional active ingredient-based delivery systems for skincare formulations: A review. *Colloids and Surfaces B: Biointerfaces*, 217, 112676. https://doi.org/10.1016/j.colsurfb.2022.112676

Lee, J., et al. (2024). Deep learning-based skin care product recommendation: A focus on cosmetic ingredient analysis and facial skin conditions. *Journal of Cosmetic Dermatology, 23*(6), 2066–2076.

Liu, S., & Zhang, Z. (2018). A hybrid recommendation system for skincare products using collaborative filtering and deep learning. *Journal of Applied Intelligence, 48*(7), 1867–1879.

Roiger, R., & Geatz, M. (2002). *Data mining: A tutorial-based primer.* Addison-Wesley.

Tristandinata, M. (2024). Consumer confusion and its impact on decision making among female consumers in Indonesia's online cosmetics and personal care industry. *JSER, 6*(1). https://doi.org/10.54783/jser.v6i1.540

Yoon, J. Y., & Joung, S. H. (2020). A big data based cosmetic recommendation algorithm. *Journal of System and Management Sciences, 10*(2), 42-50.

Yun, S., & Youn, S. (2017). Recommendation system using big data processing technique. *Journal of the Korea Institute of Information and Communication Engineering, 21*(6), 1183–1190.