Application of Data Science in Skin Care Formulation

Farkhanda Dalal
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India
7farkhanda@gmail.com

Misbah Bagwan
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India
misbahiqbalbagwan@gmail.com

Mansi Jangle
Department of Computer Engineering
Pune Institute of Computer Technology
Pune, India
mansijangle559@gmail.com

***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.

***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**

I. 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 but marketed to women of all skin colors [3]. As a result, many skincare solutions are ineffective for individuals with different ethnic backgrounds, skin tones, and conditions. 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 [3] 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.

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 melanin content, collagen, skin elasticity, pigmentation, sensitivity, and environmental impact, the skincare industry can better understand the needs of all consumers. 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.

II. NEED OF INCLUSIVE FORMULATIONS

A. *Classification of Skin Types:*

Sun-reactive Typing:
The concept of sun-reactive "skin typing" was created in 1975 to classify persons *with white skin* for treatment of psoriasis. Clinical research was conducted by T.B. Fitzpatrick from 1988. This paper which categorizes skin types based on how they react to sun exposure as shown in Fig. 1. The Fitzpatrick classification system classifies women of color as type IV, V, and VI in the table below [5]. Even though this is the most well-known and generally accepted in the somatology industry, it has limitations as not all women of color can be accommodated in this scale [3].

Skin Type	Skin Colour	Characteristics
I	White	Always burns, never tans
II	White	Usually burns, tans less than average
III	White	Sometimes burns, tans about average
IV	White	Rarely burns, tans more than average
V	Brown	Rarely burns, tans profusely
VI	Black	Never burns, deeply pigmented

Fig. 1. Fitzpatrick Classification of Human Skin

B. *Structural and Physiological Differences in Skin Characteristics Based on Ethnic Grounds:*

Human skin has structural differences based on four major characteristics namely, melanin content, TEWL, skin irritation and sebum levels [6]. These parameters have a profound impact on how skin behaves and reacts to environmental factors, treatments, and products. Hence, these differences are of significance when we think of inclusive skin care formulation.

The primary short coming of the current skincare industry is that these differences are neglected during the Research and Development phase of manufacturing. As a result, we see an array of formulations that are unfit for use, for non-Caucasian skin.

TABLE I. STRUCTURAL AND PHYSIOLOGICAL DIFFERENCES IN SKIN CHARACTERISTICS BASED ON ETHNIC GROUNDS

Property	Caucasian	Brown	Black
Melanosomes	Small, approx. (0.5nm * 0.3 nm) [4]	Larger, approx. (0.6nm * 0.3 nm) [4]	Largest, approx. (1.0nm * 0.5 nm) [4]
Higher Melanin Concentration is seen at	Chin and Neck [3]	Neck and Periauricular space [3]	Fore Head and Chin [3]
Dermis	Thinner, between 1mm to 2mm	Intermediate, between 1mm to 3mm	Thicker, between 1.5mm to 4mm
Aging Results In	Photodamage, Loss of Collagen [1]	Dyspigmentation [1]	Frequent xerosis, Hyperpigmentation [1]
Irritation Causes Following Reaction	Erythema	Erythema, Hyperpigmentation	Results in darkening of effected area instead of Erythema
Sebum Production	More on Fore Head, Periauricular space, Chin [3]	More on Fore Head, Periauricular space, Chin [3]	More on Fore Head, Periauricular space, Cheek [3]

C. Under Representation of Skin of Color:

The traditional method of formulation is deep rooted to cater to the European or Caucasian skin type. Hence, previously products were developed to cater to their needs and wants, whereas the same products were also marketed towards non-Caucasian women with less consideration to how these products would address their needs and how their skin would react to these formulations.

This problem is particularly prevalent in dermatological and cosmetic literature, including dermatology textbooks, where a racial bias is reflected in the underrepresentation of images of minorities compared to the general population, which can lead to inequalities in skin health care. [1]

In an article published by King (1998) in the USA, in which 21 cosmetic company executives were interviewed, it was found that most cosmetic companies do not market their products differently to women of color, neither do they have special cosmetic collections for these diverse ethnic groups [6].

This biasness has been prevalent even with the rise of Artificial Intelligence and Machine Learning and its implementation in the Cosmetics and Personal Care Industry.

AI has proven to inherit and adopt human biases and this is clearly evident as there is lack of focus on diverse skin types and skin concerns. Here are a few examples:

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. [1] 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.

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]

III. METHODOLOGY

A. Data Collection:

Data was collected from diverse sources, including dermatological studies, clinical trials, and skincare product databases.

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.
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.

B. Data Processing:

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

1. **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.
2. **Outlier Detection**: Statistical methods (e.g. boxplot) were employed to detect and remove outliers that could skew the analysis.

C. Data Exploration:

a. 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.

b. Visualization Techniques:

- Visualization of trends in Data Set Samples:

1. Distribution of Skin Types:

As shown in Fig. 2 A diverse amount of skin types is present in the dataset, with Normal skin type having the most samples (228) and oily having least (172).

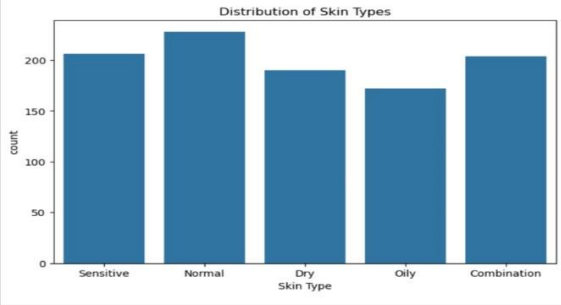


Fig. 2. Distribution of Skin Types in Data Set.

2. Distribution of Collagen Amount:

As shown in Fig. 3 A significant number of samples have 65-70 AU of Collagen.

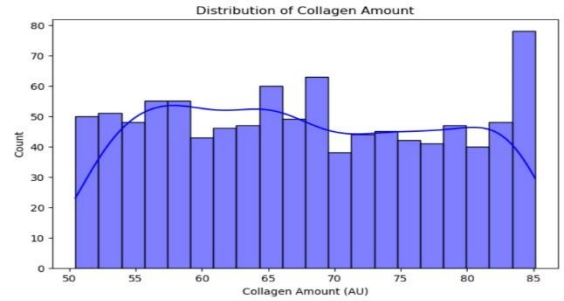


Fig. 3. Distribution of Collagen Amount in Data Set

3. Boxplot of Hydrations:

As shown in Fig. 4 majority of the samples have collagen between approximately 40 and 60.

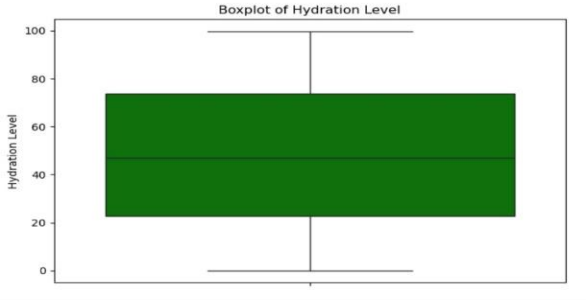


Fig. 4. Boxplot of Hydration

4. Average Collagen Amount by Age:
   As shown in Fig. 5 the average amount of collagen takes a dip as age tends to increase.

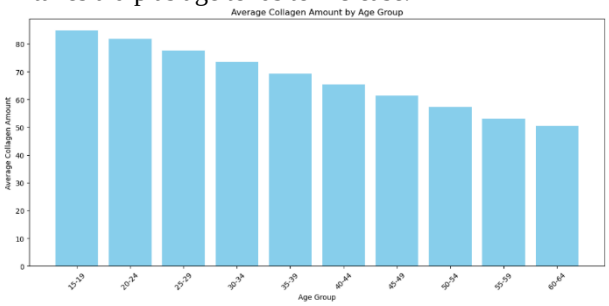


Fig. 5. Average Collagen Amount by Age

- Predicted Reaction of various skincare ingredients at different melanin concentrations:

  As shown in Fig.6 reactions of ingredients vary with variation in melanin concentration, to name a few:

  1. **Salicylic Acid & Azelaic Acid**: These ingredients show **neutral reactions** across a wide range of melanin concentrations, including lower levels.
  2. **Kojic Acid**: At **lower melanin concentrations**, Kojic Acid tends to cause a **positive reaction**, this effect tapers off as melanin concentration increases.
  3. **Vitamin C**: This ingredient results in a **brightening reaction** at **higher melanin concentrations**, and less reactive at lower melanin levels.

Fig. 6. Reaction of various skincare ingredients at different melanin concentrations

IV. 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. Three models were used to perform this, namely:

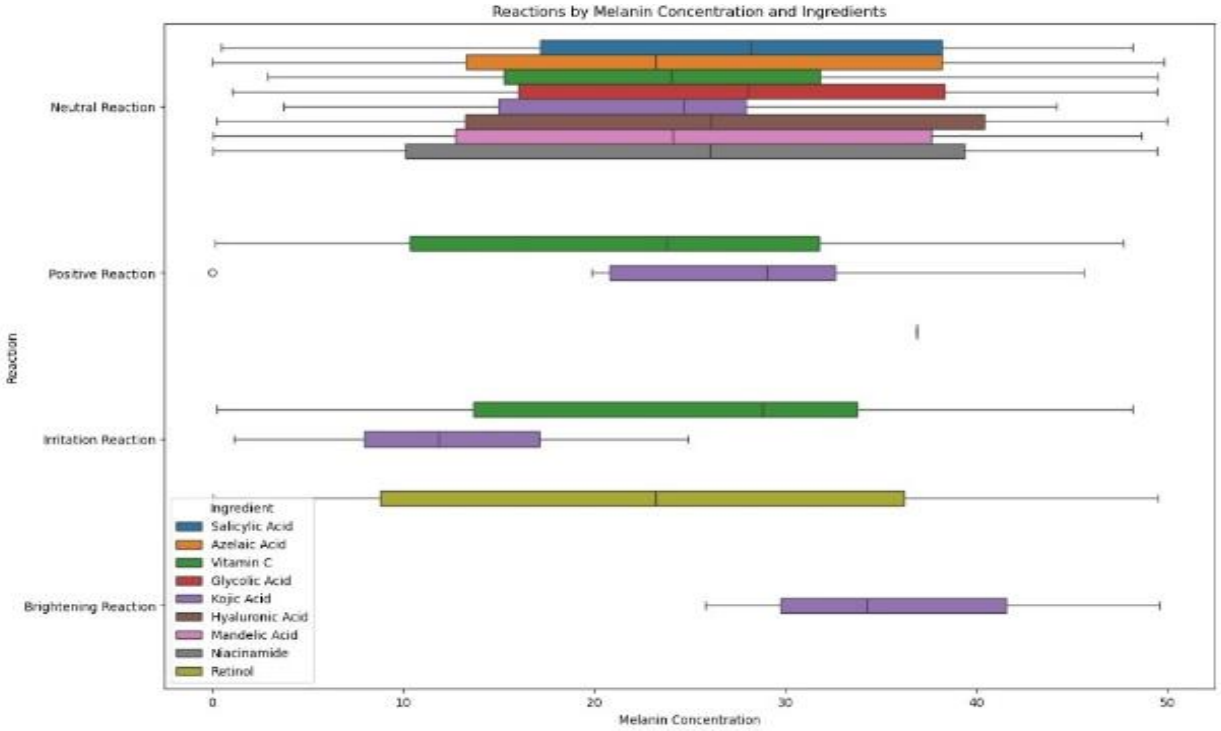
A. Random Forest Classifier (RFC):

Random Forest is a supervised machine learning algorithm that constructs multiple decision trees during training and outputs the most common prediction from the trees (majority voting).

Algorithm:

1. The dataset includes features like Age, Melanin Concentration, Collagen Amount, Epidermis Thickness, Hydration Level, Skin Type, Ingredient, and Concentration.
2. Categorical features like Skin Type and Ingredient are encoded using One-Hot Encoding to convert them into numerical form.
3. The dataset is split into training and testing sets to ensure model accuracy.
4. The Random Forest algorithm is then trained on this dataset to learn the relationships between these features and different types of skin reactions (e.g., irritation, brightening, neutral).

The trained model predicts the most likely reaction to skincare products based on user inputs.



B. Naïve Bayes Classifier:

The Naïve Bayes algorithm operates under the assumption that all features independently contribute to determining the reaction, despite potential correlations between them. The Naïve Bayes algorithm predicts a class given a set of features using probability [6]. This assumption allows it to handle multi-class prediction tasks, such as determining whether the skin's response will be "Irritation," "Brightening," or "Neutral."

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (1)$$

Algorithm:

1. **Data Preparation:** Categorical features, like skin type and ingredient type, are converted into a numerical format using one-hot encoding. This ensures that the model can properly interpret these variables during training.
2. **Model Training:** The algorithm calculates the probability of each skin reaction, given the various user-specific features. By applying **Bayes' Theorem**, it combines the likelihood of a particular skin reaction with the observed feature values (such as melanin level and age) to build a prediction model.
3. **Prediction Process:** For new data, like a person's age, melanin concentration, and chosen skincare ingredient, the model computes the probability of each possible reaction. The final prediction is based on the reaction class with the highest calculated probability.

C. Support Vector Model:

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].

Algorithm:

1. **Data Pre-processing:** Categorical variables, such as skin type and ingredient type, are converted into numerical form using one-hot encoding to allow SVM to process them. The resulting input consists of numerical values for all features, both original and encoded.
2. **Prediction:** The SVM model, trained on the skincare dataset, classifies new input data (like age and melanin concentration) into one of the skin reaction categories. It determines which class the data belongs to by calculating the position of the input relative to the learned decision boundaries in the feature space.
3. **Decision Making:** Based on the location of the new input in relation to the hyperplanes, the model assigns the most likely skin reaction category. For example, if the input includes features like age, melanin concentration, and a specific ingredient, SVM predicts the skin reaction based on the patterns it learned during training.

V. 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].

A. RFC:

It has accuracy of 99% and high precision across all classes, especially for "Brightening Reaction" (1.00) and "Neutral Reaction" (0.99).

Fig. 7 helps us draw the following conclusions:

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

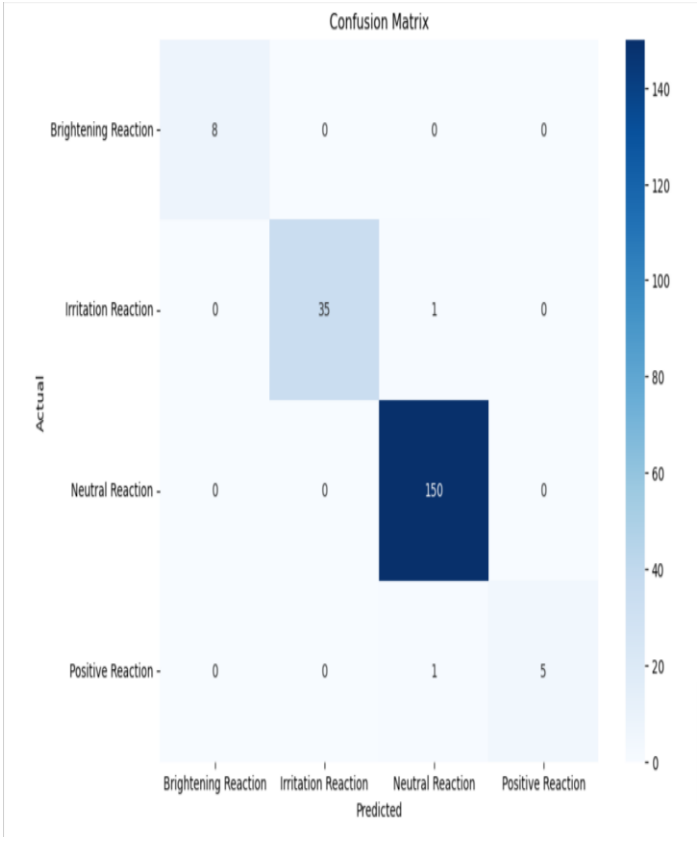


Fig. 7. RFC Confusion Matrix

B. Naïve Bayes Classifier:

It has accuracy of 86% and has Lower precision than RFC, particularly for "Brightening" (0.33) and "Positive" (0.45) reactions, meaning it made more false positive predictions.

Fig. 8 helps us draw the following conclusions:

1. Brightening Reaction: Correctly predicted all 8 instances.
2. Irritation Reaction: 20 correctly predicted, but 15 misclassified as other categories.
3. Neutral Reaction: 140 correctly identified, but 5 misclassified as Irritation and 5 as Positive.
4. Positive Reaction: 5 correctly identified, but 1 misclassified as Irritation.
5. Overall: The NB model had more misclassifications, especially with Irritation.

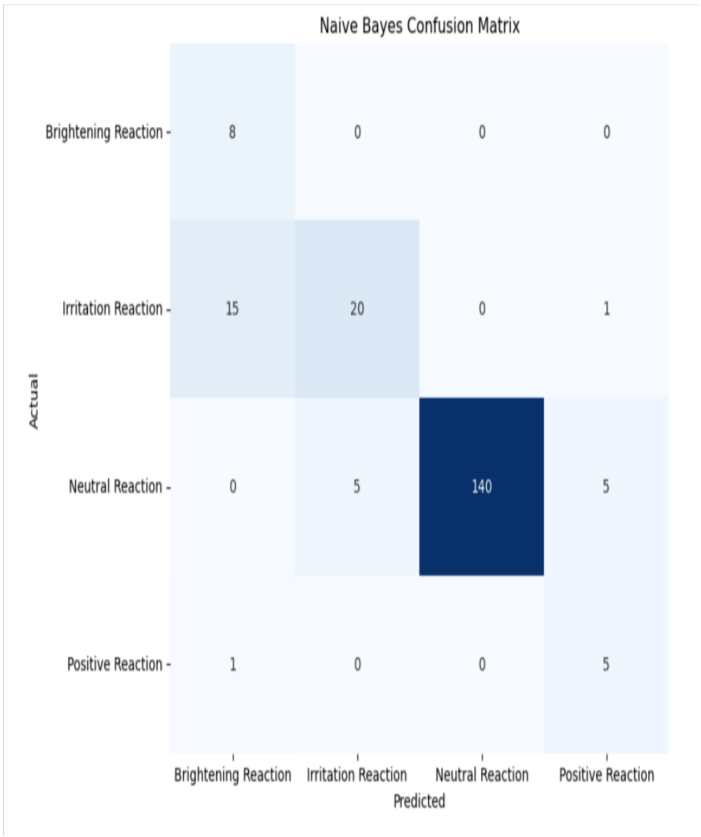


Fig. 8. Naïve Bayes Classifier

C. *Support Vector Machine:*

It has accuracy of 21% and has very low precision, particularly for "Brightening" (0.10) and "Positive" (0.05) reactions, suggesting it struggles significantly to identify these classes accurately.

Fig. 9 helps us draw the following conclusions:

1. Brightening Reaction: 7 correctly predicted, 1 misclassified as Irritation.
2. Irritation Reaction: 18 correctly identified, but 9 misclassified as Brightening and 8 as Positive.
3. Neutral Reaction: 16 correctly identified, but many misclassified as Irritation or Positive.
4. Positive Reaction: 2 correctly identified, but mostly misclassified.
5. Overall: Many misclassifications across all categories.

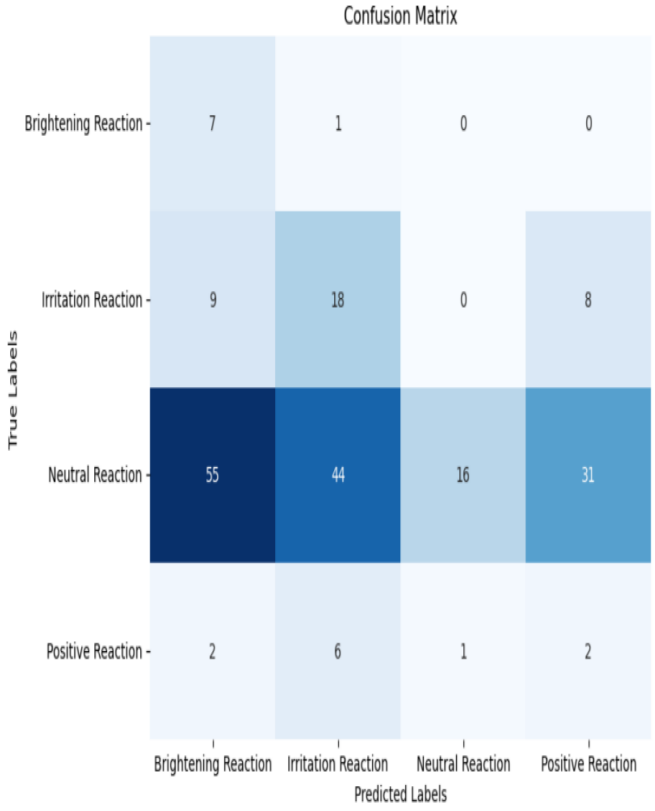


Fig. 9. Support Vector Machine

Random Forest Classification showed highest accuracy while Support Vector Machine had the lowest.

VI. FUTURE WORK

A. *Future Work:*

1. **Expanding Datasets**: Gather diverse skin data globally for inclusive skincare solutions.
2. **Inclusivity**: Brands should cater to different skin types, considering factors like melanin and elasticity.
3. **Ethical Concerns**: Ensure ethical data collection, transparency, and user privacy.
4. **Clustering Techniques**: Use K-means or hierarchical clustering to uncover ingredient efficacy insights.

VII. 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

```
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
```

VIII. 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):
   a) Accuracy: 99%
   b) Precision: Overall, RFC achieved high precision across all classes, especially for "Brightening Reaction" (1.00) and "Neutral Reaction" (0.99).
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.
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.

Therefore, we arrived to the conclusion that Random Forrest Classifier turned out to be the most accurate out of the three models used

ACKNOWLEDGMENT

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.

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 Colour. 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.