# Final Report of Traineeship Program 2025

## On

# “Analysis of Chemical Components Project”

# MEDTOUREASY

## 27th September, 2025

**By- ANUSHK SANGHVI**

# ACKNOWLEDGMENT

The traineeship program that I undertook with **MedTourEasy** has been an invaluable opportunity for me to explore the depth of **Data Analytics**. It not only enhanced my technical knowledge but also contributed significantly to my personal and professional growth. I feel fortunate to have interacted with skilled professionals and mentors who guided me patiently throughout the project, making this experience both insightful and rewarding.

I would like to extend my sincere gratitude and appreciation to the **Training & Development Team at MTE Trainings** for giving me the chance to be a part of their esteemed traineeship program. Their continuous guidance, encouragement, and willingness to share knowledge helped me build a strong foundation in analytics and ensured that I could successfully complete my project with confidence and clarity.

I am equally thankful to my colleagues at MTE Trainings for creating a collaborative and motivating environment that made learning more engaging and effective. Their support and cooperation contributed greatly to the smooth execution of my project and made the overall traineeship a truly enriching journey.

# TABLE OF CONTENTS

# ABSTRACT

Cosmetic products often contain complex chemical formulations that are difficult for consumers to interpret, especially for individuals with sensitive or dry skin. Making the right purchasing decision without understanding ingredient lists can be challenging and, at times, risky. To address this issue, this project focuses on analyzing the chemical components of cosmetics and developing a content-based recommendation system that suggests products based on ingredient similarity.

The dataset used for this project consists of 1,472 cosmetic products from Sephora, including details such as product names, brands, categories, ingredients, prices, and skin type suitability. The project methodology involved several key steps: data preprocessing and filtering (with a focus on moisturizers for dry skin), tokenization of ingredient lists, construction of a cosmetic-ingredient matrix, and dimensionality reduction using t-SNE (t-Distributed Stochastic Neighbor Embedding). To enhance interpretability, an interactive visualization was created using the Bokeh library, enabling users to explore product clusters and compare ingredient compositions.

The results demonstrated that products positioned closely on the t-SNE map shared significant overlaps in their chemical components. For example, AMOREPACIFIC Color Control Cushion SPF 50+ and LANEIGE BB Cushion Hydra Radiance SPF 50 were identified as chemically similar due to shared ingredients such as *Cyclopentasiloxane*, *Dimethicone*, and UV filters.

This project highlights how data science and visualization can transform raw ingredient lists into actionable insights, empowering consumers to make informed choices. The system lays the foundation for building intelligent, ingredient-based cosmetic recommendation tools with future potential for personalization, advanced embeddings, and integration into real-world applications.

# 1. INTRODUCTION

## 1.1. About the company

MedTourEasy is a global healthcare company specializing in providing informational resources and analytical solutions to patients seeking medical services worldwide. Founded in 2015 and headquartered in New Delhi, India, MedTourEasy facilitates access to affordable, high-quality medical treatment globally through an easy-to-use platform. The company offers services including medical second opinions, treatment package customization, and telemedicine consultations. By providing analytical solutions and rich informational resources, MedTourEasy empowers patients to make transparent and data-driven healthcare decisions across a broad spectrum of specialties—from cancer care and cardiology to cosmetic surgery and wellness therapies.

By leveraging data analytics and technology, MedTourEasy improves access to healthcare for patients globally, helping them make informed decisions about their medical journeys. The company also collaborates closely with partner healthcare providers to deliver analytical insights that enhance healthcare delivery.

MedTourEasy's MTE Training Program is a comprehensive and structured initiative designed to equip aspiring data analysts, healthcare professionals, and medical tourism facilitators with practical skills in healthcare data analysis, analytics tools, and decision-making frameworks. The MTE training program has successfully prepared numerous learners to transition into data-driven roles in healthcare and related industries. By fostering both technical expertise and domain understanding, MedTourEasy continues to support workforce development aligned with evolving healthcare technology trends.

## 1.2. About the project

The project titled **"Analysis of Chemical Components"** was carried out as part of the **Data Analytics Traineeship Program at MedTourEasy (MTE Trainings)**. The primary objective of the project was to apply data analytics and machine learning techniques to solve a real-world problem in the cosmetics industry — understanding product similarities based on their chemical ingredients.

Cosmetic products often contain long and complex lists of ingredients, which can be overwhelming for consumers to interpret. This is especially challenging for individuals with sensitive or dry skin, where choosing the wrong product can lead to skin issues. While this information is available on product labels, the lack of clarity makes it difficult for users to make informed purchase decisions.

This project leverages **data science methodologies** to create a **content-based recommendation system** that recommends products based on their chemical composition. Using a dataset of **1,472 cosmetics from Sephora**, the project focuses specifically on the category of **Moisturizers** for **Dry Skin**.

We started by collecting a **cosmetics dataset** containing product details such as **name, brand, price, rating, label (product type), ingredients, and suitability for different skin types**. Since ingredient composition is the most critical factor in product performance, our focus was on **tokenizing and analyzing ingredient lists**.

Using **Natural Language Processing (NLP)** techniques, each product's ingredients were split into tokens (words), standardized, and then represented in a **Document-Term Matrix (DTM)**. This binary matrix mapped whether a particular ingredient was present (1) or absent (0) in a given product.

As the DTM was very high-dimensional (thousands of unique ingredients), we applied **t-SNE (T-distributed Stochastic Neighbor Embedding)** to reduce the data into two dimensions. This dimensionality reduction preserved product similarities while allowing us to visualize them effectively.

Finally, we built an **interactive visualization using Bokeh**, where each point represents a product. Hovering over a product point displays useful details like its **name, brand, price, and rank**, making it easy to explore and compare products.

Through this project, we were able to:

- **Map relationships** between products based on ingredient similarity.
- **Visualize clusters** of products that share common formulations.
- **Provide product comparisons**, e.g., showing which moisturizers for dry skin are most alike in composition.

## 1.3. Objective and Deliverables

### Objectives

- **Ingredient Analysis:** Develop a method to process and analyze the chemical ingredient lists of cosmetic products to provide a clearer understanding for consumers.
- **Product Categorization:** Focus on specific product categories (such as moisturizers) and skin types (such as dry skin) to tailor the analysis.
- **Data Representation:** Tokenize ingredient lists and convert them into a structured, binary document-term matrix representing ingredient presence across products.
- **Similarity Measurement:** Utilize machine learning techniques, specifically t-distributed Stochastic Neighbor Embedding (t-SNE), to reduce high-dimensional ingredient data to an interpretable two-dimensional space while preserving product similarity.
- **Interactive Visualization:** Create an interactive visualization using the Bokeh library to allow users to explore the chemical similarity between products with detailed information via hover tools.
- **Recommendation Support:** Enable consumers to identify alternative products with similar chemical compositions, facilitating informed, safer cosmetic choices without needing specialized chemical knowledge.

### Deliverables

- **Cleaned and Filtered Dataset:** A subset of the original cosmetic products dataset, filtered by product category and skin type.
- **Tokenized Ingredient Corpus:** A list of tokenized ingredient lists suitable for constructing the document-term matrix.
- **Ingredient Index Dictionary:** A mapping from unique ingredient names to indices for matrix construction.
- **Cosmetic-Ingredient Matrix:** A binary representation (document-term matrix) indicating which ingredients are present in each product.

- **t-SNE Embeddings:** Two-dimensional coordinates obtained from dimensionality reduction of the ingredient matrix.
- **Interactive Visualization:** A scatter plot of cosmetics positioned by ingredient similarity, with hover tools showing product metadata.
- **Case Study Analysis:** A detailed comparison of two chemically similar cosmetic products demonstrating the effectiveness of the recommendation system.

# 2. METHODOLOGY

## 2.1. Flow of the Project

The flow of the project "**Analysis of Chemical Components**" is designed to systematically move from raw data to meaningful insights and visualizations. The major steps are as follows:

1. **Data Collection and Importing**
   - The cosmetics dataset was collected and imported into a Jupyter Notebook environment.
   - The dataset contains product details such as product name, brand, label (type of product), price, rating, skin type suitability, and ingredient lists.

2. **Data Cleaning and Filtering**
   - Filtered the dataset to focus specifically on moisturizers for dry skin, ensuring a targeted analysis.
   - Cleaned and prepared the ingredient lists for further processing.

3. **Tokenization of Ingredients**
   - Converted all ingredient text into lowercase for uniformity.
   - Tokenized (split) ingredient lists into individual components to build a structured corpus.
   - Created an ingredient index dictionary for mapping each unique ingredient to a numerical index.
4. **Document-Term Matrix (DTM) Creation**
   - Built a binary matrix representing the presence (1) or absence (0) of ingredients in each product.
   - This Cosmetic-Ingredient Matrix forms the foundation for product comparison.

5. **Dimensionality Reduction with t-SNE**
    - As the DTM contained thousands of features, t-SNE (T-distributed Stochastic Neighbor Embedding) was applied to reduce the data into two dimensions. Reduce the high-dimensional ingredient matrix into two dimensions while preserving local similarity patterns.
    - This preserved product similarities while making the data suitable for visualization.

6. **Visualization with Bokeh**
    - An interactive scatter plot was created using the Bokeh library.
    - Each point represents a product, positioned based on ingredient similarity.
    - A hover tool was added to display details like product name, brand, price, and rank.

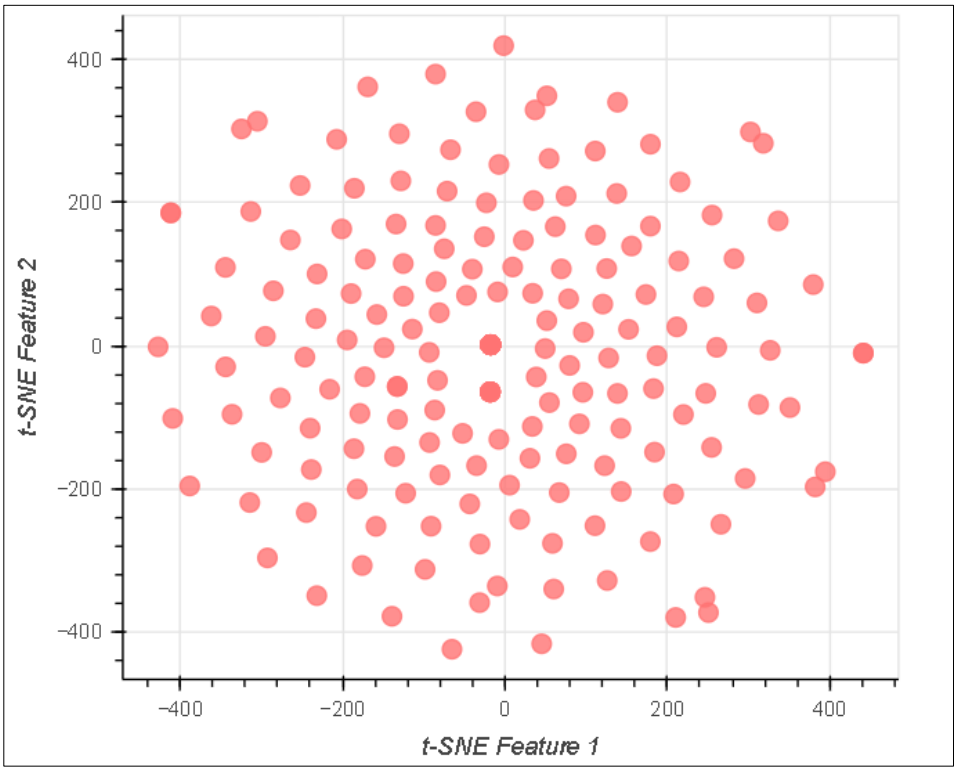


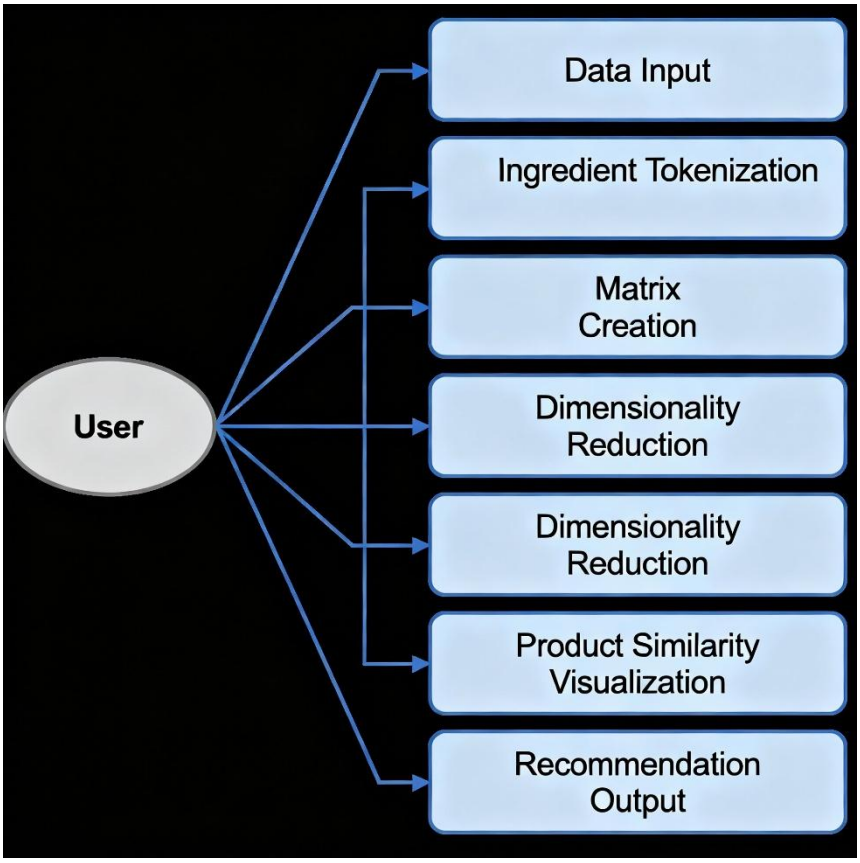

**Figure: t-SNE Scatter Plot of Cosmetic Products**

7. **Analysis and Product Comparison**
    - Products with similar ingredient compositions were identified.

- Example comparisons between two products demonstrated how ingredient similarity translates into potential alternatives.

8. **Conclusion and Insights**
   - Utilize visualization and data to guide consumers on making informed cosmetic choices based on ingredient similarity, price, and ratings.

## 2.2. Use Case Diagram




## 2.3. Language and Platform Used

The project was implemented using the following programming languages, libraries, and platforms:

- **Python –** The primary programming language used for data preprocessing, analysis, and model building.
- **Pandas & NumPy –** For handling, cleaning, and preprocessing the dataset efficiently.

- **Scikit-learn (sklearn) –** Used for dimensionality reduction with t-SNE.
- **Bokeh –** A powerful Python visualization library, used for creating interactive plots and scatter visualizations of products.
- **Jupyter Notebook –** The platform used for coding, visualization, and documenting the workflow in an interactive manner.
- **Dataset Source –** A CSV dataset (cosmetics.csv) containing product details and ingredient lists.

# 3. IMPLEMENTATION

## 3.1. Gathering Requirements and Defining Problem Statement

### Problem Statement

Selecting the right cosmetic product is often a challenging task for consumers, especially for individuals with sensitive or specific skin types such as dry skin. While the ingredient lists provided on product packaging contain the necessary information, they are often complex and difficult to interpret without specialized knowledge of chemistry or dermatology. As a result, consumers may struggle to identify which products are suitable for their needs and may end up purchasing products that cause skin irritation or fail to deliver the expected results.

To address this challenge, there is a need for a system that can analyze cosmetic ingredient lists, compare products, and provide insights into their similarities. Such a system would enable users to make informed decisions and reduce the risk of adverse reactions.

### Requirements

#### i. Functional Requirements

- The system should allow the analysis of cosmetic product ingredient lists.
- It should filter products based on category (e.g., moisturizer, cleanser) and skin type (e.g., dry, oily, sensitive).
- It must preprocess the data by tokenizing ingredient lists into a structured format.
- It should generate a Cosmetic-Ingredient matrix that maps products to their chemical components.

- It should apply dimensionality reduction to visualize similarities between products in a two-dimensional space.
- The visualization should be interactive, allowing users to hover over products to view details like name, brand, price, and rank.

**ii. Non-Functional Requirements**

- The system should be implemented in Python for flexibility and ease of data handling.
- Libraries like Pandas, NumPy, Scikit-learn, and Bokeh must be used for preprocessing, machine learning, and visualization.
- The system should be efficient enough to process and visualize over 1000+ cosmetic products.
- The interface (Jupyter Notebook) should be interactive, user-friendly, and easily modifiable for other categories or skin types.

## 3.2. Data collection and Importing

The first step in any data-driven project is to **collect and load the dataset**. For this project, the dataset cosmetics.csv is used. It contains details such as product name, brand, price, rank, label (type of product), and a list of ingredients.

We use **Pandas**, a Python library for data manipulation and analysis, to import and inspect the dataset. This allows us to understand its structure before performing any transformations.

Import and inspect the dataset.

i. Import pandas aliased as pd and numpy as np. Import TSNE from sklearn.manifold.
ii. Read the CSV file, "datasets/cosmetics.csv", into a pandas DataFrame and name it df.
iii. Display a sample of five rows of the data using the sample() method inside the display() function.
iv. Display counts of types of products using the value_counts() method on the Label column of df.

## 3.3. Data cleaning and preprocessing

Before analysis, unnecessary missing values and irrelevant attributes were handled. Products were filtered to include only moisturizers for dry skin. The ingredient lists were split into tokens for further processing.

Filter the data for moisturizers and dry skin.

i. Filter df for "Moisturizer" in the Label column and store the result in moisturizers.
ii. Filter moisturizers for 1 in the Dry column and store the result in moisturizers_dry.
iii. Drop the current index of moisturizers_dry and replace it with a new one usingthe reset_index() method, setting drop = True.

## 3.4. Tokenization of ingredients

**Tokenization** refers to breaking text into smaller parts (tokens). Here, each ingredient string is split into individual ingredients.

Tokenize the ingredients and create a bag of words.

i. Inside the outer for loop:
   - Make each product's ingredients list lowercase.
   - Split the lowercase text into tokens by specifying ', ' as the separator.
   - Append tokens (which itself is a list) to the list corpus.

ii. Inside the inner for loop, if the ingredient is not yet in ingredient_idx dictionary:
   - Add an entry to ingredient_idx with the key being the new ingredient and thevalue being the current idx value.
   - Increment idx by 1.

## 3.5. Document-Term Matrix (DTM)

A **Document-Term Matrix (DTM)** is used to represent the relationship between products and ingredients:

- Each row = a cosmetic product (document).
- Each column = an ingredient (term).
- Each value = 1 if the ingredient is present in the product, else 0.

This matrix captures the ingredient composition of each cosmetic in a structured, binary form.

i. Initialize the document-term matrix.
   - Get the total number of products in the moisturizers_dry DataFrame. Assign it to M.

- Get the total number of ingredients in the ingredient_idx dictionary. Assign it to N.
- Create a matrix of zeros with size MxN. Assign it to A.

ii. One-hot encoding function
- Get the binary value of the tokens for each row of the matrix A.
- Inside the for loop:
- Apply oh_encoder() to get a one-hot encoded matrix for each list of tokens in corpus (i.e., each product's ingredients list).
- Increment i by 1.

iii. Fill the matrix.

## 3.6. Dimension Reduction with t-SNE

The DTM created has very high dimensionality (e.g., 190 products × 2233 ingredients). Visualizing this in raw form is not possible.

To solve this, we use **t-SNE (T-distributed Stochastic Neighbor Embedding)**:

- It reduces high-dimensional data into **2D coordinates**.
- Maintains **similarity structure** between products.
- Products with similar ingredients will appear **closer** in the 2D plot.

This enables clear visualization of product similarity.

Reduce the dimensions of the matrix using t-SNE.

i. Create a TSNE instance with n_components = 2, learning_rate = 200, and random_state = 42. Assign it to model.
ii. Apply the fit_transform() method of model to the matrix A. Assign the result to tsne_features.
iii. Assign the first column of tsne_features to moisturizers_dry['X'].
iv. Assign the second column of tsne_features to moisturizers_dry['Y']

## 3.7. Visualization with Bokeh

The reduced t-SNE coordinates are plotted using **Bokeh**, an interactive visualization library.

- Each cosmetic product is represented as a **circle** on the scatter plot.
- The **HoverTool** feature is added, so when a user moves the cursor over a product, they can see details like Name, Brand, Price, and Rank.

- This helps in exploring similarities visually and interactively.

### 3.8 Interactive Exploration and Insights

- The interactive Bokeh plot allows users to explore clusters of similar products visually.
- Users can identify groups of closely related products and investigate ingredient overlap to discover safer or more affordable alternatives.
- This visual exploration is instrumental for consumers lacking detailed chemical knowledge to make sound decisions using ingredient similarity maps.
- The plotted results serve as a foundation for developing more advanced recommendation engines and personalized skincare suggestions.

# 4. PROJECT IMPLEMENTATION CODE

The complete implementation of this project, including data preprocessing, Cosmetic-Ingredient matrix creation, t-SNE analysis, and interactive visualization using Bokeh, is available in the Jupyter Notebook.

**Jupyter Notebook Link:**
https://github.com/anushksanghvi/MedTourEasy/blob/main/notebook.ipynb

**Complete Project Link:**
https://github.com/anushksanghvi/MedTourEasy.git

# 5. RESULT AND OBSERVATIONS

### Result

The project successfully processed and analyzed cosmetic ingredient data, focusing on moisturizers for dry skin.

- The t-SNE plot successfully visualizes the similarity between cosmetic products based on their ingredients.
- Products that are close together on the plot have similar chemical compositions.
- Example Pair:

- **AMOREPACIFIC Color Control Cushion SPF 50+**
- **LANEIGE BB Cushion Hydra Radiance SPF 50**

These two products share many core ingredients like Cyclopentasiloxane, Dimethicone, and UV filters.

- This analysis helps users find alternative products with similar formulations and make informed purchase decisions.

### Observations

- The ingredient-based similarity measure clusters products with overlapping chemical profiles effectively.
- Proximity in the t-SNE plot closely matches ingredient list similarity and product characteristics.
- Hover tools improve usability for consumers without specialized chemical knowledge.
- This system supports personalized cosmetic recommendations, especially useful for sensitive skins.
- Limitations include missing ingredient concentration and allergen data, which could refine recommendations.

# 6. CONCLUSION

This project demonstrates the successful application of data science methods to analyze and visualize complex cosmetic ingredient data for personalized skincare recommendations. By structuring ingredient information and applying dimensionality reduction and interactive visualization techniques, the system translates chemical complexity into user-friendly insights. It empowers consumers to make safer and better-informed cosmetic choices. Future enhancements may include integration of ingredient toxicity, allergen warnings, and user preferences to build a comprehensive decision-support platform in cosmetics.

# 7. FUTURE SCOPE

- Incorporate ingredient concentration levels to refine similarity measures and recommendations.

- Expand analysis to cover multiple cosmetic product categories beyond moisturizers.
- Develop a user-friendly mobile or web application interface for broader accessibility.
- Implement machine learning-based personalized recommendation engines leveraging user preferences and purchase history.
- Include sentiment analysis from user reviews to complement ingredient-based similarity.
- Utilize automated ingredient synonym resolution to improve tokenization accuracy.
- Enable real-time dataset updates from cosmetic brands and retailers for current recommendations.

# 8. REFERENCE


[1] Serb, A. F. (2024). Mass-spectrometry-based research of cosmetic ingredients. *Journal of Cosmetic Analysis*. https://pubmed.ncbi.nlm.nih.gov/38542972/

[2] Klaschka, U. (2016). Natural personal care products—analysis of ingredient lists. *International Journal of Environmental Research*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5044959/

[3] Lee, J. (2024). Deep learning-based skin care product recommendation. *Journal of Cosmetic Dermatology*.

[4] Maximizer Market Research. (2024). Global cosmetic ingredient market report. https://www.maximizemarketresearch.com/

[5] Gong, L., et al. (2023). CCIBP: A comprehensive cosmetic ingredients bioinformatics platform. *Bioinformatics Journal*. https://pmc.ncbi.nlm.nih.gov/articles/PMC10345691/

[6] Sephora Cosmetic's Ingredient Analysis. (2025). Kaggle. https://www.kaggle.com/code/willianleite/sephora-cosmetic-s-ingredient-analysis

[7] Content-Based Recommendation Cosmetics. (2024). Kaggle. https://www.kaggle.com/code/huseyincosgun/content-based-recommendation-cosmetics

[8] Shang, Y. (2023). Applications of mass spectrometry in cosmetic analysis. *Analytical Chemistry*. https://www.sciencedirect.com/science/article/abs/pii/S0021967323004016

[9] Precedence Research. (2024). Natural cosmetic ingredient market size report. https://www.precedenceresearch.com/natural-cosmetic-ingredient-market