Pune Institute of Computer Engineering, Dhankawadi, Pune

A SEMINAR REPORT ON

"Application of Data Science to aid formulation of Skincare Products"

SUBMITTED BY

Name: Farkhanda Dalal
Roll No: 31418
Name: Mansi Jangle
Roll No: 31438
Name: Misbah Bagwan
Roll No: 31455

Under the guidance of
Prof. Rutuja Kulkarni

DEPARTMENT OF COMPUTER ENGINEERING

Academic Year 2024-25

DEPARTMENT OF COMPUTER ENGINEERING

Pune Institute of Computer Technology
Dhankawadi, Pune-43

CERTIFICATE

This is to certify that the Seminar report entitled
"Application of Data Science to aid formulation of Skincare Products"

Submitted by

Farkhanda Dalal	Roll No: 31418
Mansi Jangle	Roll No: 31438
Misbah Bagwan	Roll No: 31455

have satisfactorily completed a seminar report under the guidance of Prof. Rutuja Kulkarni towards the partial fulfillment of third year Computer Engineering Semester I, Academic Year 2024-25 of Savitribai Phule Pune University

Prof. Rutuja Kulkarni
Internal Guide
Department of Computer Technology

G. V. Kale
Head of Department
Department of Computer Technology

Place: Pune
Date: 21/10/2024

ACKNOWLEDGEMENT

It was indeed a great learning for me to write this report on “Application of Data Science to aid formulation of Skincare Products”. We would like to thank our Seminar Coordinator Prof N. Y. Kapadnis, Head of Department Dr. G. V. Kale and Principal Dr. S. T. Gandhe for their encouragement and support.

We would also like to thank my guide Prof. Rutuja Kulkarni, Department of Computer Engineering for her guidance and help. She has continuously motivated me throughout the seminar project and thus was of great help to make it successfully.

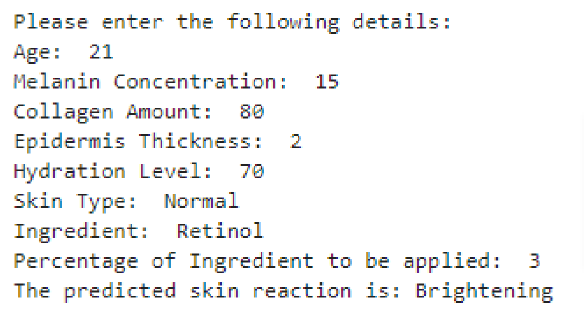
CONTENTS:

Abstract

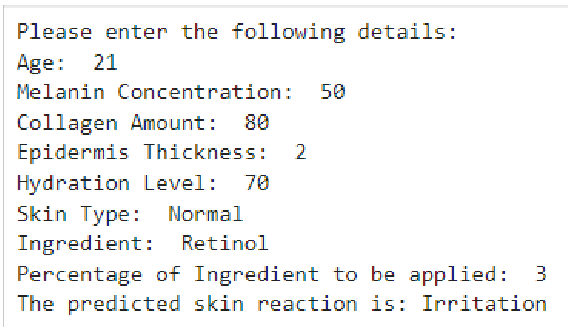
Skincare formulations have predominantly been based on research and data derived from Caucasian skin, leading to a lack of comprehensive studies on skin color. This oversight has resulted in limited product efficacy and relevance for individuals with diverse skin tones and conditions. To create better skincare products, developing formulations that are tailored to the skin color, skin type, age, and skin condition is essential.

Three machine learning algorithms, namely Random Forest Classifier, Naïve Bayes and Support Vector Machine, are trained on skincare data to model and predict reactions to variations of ingredient concentrations to find out which ingredients and percentage is safe for use. Users are prompted to input key skin parameters and ingredient percentages, after which the model predicts likely skin reactions. This approach allows skincare companies and individuals to make data-driven decisions about ingredient suitability for specific skin types, reducing the risk of adverse effects and enhancing overall skincare outcomes. It helps skincare formulators to better customize products for diverse skin types while minimizing adverse reactions.

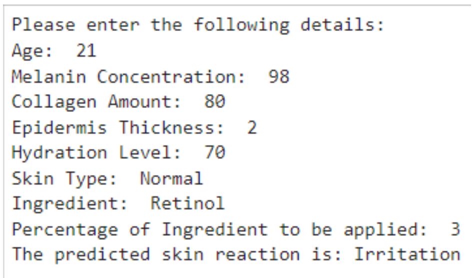
While implementing and comparing the three models. It was verified that Random Forest Classifier displayed highest accuracy and precision (of 99%). The following are the results obtained if 3% retinol is applied by test subjects with same age, epidermal thickness, collagen and hydration levels but with varying melanin concentrations.



```
Please enter the following details:
Age:  21
Melanin Concentration:  15
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Brightening
```



```
Please enter the following details:
Age:  21
Melanin Concentration:  50
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Irritation
```



```
Please enter the following details:
Age:  21
Melanin Concentration:  98
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Irritation
```

Keywords

Data Processing, Data Visualization, Random Forest Classifier, Support Vector Machine, Naïve Bayes, Exploratory Data Analysis, One-hot Encoding, Predictive Modelling, Skincare Dataset, Skincare Formulation using Machine Learning

1 INTRODUCTION

In recent years, the skincare industry has faced significant challenges in addressing the diverse needs of consumers with different skin types. Historically, the formulation of skincare products has been guided by limited datasets, predominantly focusing on Caucasian skin types [9]. As a result, many skincare solutions are ineffective for individuals with different ethnic backgrounds, skin tones, and conditions [10]. This lack of diversity in data collection and analysis has led to the development of one-size-fits-all products that fail to cater to the specific needs of underrepresented demographic groups [8]. Additionally, traditional skincare research has often relied on small sample sizes, subjective assessments, and outdated methodologies, further limiting the accuracy and inclusivity of product formulations. The pressing issue is the skincare industry's inability to address the full spectrum of skin types, which has left a gap in efficacy and consumer satisfaction, particularly for individuals with non-Caucasian skin [10][8].

With the growing awareness of this diversity, there is an urgent need for a more comprehensive, data-driven approach that uses large and diverse datasets. By collecting and analyzing data on multiple skin types, including factors like age, melanin content, collagen, hydration, and epidermal thickness, the skincare industry can better understand the needs of all consumers [5][4]. This shift toward a data-centric methodology can lead to the creation of more inclusive and effective products that cater to a wider audience, thereby reducing the disparities in skincare outcomes [1].

To address this issue, our research applies advanced data science techniques to analyze and interpret complex datasets related to various skin characteristics [1]. By leveraging various methods, such as machine learning and predictive analytics, we can uncover hidden patterns in skin behavior across diverse demographics. Our approach involves collecting a vast range of data points, including skin type (dry, oily, combination, normal, sensitive), pigmentation, sensitivity, and reactions to different ingredients, allowing for the creation of personalized skincare formulations [5].

Our goal is to bridge the gap in the current skincare industry by providing a scientifically validated, data-driven solutions and predictions that can be used to develop targeted skincare products [5]. This will ultimately enhance consumer satisfaction by providing effective skincare solutions that are tailored not just to one demographics but to the full spectrum of skin types globally [10][8].

2 MOTIVATION

Developing effective skincare products for diverse skin types has always been a challenge because of variations in skin composition, tone, and condition. Historically, skincare research has focused on Caucasian skin, resulting in a gap in product efficacy for individuals with darker skin tones or other skin concerns. The motivation for this research lies in addressing these gaps by leveraging data science techniques and machine learning to better understand and cater to the needs of all skin types, especially for those with different ethnicity and skin color.

With advancements in data collection, processing, and machine learning, it is now possible to analyse skin-related data in ways that were previously inaccessible. By combining data science with dermatological research, we can enhance product formulation processes, leading to better outcomes for consumers across all demographics.

3 LITERATURE SURVEY

The Following table shows the literature survey by comparing techniques propose in various references:

Table 1: Literature survey

No.	Paper	Summary	Limitations
1	Personalized skincare: from molecular basis to clinical and commercial applications	Differences in skin types (Asian, African, Caucasian, Oriental) include variations in melanin levels, dermal thickness, and hydration.	Each skin type, faces unique challenges, such as pigmentation issues, TEWL, and sensitivity to photodamage, requiring targeted skincare solutions.
2	Skin care assessment on the basis of skin hydration, melanin, erythema, and sebum at various body sites.	Skin concerns vary by skin tone, as some are had sebum or melanocytes or collagen, etc.	Skincare products must be tailored to both skin tone and body region, as hydration, oil production, and sun protection needs differ.
3	How artificial intelligence adopts human biases: The case of cosmetic skincare industry	AI/ML in skincare often inherits human biases due to lack of diverse data. Biases include sampling, confounding, measurement, and label bias.	Algorithms trained on non-diverse datasets can fail to generalize across skin tones, leading to inaccurate predictions for hyperpigmentation, age, and other skin concerns, especially for darker tones.

4 A SURVEY ON PAPERS

The following points were explored during the literature survey of different papers:

4.1 Structural and physiological differences in skin characteristics based on ethnic backgrounds:

- **Asian Skin:**

  1. **Defining Features:** Greater melanin production, enhanced photoprotection. Thicker dermis and more collagen [5].

  2. **Concerns:** Despite these protective features, hyperpigmentation can still occur [5]. It also faces problem of increased melanin production when subject to irritation [9].

- **African Skin:**

  1. **Composition:** High melanin levels, superior photoprotection. Its stratum corneum is thicker, and it experiences increased TEWL [5][8].

  2. **Structure:** The dermis is thicker, with more fibroblasts and compacted collagen, contributing to better skin elasticity [8].

  3. **Challenges:** This skin type frequently encounters issues such as Xerosis (dry skin) and pigmentation irregularities [5].

- **Caucasian Skin:**

  1. **Characteristics:** Features smaller melanosomes, lower melanin levels, and a thinner stratum corneum, resulting in decreased trans epidermal water loss (TEWL) [5][4].

  2. **Age-Related Changes:** With aging, there is a noticeable thinning of the dermis and decreased elasticity. Additionally, photodamage from UV exposure leads to a reduction in collagen [4][1].

- **Oriental Skin:**

  1. **Core Properties:** This skin type maintains a thinner stratum corneum and reduced TEWL, coupled with densely packed eccrine glands [5].

  2. **Resilience:** It preserves good dermal elasticity, although individuals may experience issues with hyperpigmentation or dyschromia (uneven skin color) [5].

4.2 Variation in skin concerns according to difference in skin tone:

Skincare products need to be tailored to individual skin types and body regions, as each has distinct needs. Lighter skin tones are more vulnerable to sun damage, including redness and tanning, and therefore require formulations with enhanced sun protection [7]. On the other hand, darker skin tones benefit from products focused on moisture retention, often containing humectants to prevent dryness [8]. Additionally, areas like the cheeks and forehead exhibit varying levels of hydration and oil production, requiring targeted care to maintain overall skin health [3]. These differences highlight the importance of developing precise skincare solutions that address both the skin type and specific body area [3].

The following trends were observed while exploring skin parameters such as melanin, erythema, hydration, and sebum across different skin tones and body sites in an Indian population:

1. **Melanin Content**: The concentration of melanin is greater in facial areas, particularly the chin and forehead, compared to the volar forearm. Individuals with darker skin tones (Group III) had higher melanin levels than those with lighter skin (Group I) [3].

2. **Erythema**: Erythema was observed to be most prominent in facial regions across all skin tones. Similar to melanin, erythema levels were elevated in areas with increased exposure to UV radiation [3].

3. **Sebum Levels**: The forehead exhibited the highest levels of sebum, while the volar forearm had the lowest. Individuals in Group I (lighter skin) demonstrated the greatest overall sebum production [3].

4. **Skin Hydration**: The area near the ear (periauricular) had the highest hydration levels, whereas the cheek was the driest. Lighter-skinned individuals exhibited higher hydration compared to those with medium and darker [3].

4.3 Under representation of skin of color in AI and ML:

As Artificial Intelligence and Machine Learning are being integrated in production chains they also inherit and adopt human biases and the same is also seen when AI is applied in the field of cosmetics and skin care [1][9].

1. **Sampling Bias:**

   Sampling bias arises when the input data lacks diversity or is skewed. In cosmetic skincare research, panelists with light skin tones are often overrepresented, while those with darker skin tones are underrepresented. This imbalance in datasets can lead to algorithms that are non-generalizable across race, ethnicity, sex, gender, or age [1].

2. **Confounding Bias:**

   Confounding bias occurs when the relationship between variables is obscured by the presence of additional variables not considered in the model. An example related to skin of color is when predicting hyperpigmentation using UV exposure data but failing to account for melanin levels. Melanin significantly influences how skin reacts to UV light and excluding it from the model can lead to inaccurate predictions, particularly for individuals with darker skin tones [1].

3. **Measurement Bias:**

   Measurement bias occurs due to discrepancies in how skin properties are measured across different devices and methods. For instance, skin color measurements from spectrometers might vary depending on ethnicity and the device used, leading to inconsistencies. Additionally, post-processing features in cameras can introduce biases, particularly in representing different skin tones [1].

4. **Label Bias:**

   Label bias arises when numerical values obtained for skin properties do not account for individual differences. For instance, redness values may vary based on pigmentation, leading to inaccurate classifications. Inconsistent human visual assessments for features like age prediction can also introduce bias, as cultural and individual beliefs may affect judgment [1].

5 PROBLEM DEFINITION AND SCOPE

5.1 Problem Definition:

The primary challenge in developing skincare formulations lies in the limited scope of existing studies, which have predominantly focused on Caucasian skin types. This has led to gaps in product efficacy for individuals with diverse skin tones, textures, and conditions. The current goal is to use data science to bridge this gap by leveraging data analysis, classification using Machine Algorithms like Random Forest Classifier, Naïve-Bayes and Support Vector Machine to predict reactions that are likely to occur when various ingredients in different concentrations are applied on different skin type and varying melanin concentrations. Thus, resulting in better understanding of formulations that are inclusive and effective for several skin types.

5.2 Scope:

1. Employ algorithms to identify trends and patterns in skin data.

2. Predicting reaction of ingredients on different skin types using data-driven insights.

3. Finding out how reactions change with change in parameters like age, melanin concentration, epidermal thickness, etc.

4. Finding out irritants for diverse skin types, to prevent unsuitable formulations.

6 METHODOLOGY

6.1 Data Collection:

Data was collected from diverse sources, including dermatological studies, clinical trials, and skincare product databases. The dataset consists of attributes such as:

1000 records of different individuals with diverse skin types are noted across 13 parameters, namely:

1. **Age**: The age of participants.
2. **Melanin Concentration**: Measured on a scale from 0 to 100.
3. **Collagen Amount**: Quantified in arbitrary units.
4. **Epidermis Thickness**: Measured in millimetres.
5. **Hydration Level**: Assessed using standardized techniques in percentage.
6. **Skin Type**: Categorized as Dry, Normal, Oily, Combination, or Sensitive.
7. **Ingredient**: Type of ingredient applied (e.g., Brightening Ingredients like: Vitamin C, Kojic Acid, Exfoliating Ingredients like Glycolic acid, Azelaic acid and Salicylic acid, Hydrating and Barrier Strengthening Ingredients like Hyaluronic and Niacinamide).
8. **Concentration**: Percentage concentration of the ingredient.
9. **Reactions**: Assessed across multiple dimensions, including Neutral, Positive, Brightening, Irritation, and Negative reactions. Binary 0 represents no reaction while Binary 1 represents a reaction was observed

6.2 Data Processing:

Prior to analysis, the dataset underwent several preprocessing steps to ensure its integrity and reliability:

- **Data Cleaning**: Missing values were identified and handled appropriately. For continuous variables like Melanin Concentration and Hydration Level, missing values were imputed using mean values based on skin type, while categorical variables were encoded using one-hot encoder for ease of analysis.
- **Outlier Detection**: Statistical methods (e.g. boxplot) were employed to detect and remove outliers that could skew the analysis.

6.3 Data Exploration:

- **EDA:**

Exploratory Data Analysis (EDA) was conducted using **Seaborn** and **Matplotlib** for visualizing patterns and relationships within the dataset.

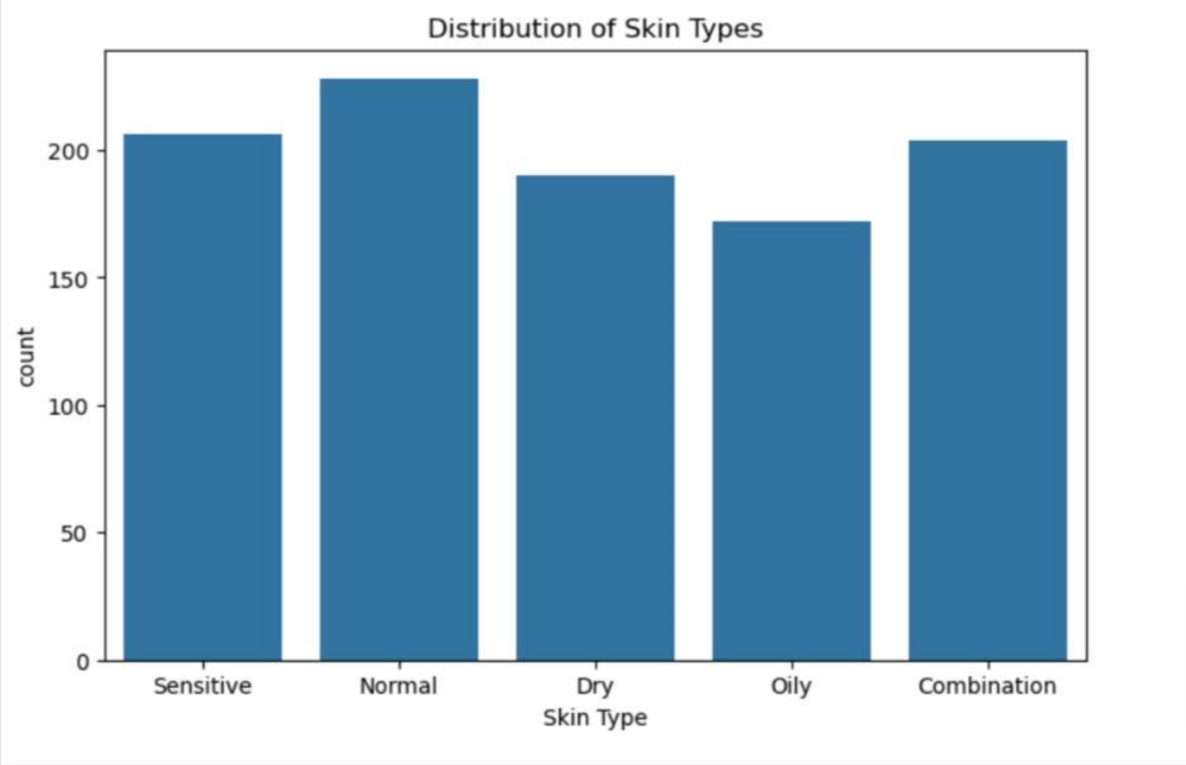
1. **Initial Data Inspection**: The dataset includes 1000 rows and 13 columns, with both numerical and categorical variables. Key columns include age, melanin concentration, collagen amount, epidermis thickness, hydration level, skin type, ingredient, and various skin reactions.

2. **Handling Missing Values**: Missing values in 'Collagen Amount' and 'Epidermis Thickness' were filled using the mean of their respective columns.

3. **Summary Statistics**: Summary statistics provided insights into numerical data, such as an average age of 39 and a melanin concentration range of 1 to 100.

- **Visualization Techniques:**

1. **Distribution of Skin Types:**

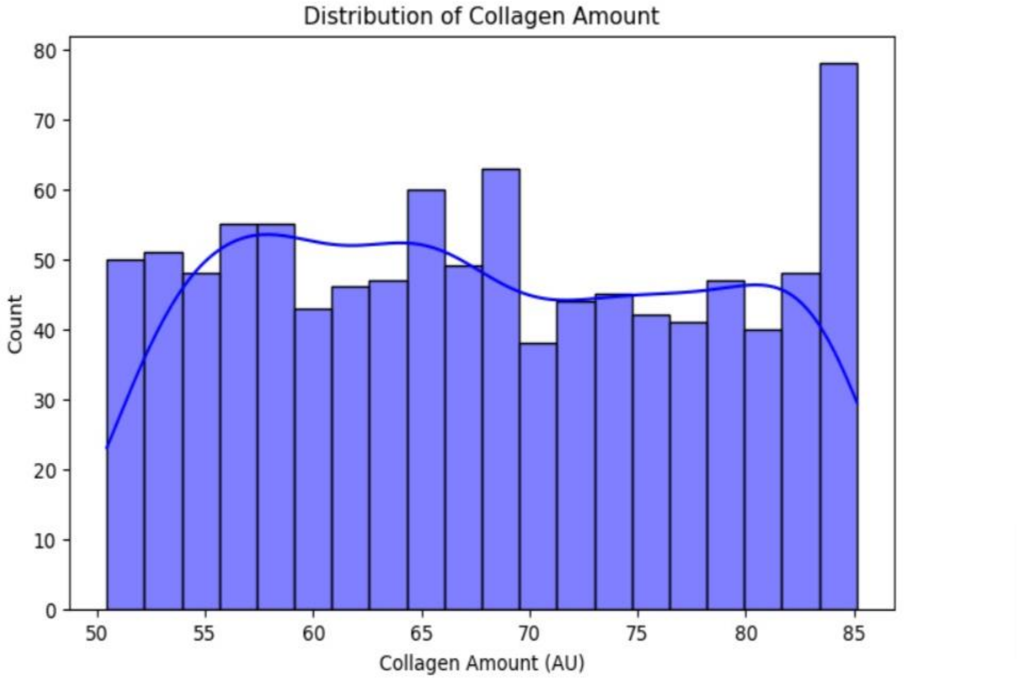


A diverse amount of skin types is present in the dataset, with Normal skin type having the most amount of samples present.

The following frequency was noted for samples in the dataset:

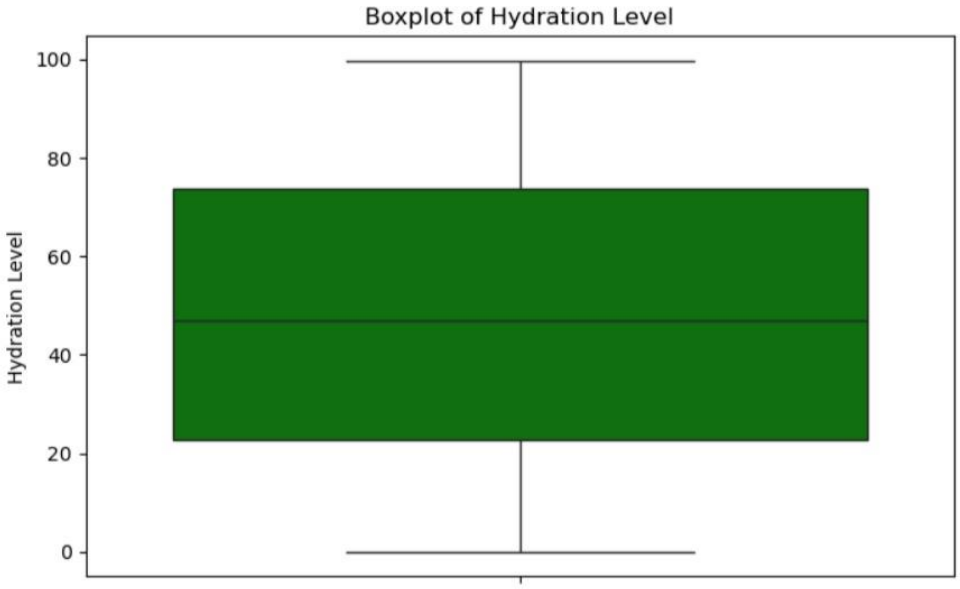
1. Normal Skin: 228
2. Sensitive Skin: 206
3. Combination Skin: 204
4. Dry Skin: 190
5. Oily Skin: 172

2. Distribution of Collagen Amount:



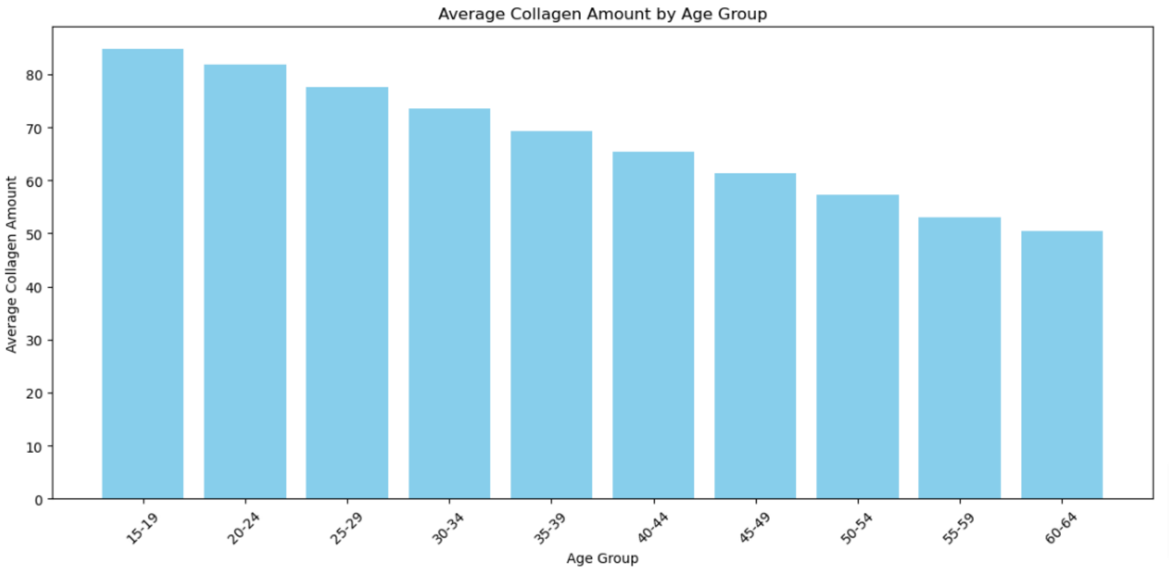
a. Noticeable peak can be observed between the range 65-70, this shows that a significant number of samples have 65-70 AU of Collagen.

3. Boxplot of Hydration Level:



The hydration levels range from 0 to 100, with the box indicating that the majority of the data falls between approximately 40 and 60. The median is positioned around 50, suggesting that half of the observations are below this value and half are above.

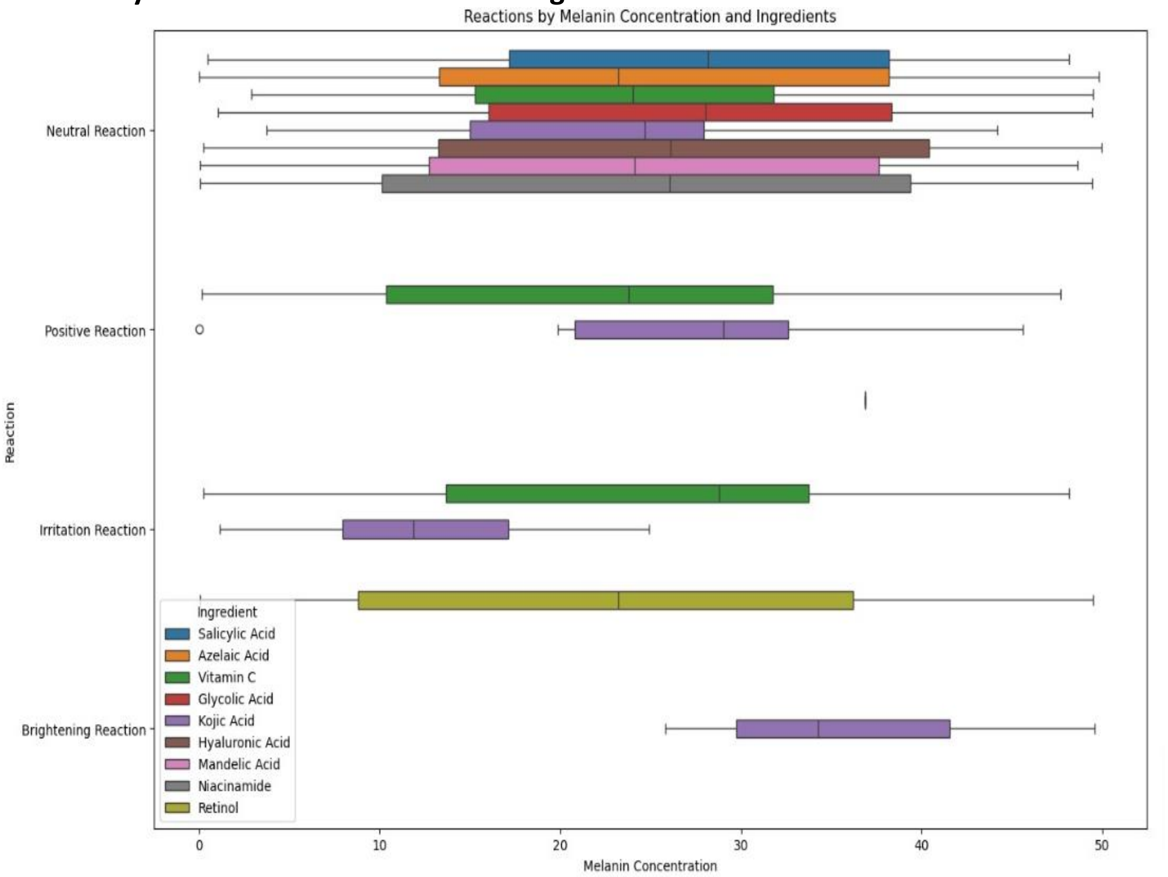
4. Average Collagen Amount by Age Group:



Collagen Levels: The average collagen amount is highest in the younger age groups (15-19 and 20-24), with values around 80-90 units. As age increases, there is a noticeable decline in collagen levels.

Trends: The data shows a gradual decrease in collagen from the age group of 25-29 onward, with the lowest levels observed in the 60-64 age group, which suggests that collagen production diminishes with age.

5. Reactions by Melanin Concentration and Ingredients:



This graph visually represents the reactions of various skincare ingredients at different melanin concentrations.

1. Neutral Reaction:

   Ingredients such as Salicylic Acid, Azelaic Acid, and Niacinamide show a range of neutral reactions across varying melanin concentrations, indicating minimal adverse effects.
2. Positive Reaction:
   Vitamin C and Glycolic Acid exhibit positive reactions, particularly at lower melanin concentrations, suggesting their effectiveness in improving skin condition.
3. Irritation Reaction:
   Kojic Acid show irritation reactions, especially at higher melanin concentrations, indicating potential adverse effects on the skin.
4. Brightening Reaction:
   Mandelic Acid and Retinol demonstrate varying effectiveness in brightening the skin, with their impact depending on the melanin levels.

7 IMPLEMENTATION

Predictive Analysis is performed to project the reactions that will occur when an ingredient with specific concentration will be applied on skin of different ethnicities i.e. with varying melanin concentration. The following models were used to find these results:

7.1 Random Forest Classifier:

1. **Input Data Representation:**
   Input data **consists of user-specific features such as age, melanin concentration, collagen amount, hydration, skin type and formulation-specific features like ingredient composition and concentration**. To prepare this data for modelling, categorical variables (e.g., skin type) are encoded, while numerical variables (e.g., ingredient concentration) are normalized.

2. **Model Training with Random Forest:**

a) Bagging OR Bootstrap Aggregation:
It involves **creating multiple decision trees**, each trained on a random subset of our input data. This is done by sampling the dataset with replacement, which means each tree sees a slightly different subset of the data. This is done by sampling the dataset with replacement, which means each tree sees a slightly different subset of the data. As a result, the model gains diversity and robustness, ensuring that the final predictions are less likely to be influenced by any specific patterns or noise present in the training data.

Example:
Each decision tree is trained on slightly different user-product combinations.
One tree might be trained on data from users with sensitive skin and Vitamin C at 15%, while another may see data with normal skin and Hyaluronic Acid at 2%.

Bagging helps reduce overfitting, ensuring that predictions generalize well to unseen data.

b) Random Feature Selection:
During the construction of each decision tree, Random Forest randomly **selects a subset of input features to split at each node.** This ensures that no single tree can overly rely on one feature. By introducing this randomness, the model captures a wider range of patterns and relationships within the data, leading to more balanced and generalized predictions without relying too heavily on any one factor like age or melanin concentration.

Example:
In one tree, the model might first split based on ingredient concentration, whereas another tree might prioritize skin type first.

By doing so, Random Forest ensures that multiple relationships between features (e.g., age, skin type, ingredient) are considered across all trees, which **reduces the bias of any one feature dominating the prediction process.**

c) Recursive Splitting:
Within each tree, Recursive Splitting is used to **divide the input data** into smaller, more homogeneous groups. Each tree splits the data at decision points (called nodes) based on the values of the features.

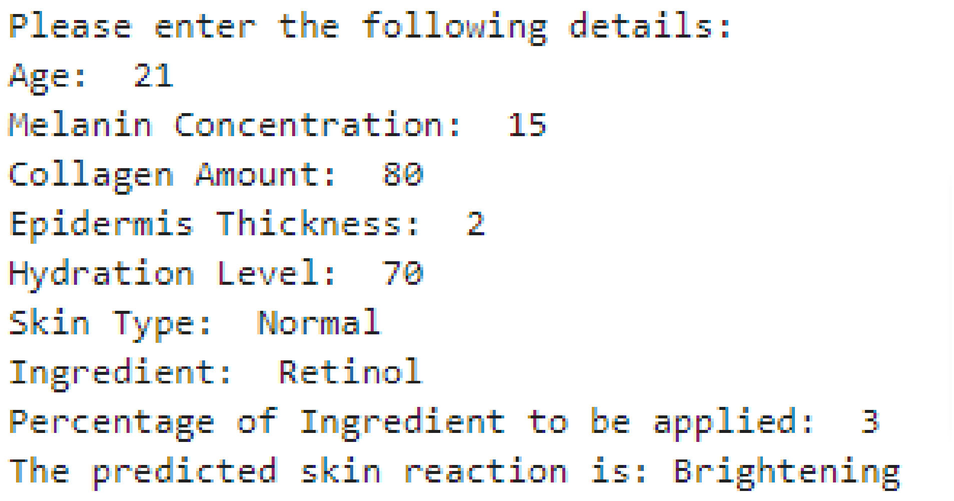
Example:
One tree might first split based on whether **retinol concentration** is higher than 1%. It will then further split those users by age, say, if they're above 30 years old. The tree keeps making splits until the groups are small enough (or a stopping condition is met). Each split is based on thresholds (e.g., "Is age above 30?", "Is melanin concentration greater than 25?"), and the final leaf node will assign a label (e.g., positive reaction, irritation, etc.).

3. **Making Predictions:**
Once trained, the Random Forest can **predict outcomes for new user-product combinations.** Each tree makes an independent prediction based on its learned patterns.

Example:
Tree 1 might predict irritation due to high retinol concentration and dry skin.
Tree 2 might predict a positive outcome based on the user's high melanin levels.
These predictions are aggregated using majority voting—the most common outcome across all trees becomes the final prediction, ensuring robustness and minimizing individual tree biases.

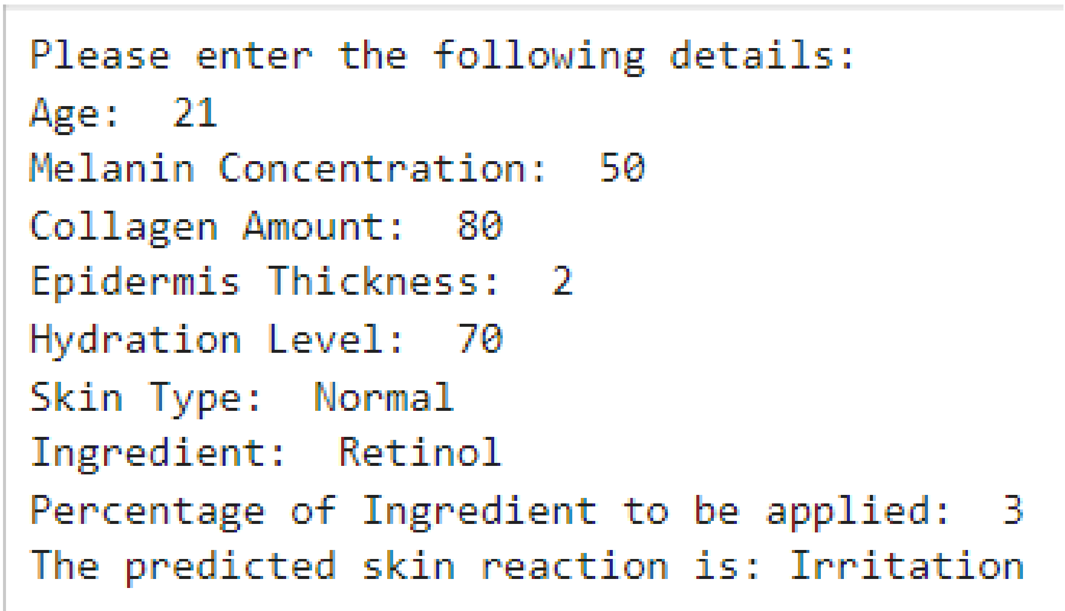
4. **Results:**
Retinol with 3% concentration was tested for skin samples of different melanin concentration to predict reaction on skin of colour. The following results were observed:

Caucasian Skin:



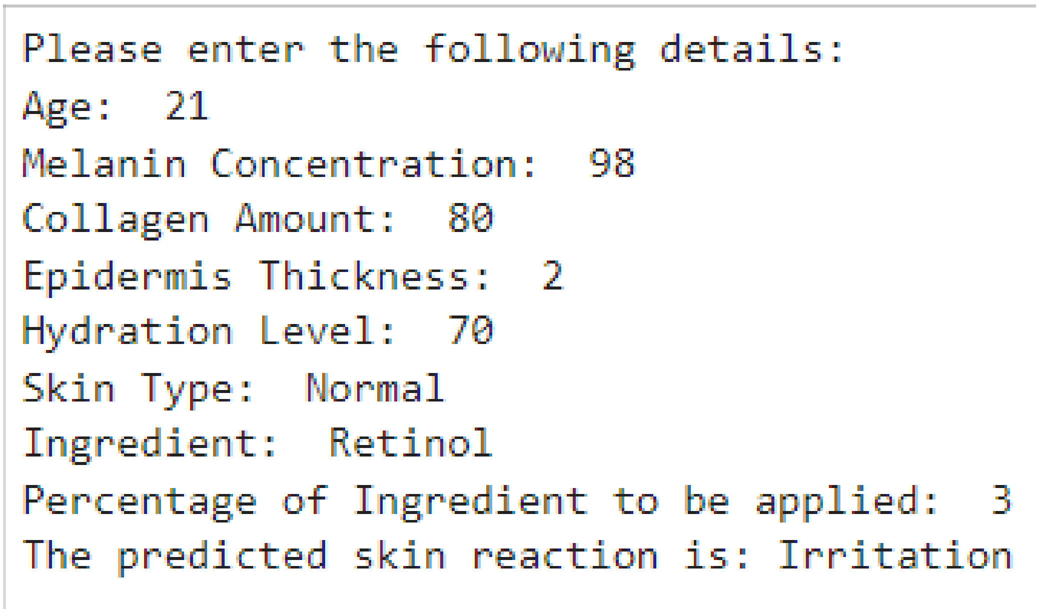
```
Please enter the following details:
Age:  21
Melanin Concentration:  15
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Brightening
```

Brown Skin:



```
Please enter the following details:
Age:  21
Melanin Concentration:  50
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Irritation
```

Black Skin:



```
Please enter the following details:
Age:  21
Melanin Concentration:  98
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Irritation
```

The results accurately predicted that normal skin type of Caucasian women will have a Brightening Reaction, while Brown and Black Skin will have irritation as it is known that skin with higher melanin need to be careful while using highly potent and reactive ingredients such as Retinol, Kojic acid and other exfoliating ingredients like Glycolic and Salicylic acid

7.2 Naïve Bayes:

The Bayes Theorem, based on Naive Bayes, assumes that all attributes, regardless of any relationships, independently contribute to the dataset’s probability. In other words, the Naïve Bayes algorithm predicts a class given a set of features using probability [6].

Naive Bayes is a classification algorithm grounded in Bayes' Theorem, where it assumes that all input features (such as age, melanin levels, and ingredients) are conditionally independent given the class label (e.g., "Irritation" or "Brightening"). Here's a breakdown of how this works:

1. **Data Preparation:**
   The input data is first pre-processed. For features such as skin type and active ingredients (e.g., retinol), which are categorical in nature, one-hot encoding is applied. This process converts each category into binary form—1 if the feature is present and 0 if it's absent—so the model can handle these categorical variables effectively.

2. **Training:**
   During training, the model learns the likelihood of each reaction (like "Irritation" or "Brightening") occurring given specific input features.

   It calculates the probability of each class using the formula:

$$\mathbf{P\ (Class \mid Features) = \frac{[P(Class) \times P\ (Features \mid Class)]}{P(Features)}}$$

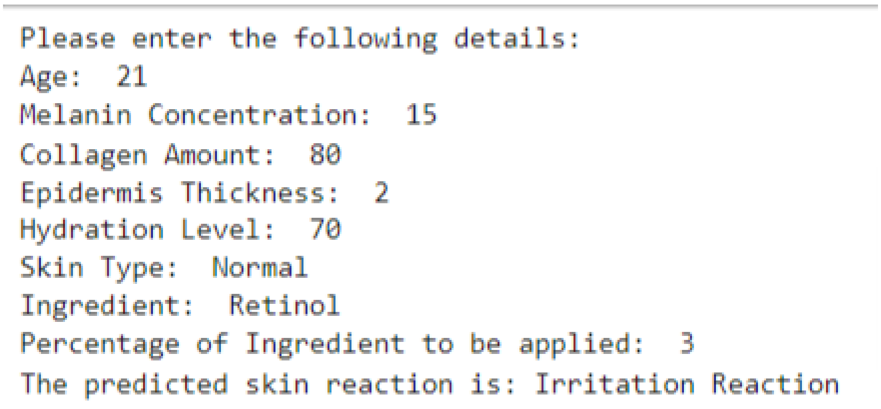
   The likelihood of specific input features (e.g., age=21, melanin=15, ingredient=Retinol) given each class. Using these, it applies Bayes' Theorem to compute the probability of a reaction class based on the observed features.

3. **Prediction:**
   Once trained, the model calculates the posterior probability for each possible reaction class for a new set of input data. For instance, if the input is:
   Age = 21
   Melanin = 15
   Ingredient = Retinol

   The model multiplies the prior probability of each skin reaction with the likelihood of observing those feature values for that reaction. The class with the highest probability is chosen as the predicted reaction.

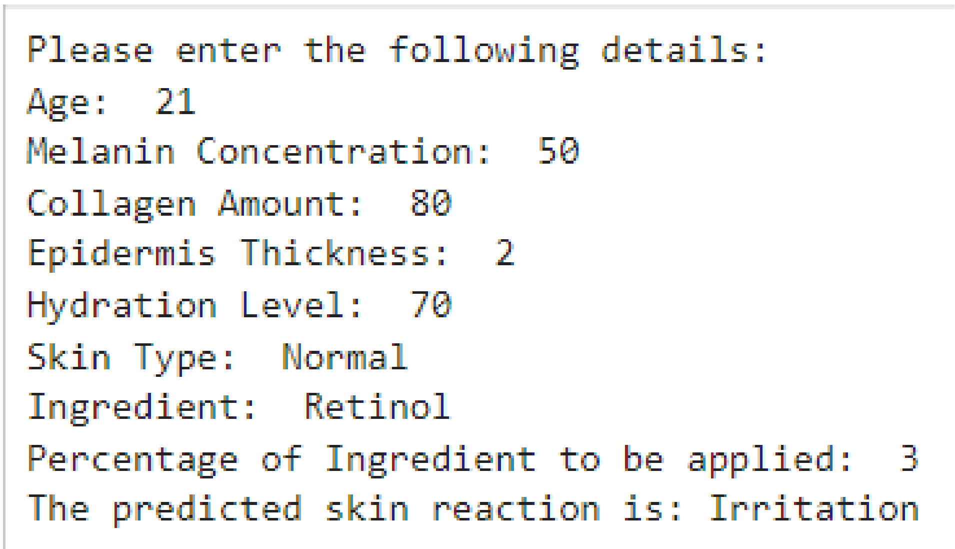
4. Results:

Caucasian Skin:



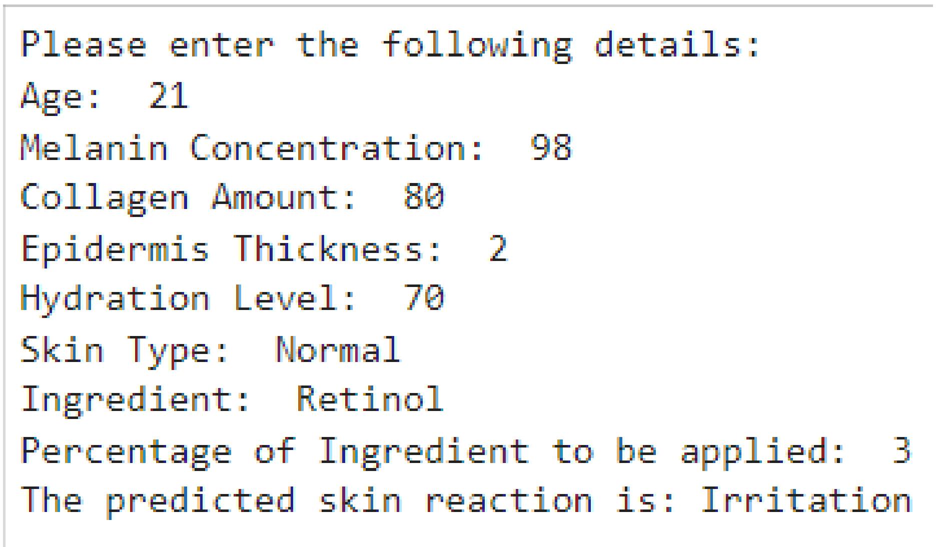
```
Please enter the following details:
Age:  21
Melanin Concentration:  15
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Irritation Reaction
```

Brown Skin:



```
Please enter the following details:
Age:  21
Melanin Concentration:  50
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Irritation
```

Black Skin:



```
Please enter the following details:
Age:  21
Melanin Concentration:  98
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Irritation
```

7.3 Support Vector Machine:

The Support Vector Machines Classifier (SVM), a discriminative classifier formally described by a separating hyperplane, created procedures for an ideal hyperplane using training datasets. SVM is frequently employed to solve classification and regression issues [6]. SVM is effective for both linear and non-linear classification tasks using kernel functions.

1. **Data Pre-processing:**

   Before feeding the data to the SVM model, the categorical variables (e.g., skin type and ingredient) are transformed into numerical form using one-hot encoding. This ensures that the model can process them effectively.

   After encoding, the final input consists of a combination of numerical values representing both the original features and the encoded categorical variables.

2. **Prediction:**

   The Support Vector Machine (SVM) model, which has already been trained on your skincare dataset, is used to predict the skin reaction.

   Given the input features, SVM tries to classify the skin reaction into one of the categories (e.g., "Irritation", "Brightening", "Neutral", etc.).

   SVM works by finding the best boundary (or hyperplane) that separates different reaction classes in the feature space. This boundary is chosen to maximize the margin between the classes (i.e., how far the data points are from the decision boundary).

3. **Decision Making:**

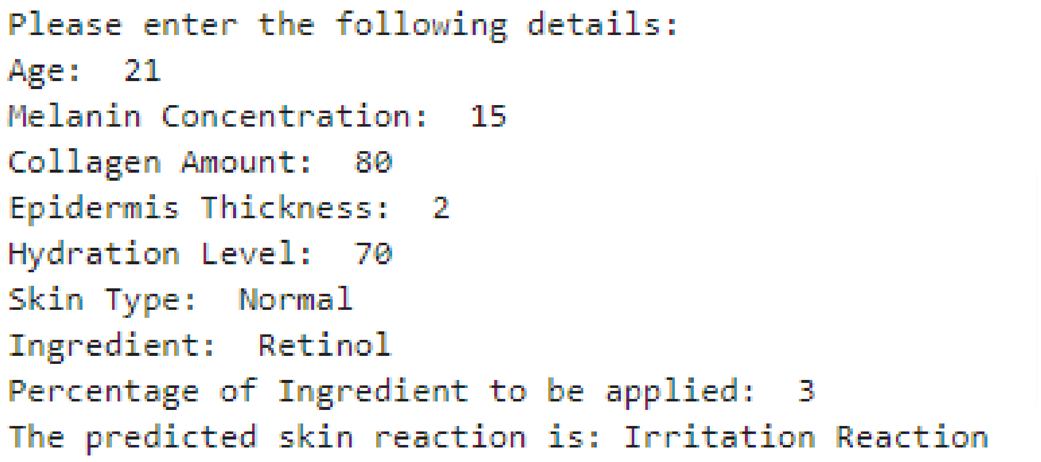
   The model uses these feature values to compute where the new input lies in relation to the learned boundaries of each class.
   Based on this, it assigns the most likely skin reaction category.

   For example, if the input is:
   Age = 21
   Melanin Concentration = 15
   Ingredient = Retinol

   The model uses the values of these features to determine whether the skin reaction will be "Irritation", "Brightening", "Neutral", etc., based on patterns learned from the training data.

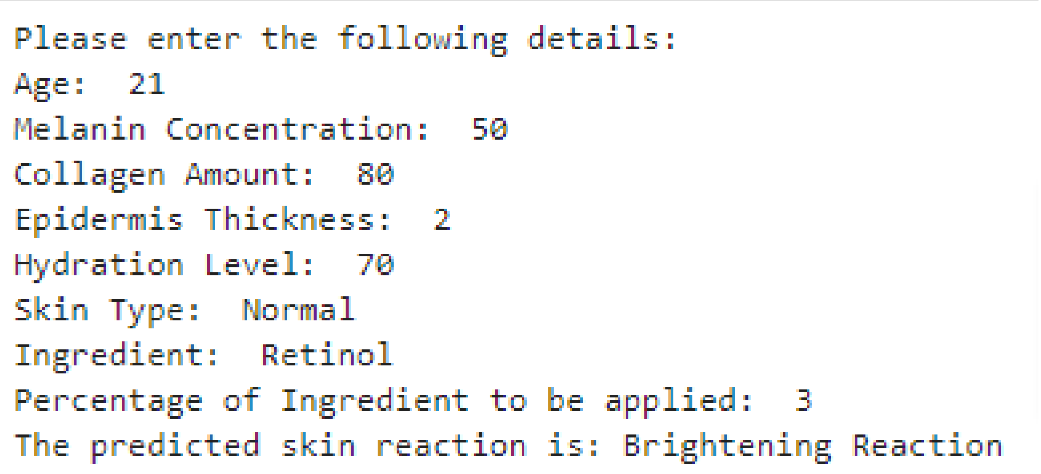
4. Results:

Caucasian Skin:



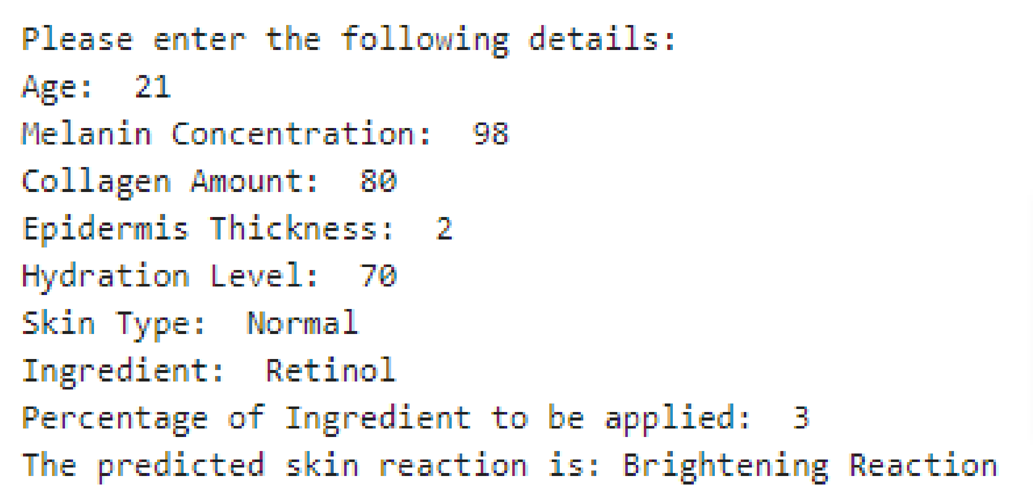
```
Please enter the following details:
Age:  21
Melanin Concentration:  15
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Irritation Reaction
```

Brown Skin:



```
Please enter the following details:
Age:  21
Melanin Concentration:  50
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Brightening Reaction
```

Black Skin:

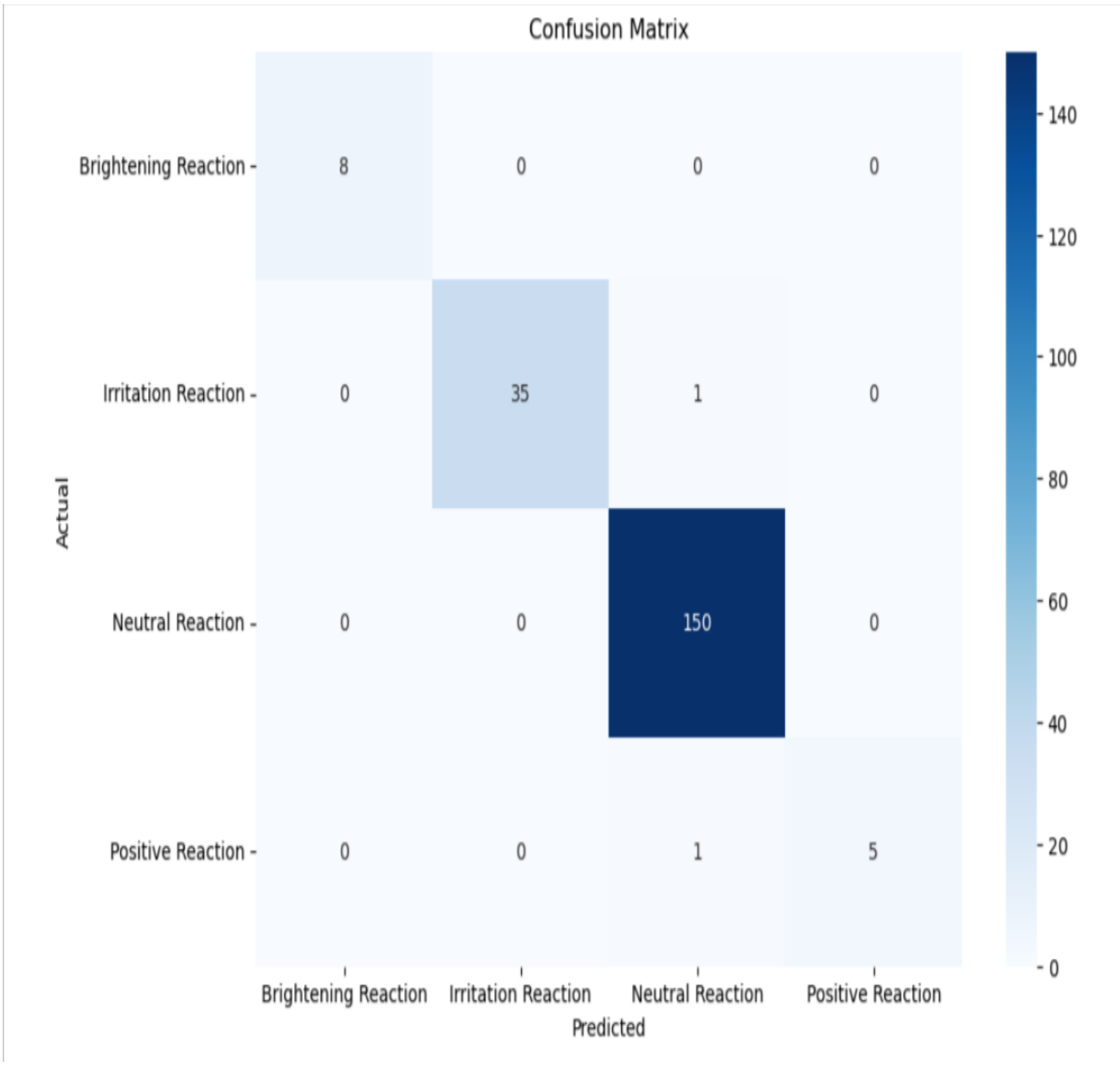


```
Please enter the following details:
Age:  21
Melanin Concentration:  98
Collagen Amount:  80
Epidermis Thickness:  2
Hydration Level:  70
Skin Type:  Normal
Ingredient:  Retinol
Percentage of Ingredient to be applied:  3
The predicted skin reaction is: Brightening Reaction
```

7.4 Model Accuracy Comparison:

To further evaluate the performance of the models, we used confusion matrix function to calculate the matrix of the predicted classes and the true classes. The confusion matrix is a table that illustrates the number of true positives, true negatives, false positives, and false negatives across each class [2].

1. **Random Forest Classifier (RFC):**
    a) Accuracy: 99%
    b) Precision: Overall, RFC achieved high precision across all classes, especially for "Brightening Reaction" (1.00) and "Neutral Reaction" (0.99).
    c) Confusion Matrix:

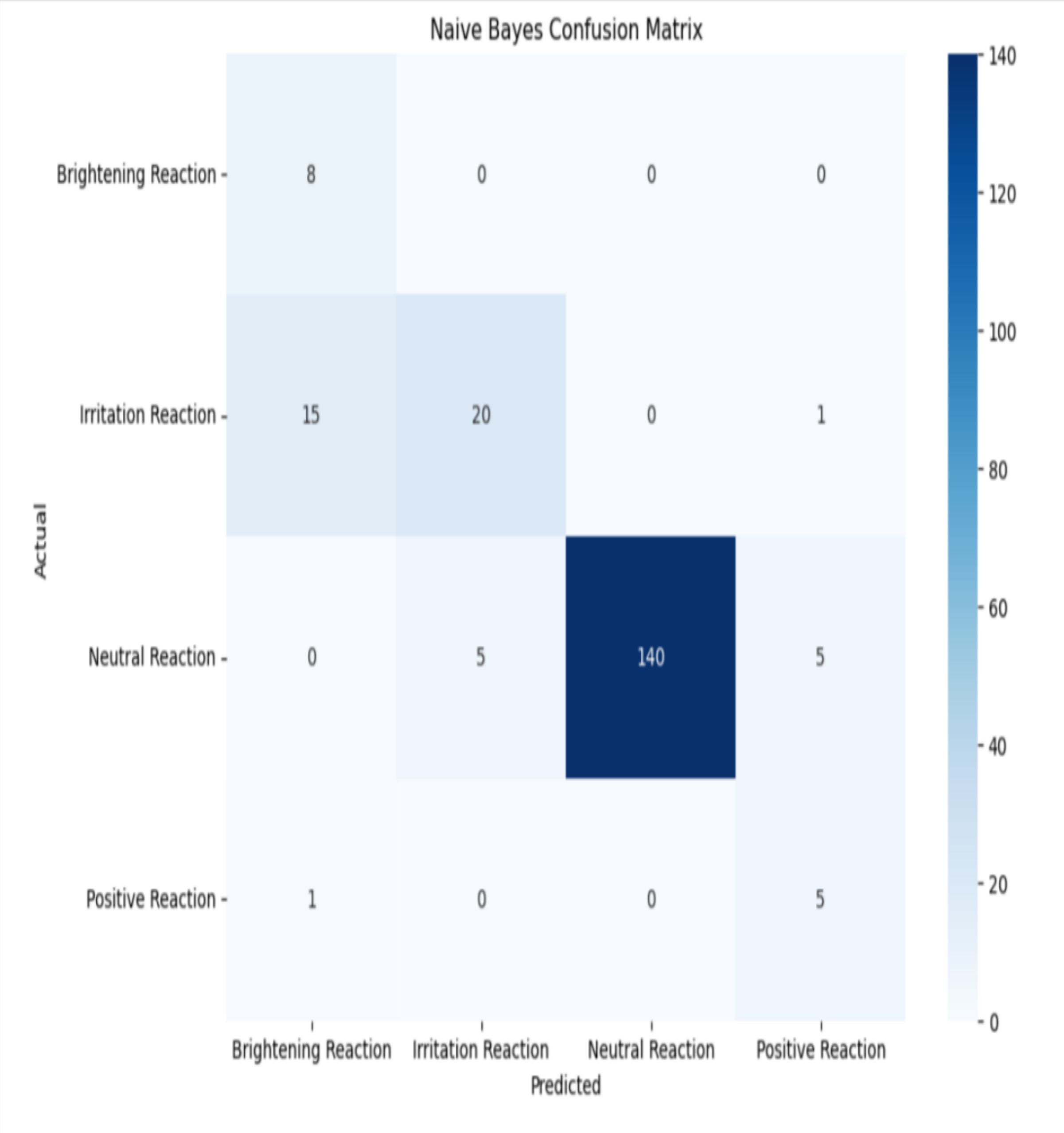


Actual \ Predicted	Brightening Reaction	Irritation Reaction	Neutral Reaction	Positive Reaction
Brightening Reaction	8	0	0	0
Irritation Reaction	0	35	1	0
Neutral Reaction	0	0	150	0
Positive Reaction	0	0	1	5

Explanation:

a) Brightening Reaction: Correctly predicted all 8 instances.
b) Irritation Reaction: 35 correctly predicted, 1 misclassified as Neutral.
c) Neutral Reaction: Perfectly identified all 150 cases.
d) Positive Reaction: 5 correctly identified out of 6, with 1 misclassified as Irritation.
e) Overall: The RFC model performed excellently, with very few misclassifications.

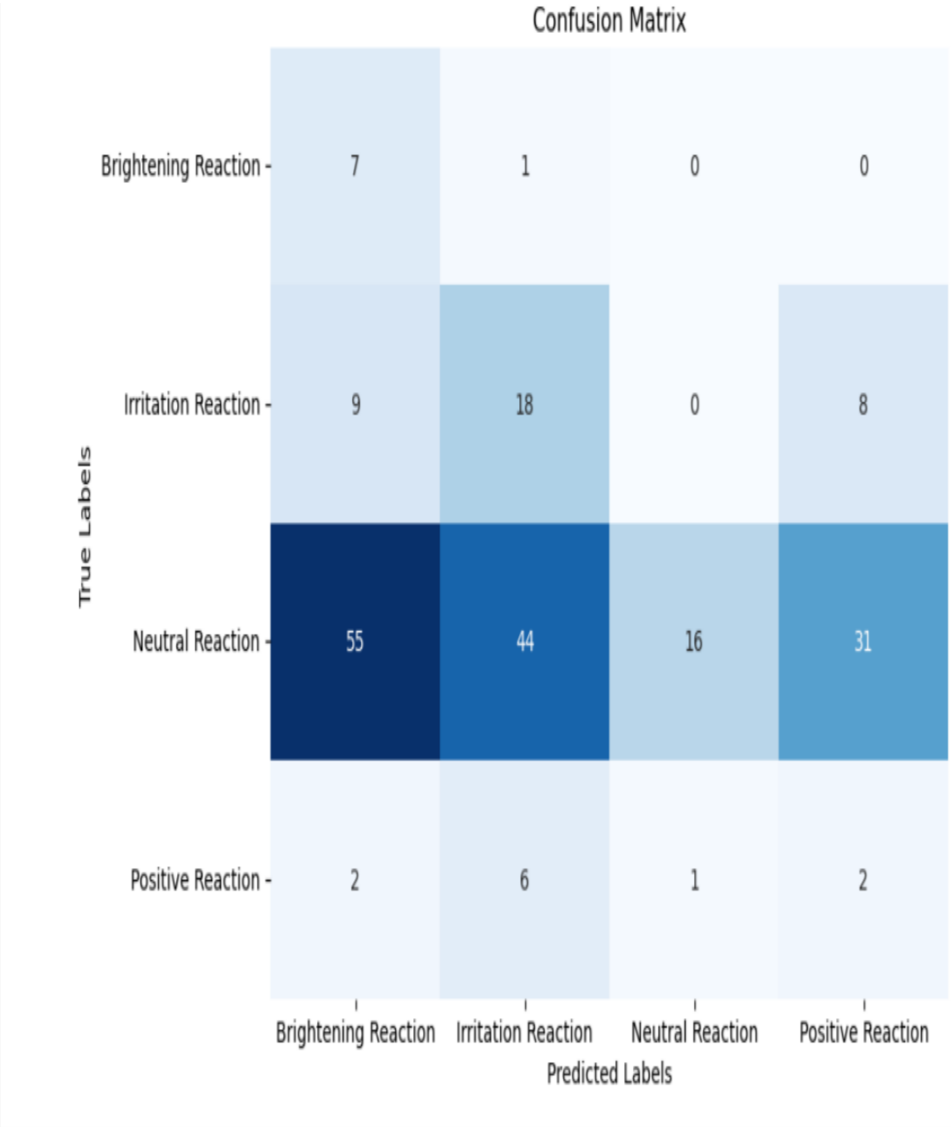
2. **Naive Bayes (NB):**
   a) Accuracy: 86%
   b) Precision: Lower precision than RFC, particularly for "Brightening" (0.33) and "Positive" (0.45) reactions, meaning it made more false positive predictions.
   c) Confusion Matrix:



Explanation:

a) Brightening Reaction: Correctly predicted all 8 instances.
b) Irritation Reaction: 20 correctly predicted, but 15 misclassified as other categories.
c) Neutral Reaction: 140 correctly identified, but 5 misclassified as Irritation and 5 as Positive.
d) Positive Reaction: 5 correctly identified, but 1 misclassified as Irritation.
e) Overall: The NB model had more misclassifications, especially with Irritation.

3. **Support Vector Machine (SVM):**
   a) Accuracy: 21%
   b) Precision: Very low precision for all classes, particularly "Brightening" (0.10) and "Positive" (0.05), suggesting it struggles significantly to identify these classes accurately.
   c) Confusion Matrix:



Explanation:

a) Brightening Reaction: 7 correctly predicted, 1 misclassified as Irritation.
b) Irritation Reaction: 18 correctly identified, but 9 misclassified as Brightening and 8 as Positive.
c) Neutral Reaction: 16 correctly identified, but many misclassified as Irritation or Positive.
d) Positive Reaction: 2 correctly identified, but mostly misclassified.
e) Overall: Many misclassifications across all categories.

It was found that RFC had the highest accuracy, while SVM displayed lowest accuracy.

8 LIMITATION AND FUTURE WORK:

8.1 Limitations:

The use of data science in skincare formulations is indeed promising, but several key limitations must be addressed to ensure inclusivity and efficacy for all skin types. Let us now elaborate on each limitation:

1. **Lack of Research Targeting Non-Caucasian Skin:**

   A significant challenge in the skincare industry is the historical focus on Caucasian skin. Most research and product development have been conducted to address the specific needs and characteristics of White skin. This leads to the creation of products that may be less effective or even unsuitable for people with skin of color (including those of African, Asian, Middle Eastern, or Latin descent).

   The lack of focus on non-Caucasian skin can manifest unintended side effects. For example, certain skincare ingredients or percentage ingredients in the formulations might work well on fair skin but could cause hyperpigmentation, irritation, or reduced efficacy in darker skin due to differences in melanin production, collagen density, and other physiological factors.

2. **Availability of Diverse Skin Data:**

   A crucial limitation in developing personalized skincare solutions lies in the scarcity of comprehensive datasets that account for the structure and physiological makeup of different skin types. Skin color is more diverse in terms of melanin content, response to environmental stressors, propensity for hyperpigmentation, and other factors. However, most of the available datasets focused heavily on Caucasian skin, with limited representation of non-Caucasian skin types.

   Without robust datasets representing diverse skin populations, machine learning models and algorithms cannot be trained adequately to predict the responses of different skin types to various skincare formulations. This lack of diverse skin data results in biases that can hinder the development of effective and personalized skincare products for all consumers.

3. **Model Accuracy:**

   The accuracy of the machine learning models in skincare formulations is another significant limitation. The current models may not account for subtle differences in skin types because they are often trained on imbalanced or homogeneous datasets that underrepresent people with darker skin tones. To improve model accuracy, skincare companies must refine their algorithms by training on more diverse datasets that account for intricate and sometimes subtle variations in skin biology.

4. **Ethical Considerations in Data Collection and Use:**

   The collection and use of personal data, including skin characteristics, poses ethical challenges. Collecting skin data from diverse populations can raise concerns regarding privacy, consent, and data protection. There is a risk of exploitation, especially in communities that have historically been marginalized or underrepresented in research. Transparency in data collection processes, along with stringent privacy measures, is essential to ensure that individuals' data are used responsibly and ethically.

   Furthermore, the use of AI and machine learning to personalize skincare should not lead to the reinforcement of harmful biases or stereotypes. It is crucial that developers take care to mitigate algorithmic biases that could perpetuate disparities in skincare solutions among different populations.

8.2 Future Work:

1. **Expanding Datasets**: It is essential to collect more diverse skin data through clinical trials, consumer research, and collaborations with dermatologists worldwide. This will ensure a more inclusive approach that benefits all skin types.

2. **Improving Model Accuracy**: By using diverse data and refining algorithms, machine learning models can be fine-tuned to better account for the complexities of different skin types, leading to more accurate and personalized skincare formulations.

3. **Addressing Ethical Concerns**: Establishing clear guidelines for ethical data collection and usage, ensuring transparency, and protecting users' privacy will be paramount in ensuring the responsible use of data science in this space.

4. **Pushing for Inclusivity:** It is of great importance for brands to consider different skin types and unique needs to create more inclusive products by considering factors such as melanin content, skin elasticity

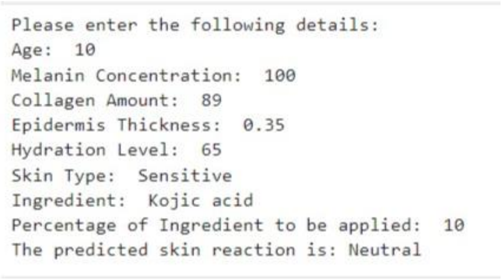
5. **Machine Learning**: Although this project primarily focuses on exploratory analysis and data visualization, potential future applications of machine learning algorithms could enhance the predictive capabilities of the model:

6. **Clustering Techniques**: K-means clustering or hierarchical clustering may be applied to identify natural groupings within the data, potentially revealing new insights into ingredient efficacy across different skin types.

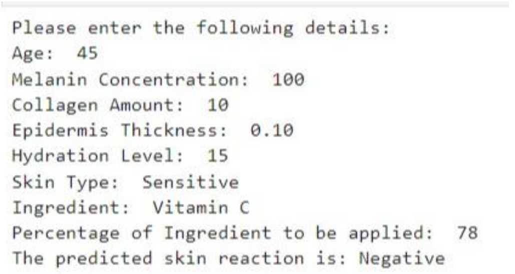
9 RESULT

The following results were seen by using Random Forest Classifier for skin with varying melanin concentrations, age, collagen, etc. for ingredients with different concentrations. It was found out that Random Forest Classifier model predicted the most accurate outcomes among all models. The following testcases were used to portray Random Forest Accuracy with variations in Age, Melanin Concentration, Epidermis Thickness, Hydration Level and Skin type as input parameters to predict the reaction

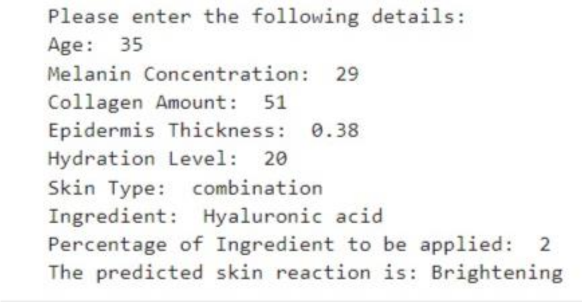
The following results were seen by using Random Forest Classifier for skin with varying melanin concentrations, age, collagen, etc. for ingredients with different concentrations.



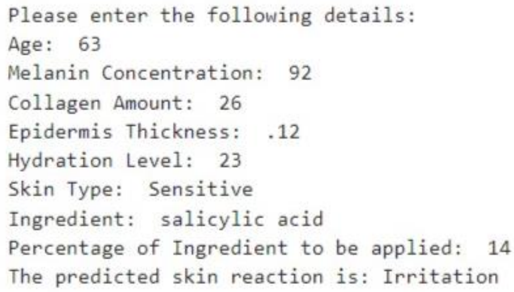
```
Please enter the following details:
Age:  10
Melanin Concentration:  100
Collagen Amount:  89
Epidermis Thickness:  0.35
Hydration Level:  65
Skin Type:  Sensitive
Ingredient:  Kojic acid
Percentage of Ingredient to be applied:  10
The predicted skin reaction is: Neutral
```



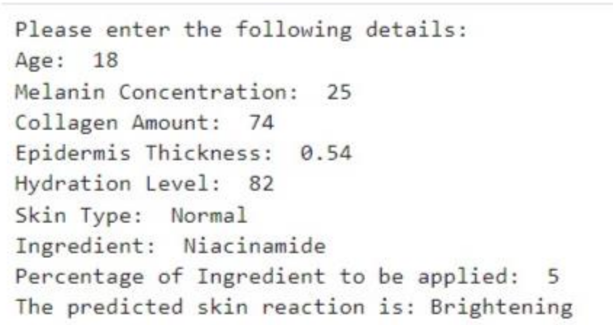
```
Please enter the following details:
Age:  45
Melanin Concentration:  100
Collagen Amount:  10
Epidermis Thickness:  0.10
Hydration Level:  15
Skin Type:  Sensitive
Ingredient:  Vitamin C
Percentage of Ingredient to be applied:  78
The predicted skin reaction is: Negative
```



```
Please enter the following details:
Age:  35
Melanin Concentration:  29
Collagen Amount:  51
Epidermis Thickness:  0.38
Hydration Level:  20
Skin Type:  combination
Ingredient:  Hyaluronic acid
Percentage of Ingredient to be applied:  2
The predicted skin reaction is: Brightening
```



```
Please enter the following details:
Age:  63
Melanin Concentration:  92
Collagen Amount:  26
Epidermis Thickness:  .12
Hydration Level:  23
Skin Type:  Sensitive
Ingredient:  salicylic acid
Percentage of Ingredient to be applied:  14
The predicted skin reaction is: Irritation
```



```
Please enter the following details:
Age:  18
Melanin Concentration:  25
Collagen Amount:  74
Epidermis Thickness:  0.54
Hydration Level:  82
Skin Type:  Normal
Ingredient:  Niacinamide
Percentage of Ingredient to be applied:  5
The predicted skin reaction is: Brightening
```

10 CONCLUSION

The use of data science in skincare formulation offers immense potential to create more inclusive and effective products. By leveraging the power of Data Science, skincare companies can optimize their formulations for a wide variety of skin types and conditions. The continuous feedback loop ensures that these formulations evolve with new data, improving product efficacy and personalization.

The following accuracy for predictions were noted for the three models:

1. **Random Forest Classifier (RFC):**
   Accuracy: 99%
   Precision: Overall, RFC achieved high precision across all classes, especially for "Brightening Reaction" (1.00) and "Neutral Reaction" (0.99).
2. **Naive Bayes (NB):**
   Accuracy: 86%
   Precision: Lower precision than RFC, particularly for "Brightening" (0.33) and "Positive" (0.45) reactions, meaning it made more false positive predictions.
3. **Support Vector Machine (SVM):**
   Accuracy: 21%
   Precision: Very low precision for all classes, particularly "Brightening" (0.10) and "Positive" (0.05) suggesting it struggles significantly to identify these classes accurately.

References

[1] A. Georgievskaya, T. Tlyachev, D. Danko, K. Chekanov, and H. Corstjens, "How artificial intelligence adopts human biases: The case of cosmetic skincare industry", *AI and Ethics*, vol. 4, no. 2, pp. 85-95, 2023

[2] B. Lokesh, A. Devarakonda, G. Srinivas, and N. K. Naik, "Intelligent facial skin care recommendation system”, Afr. *J. Bio. Sci.*, vol. 6, no. Si2, pp. 1822-1830, 2024.

[3] C. D. Kaur and S. Saraf, "Skin care assessment on the basis of skin hydration, melanin, erythema, and sebum at various body sites", *Asian Journal of Pharmaceutical and Clinical Research*, vol. 4, no. 2, pp. 40-45, 2011.

[4] Dr. Shahana Tanveer, Sama Khatoon, H. U. Begum, and U. Zainab, “Korean Skin Care Recommendation System”, *International Journal of Information Technology and Computer Engineering*, vol. 12, no. 2, pp. 797, ISSN 2347–3657, 2024.

[5] E. Markiewicz and O. C. Idowu, "Personalized skincare: from molecular basis to clinical and commercial applications", 2018.

[6] Fadly, D. Marlina, T. B. Kurniawan, M. Z. Zakaria, and S. F. Abdullah, "Sentiment analysis on natural skincare products", *Journal of Data Science*, vol. 12, pp. 1-17, 2022, ISSN 2805-5160, 2022.

[7] FITZPATRICK, T.B., “The Validity and Practicality of Sun Reactive Skin Types I Through VI.”, *Archives of Dermatology*, 124: 869-871, 1988.

[8] GRIMES, P.E., “Skin and Hair Cosmetic Issues in Women of Color.”, *Dermatological*

Clinics, 18(4), Oct: 659-665, 2000

[9] Jinhee Lee, Huisu Yoon, Semin Kim, Chanhyeok Lee, Jongha Lee, Sangwook Yoo, “Deep learning-based skin care product recommendation: A focus on cosmetic ingredient analysis and facial skin conditions”, *Journal of Cosmetic Dermatology,* vol. 23, no. 6, pp. 2066-2077, 2024.

[10] L. J. Teixeira, "Specific cosmetic and skincare needs of women of color in South Africa", *South African Journal of Dermatology*, vol. 12, no. 1, pp. 60-72, 2006.