Evaluating Database Systems and Analyzing Semi-structured Dataset

Bolormaa Mendbayar- x23176725

National College of Ireland

MSc in Data Analytics

Database and Analytics Programming

Semester 1, 2023/2024

Abstract— *Cosmetics play a significant role in daily personal care routines, fashion, and self-expression, catering to diverse needs and preferences. The industry's dynamism and continual innovation ensure a wide array of choices for consumers worldwide.*

This study examines a comprehensive dataset of cosmetic products to analyze industry trends, consumer preferences, brand performance and Positioning. Our analysis reveals notable trends in product categories and pricing strategies, shedding light on shifting consumer preferences. The findings from this study provide valuable insights for industry stakeholders, including marketers, brand managers, and consumers. Understanding these trends and consumer preferences is crucial for strategic decision-making, market positioning, and product development in the cosmetics industry.

I. INTRODUCTION

The makeup and skincare world is always changing, offering tons of options for people everywhere. This study looks at a bunch of makeup and skincare products to see what's popular, how much things cost, and which brands people like. By doing this, we can figure out important things that help companies make better decisions about what to sell and how to sell it.

II. RELATED WORK

Various experts have taken a close look at different aspects of the cosmetic industry.

Economical Perspectives

1. How Beauty Business Works: Alexis (2019) dug into how money moves in the beauty world, showing how it all adds up [1].

2. Marketing Strategies in Cosmetics: Karchin and Horvath (2023) explored new ways to sell makeup, looking at the tricks and strategies brands use [2].

3. Creating and Selling Makeup: Salvador and Chisvert (2011) explained how makeup is made and sold, giving insights into why people like certain products [3].

The Technical Side

1. Dealing with Big Data: Abadi and Madden (2016) introduced HadoopDB, a big technology mix for handling tons of data [4].

2. Big Data in Health: Chen and Qin (2018) showed how big data helps doctors and patients, not just in makeup but also in healthcare [5].

3. Keeping Data Neat: Wickham (2014), Wilson et al. (2017), and McKinney (2017) talked about keeping data tidy, doing good research, and using Python for data jobs [6] [7] [8].

Consumer-Centric side

1. Picking the Right Makeup: Sheng et al. (2019) and Sharma and Rai (2018) studied how to suggest makeup that people would love, based on what they like [9] [10].

2. Guessing What People Might Do: Gupta and Kumar (2017) tried to figure out what people might buy or like in the makeup world [11].

3. Fancy Tech Predicting Makeup: Li et al. (2020) used smart tech to guess details about makeup products, telling us more about what they're made of [13].

CRISP-DM (Cross-Industry Standard Process for Data Mining) and KDD (Knowledge Discovery in Databases) are methods that help researchers find patterns in data. They are like step-by-step guides used in different types of data research. Although not directly mentioned in the studies, these methods help structure and guide how researchers dig into data [12] [14].

All these studies give us a mix of ideas about the makeup world, from money matters to using cool tech, and understanding what customers prefer. They help us look deeper into how the cosmetic industry works in terms of money, technology, and what customers like.

III. Methodology

3.1 Dataset Description

https://www.kaggle.com/datasets/shivd24coder/cosmetic-brand-products-dataset?select=makeup_data.json

This dataset fits in perfectly with the project goals of investigating the cosmetic sector and market research.

A comprehensive Makeup Product Dataset: The data covers cosmetic brand and products, price, rating and description etc. Raw data was collected in JSON formats from the Kaggle. The dataset has 20 columns and 931 records.

The variables in the dataset are:

No	Variable	Type
1	id	int64
2	brand	object
3	name	object
4	price	object
5	price_sign	object
6	currency	object
7	image_link	object
8	product_link	object
9	website_link	object
10	description	object
11	rating	float64
12	category	object
13	product_type	object
14	tag_list	object
15	created_at	object
16	updated_at	object
17	product_api_url	object
18	api_featured_image	object

Table 1: Data ***set variables***

The dataset offers a wide range of attributes shown as Table 1, it is possible to conduct a complex analysis and investigate several aspects of the cosmetics business such as pricing and rating patterns, and product categories.

3.2 Detailed Description of Data Preprocessing

The action process of the Extract Transform and Load or ETL process has been carried out as below shown. This whole process has been done to get the final analysis what the trend for makeup brands and products.

- Find the dataset on the internet
- Connect and insert dataset to the MongoDB
- Preprocessing dataset
- Visualizations
- Connect to Postgresql and convert to CSV
- Modelling

Downloading dataset from the Kaggle, Java Script Object Notation or JSON data has been stored to the MongoDB that is no Structured Query Database or NoSQL database The explanation for using the MongoDB database—which is neither a NoSQL nor a structured query database—is that it allows for the flexibility and scalability of data, which is neither allocated nor kept in memory in a sequential manner. Wherever there was the least amount of memory demand, that is where the data was saved. MongoDB is also a schema less database. Additionally, this has the benefit of allowing for document size and content variations. Because most of the data in this database is kept in RAM, it performs queries significantly more quickly than other databases that do not use structured query languages. Since this database is schema-less, there is no need to create a schema, which will save the user a great deal of time.

After fetching the data has been converted to key-value pairs for the purpose of storing MongoDB. After this, created data frame, firstly, in data cleaning same data frame has been done where unwanted columns such as price_sign, image_link, product_link, website_link, created_at, updated_at, product_api_url, api_featured_image, product_colors, _id have been removed, and changed data frame into 931 records and 10 columns. When checking datatype, price datatype was object, so changed to numeric type.

Secondly, checked NULL, NA, and junk values, as a result no NULL values were in the data frame. And then created boxplot to check outliers using all numeric variables, see the figure below.

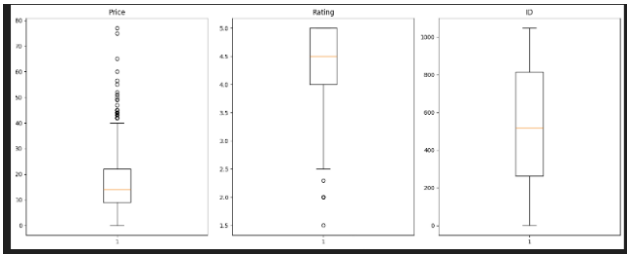

Figure 1. Before dropping outliers

Thirdly, verified outliers and dropped them, after dropping the outliers got 334 records. As shown below in Figure 2, we can see that outliers were removed.

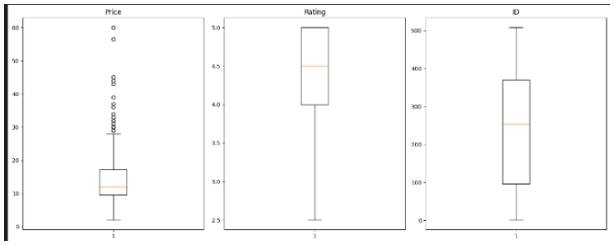

Figure 2. After dropping outliers

Fourthly, created correlation matrix using all numeric variables to see relationship between them, as shown below variables were not highly correlated.

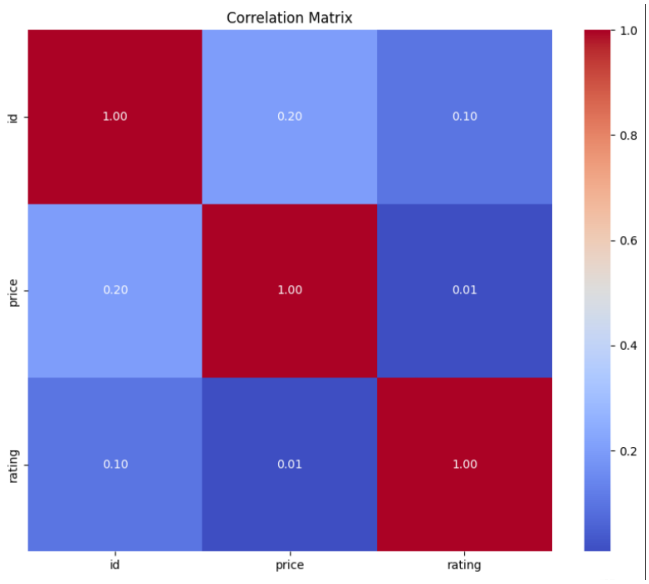

Figure 3. Correlation matrix

3.3 Technologies used

In this project I used Python programming language, and as databases MongoDB and Postgresql. Justification of choosing databases is as I mentioned in *3.2, MongoDB* is schema-less, allowing for dynamic and flexible data modelling without predefined schemas. It allows for handling semi-structured and unstructured data. Moreover, the reason of choosing Postgresql is an open-source relational database management system(RDBMS), offers several advantages that make it strong choice for various applications:

- Reliability and Robustness
- Extensive Features and SQL Support
- Scalability Performance
- Compatibility and Ecosystem

In terms of Python language, it has extensive libraries, visualization capabilities, ease of learning and usability. Recently Python has become an essential requirement in the analysis industry. Additionally, countless python libraries have been imported such as json, numpy, os, pandas, pymongo, bson, seaborn, psycopg2, matplotlib, scikit-learn, itertools.cycle, and sqlalchemy.create_engine.

IV. RESULTS AND EVALUATION

The project's outcomes involved an evaluation based on a refined DataFrame, labeled as 'df', showcasing the cleaned dataset.

4.1 Mean price by product type

Displayed in Figure 4 below, it's noticeable that the most expensive product type is bronzer, priced at more than 20 dollars. Conversely, the other types are priced below 20 dollars, making them more affordable.

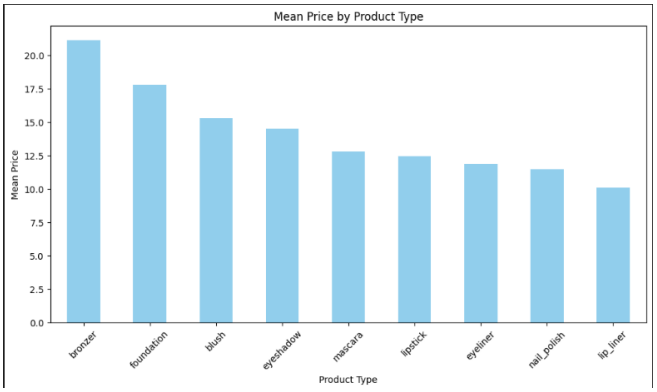

Figure 4. Mean price by product type

4.2 Distribution of product types

The distribution of product types within the dataset was visualized using a pie chart, as illustrated in Figure 5. Based on their occurrences, this graphic offers an interesting overview of different product types. A percentage is used to indicate the percentage of each product type that is represented by each segment of the pie. The dataset reveals that foundation and eyeliner consist equally, with both accounting for 15.9% of the total distribution.

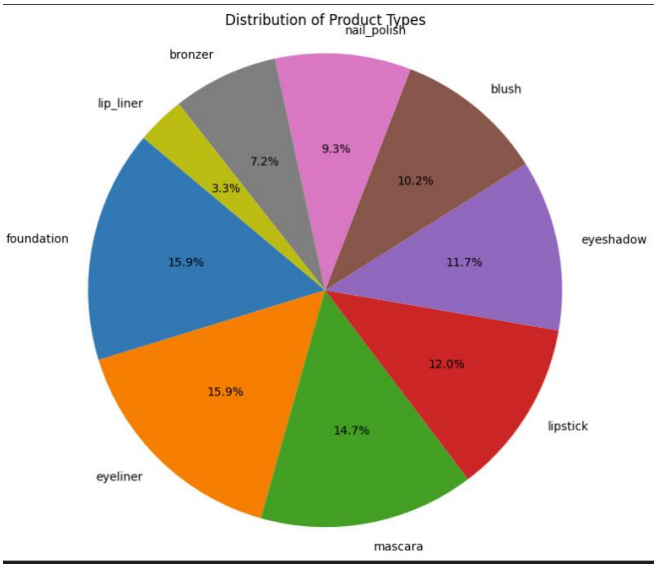

Figure 5. Distribution of Product Types

4.3 Distribution of price, rating, and id

Using the matplotlib and seaborn libraries, subplots were created to examine the dataset's numerical column distributions. The figures in Figure 6 show the histograms for the three chosen columns, "price," "rating," and "id." A visual evaluation of the data distribution patterns is made possible by each subplot, which displays the distribution of the corresponding column.

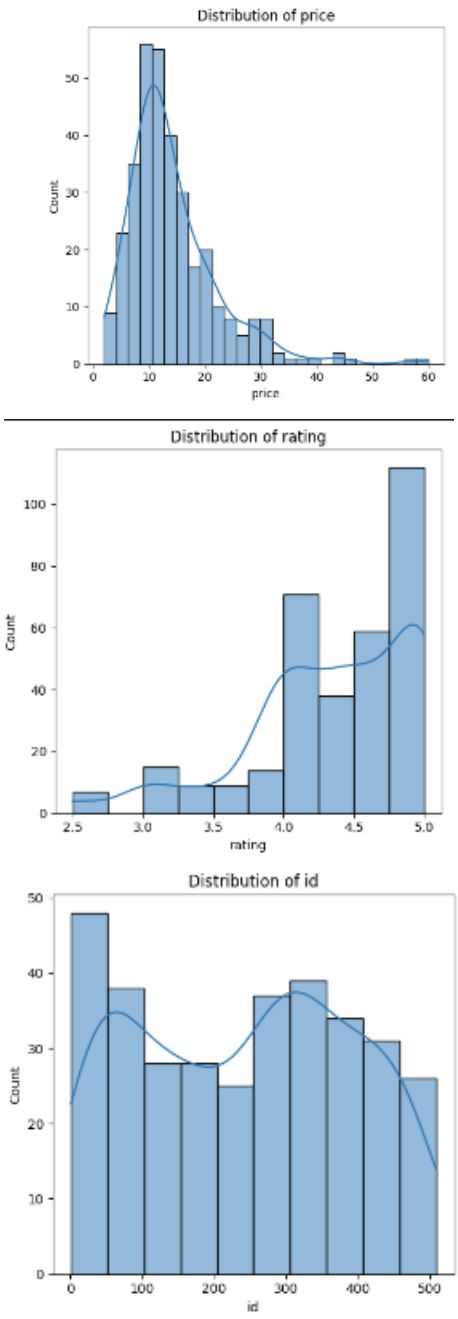

Figure 6. Distribution of price, id and rating

4.4 Average rating by brand and products

Using Seaborn's bar plot capabilities in matplotlib, visualizations were created to assess the average ratings across brands and specific goods. The resulting subplots, as shown in Figure 7 and 8, provide information on the dataset's average ratings.

The average ratings categorized by brand shows that Butter London and Misa have 5 rating level which is most satisfying brands. However, Moov and Mistura are less satisfied brands have less than 3 rating level.

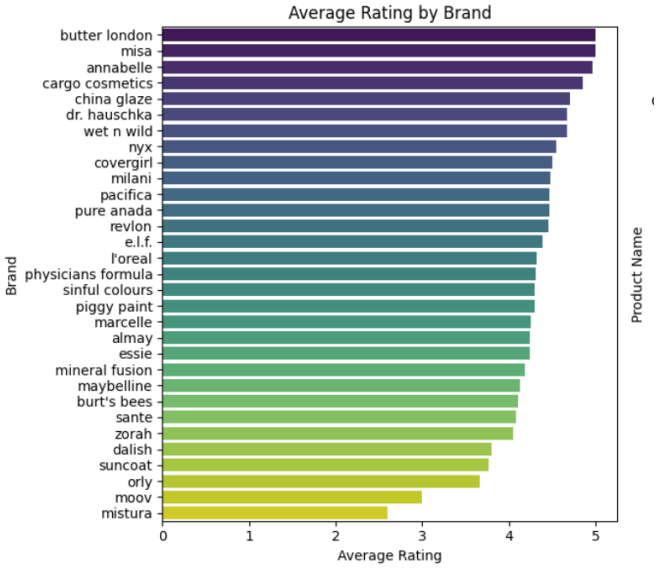

Figure 7. Average rating by brand

As well as the average rating by products graph shows top 10 products which has 5 rating level.

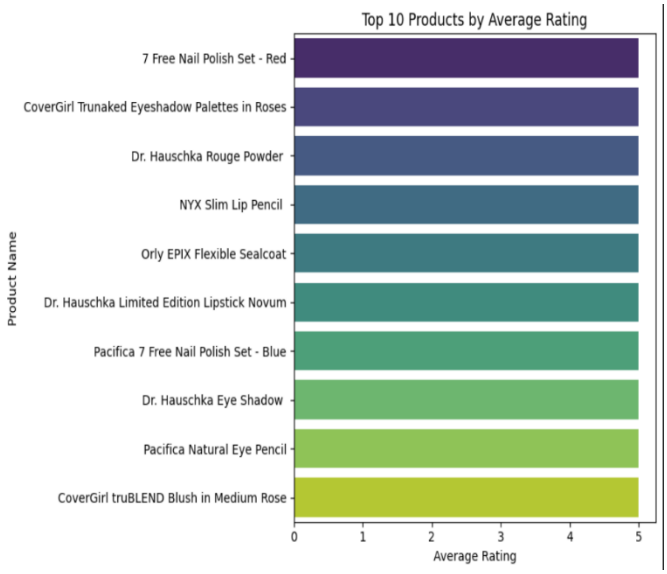

Figure 8. Average rating by products

4.5 Average rating by product type

As below illustrates figure 9 provides a comparative perspective by showing differences in average ratings between several product categories. In general, all products' rating is more than 4 which is very good, and we can say that high quality products and most satisfactory.

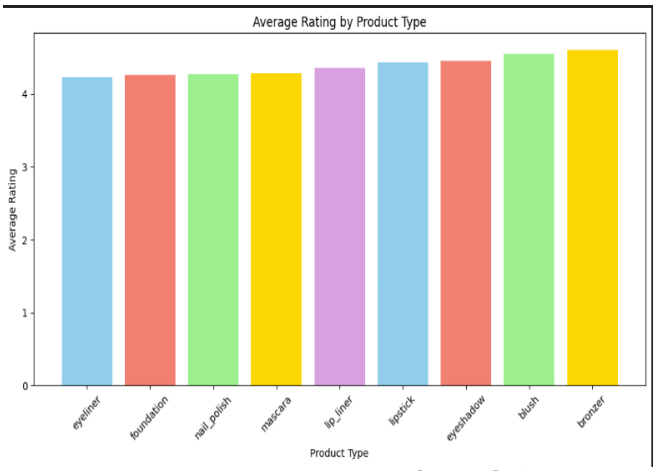

Figure 9. Average rating by product type

4.6 Prices and Ratings by Brand and Products

The association between ratings and prices for various brands and products in the dataset is shown in the below Figure 10 and 11. In Figure 10, the most expensive brand is Physicians Formula, totaling roughly 700 dollars. In contrast, Sinful Colors is the cheapest brand, priced at less than 10 dollars. Additionally, the graph suggests an inverse correlation between price and ratings, indicating that higher prices are associated with lower ratings.

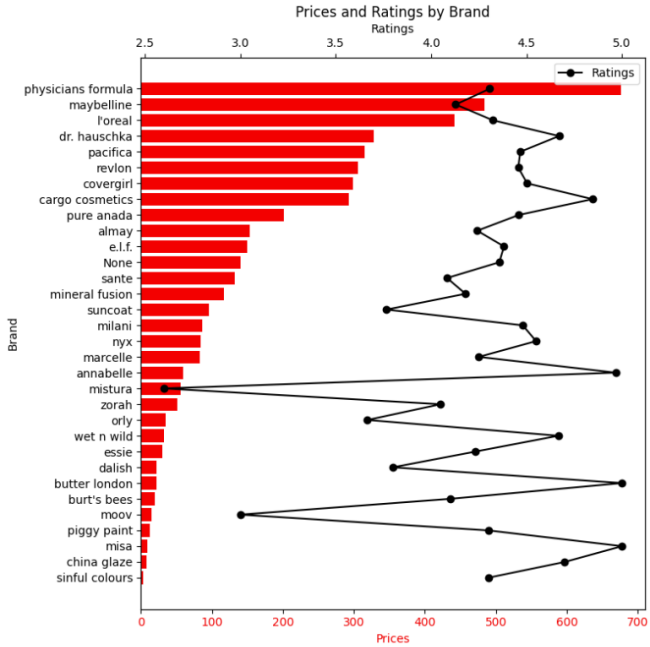

Figure 10. Total Prices and Ratings by brand

In Figure 11, the Pacifica Natural Minerals color palette emerges as the most expensive product with an approximate rating of 5. This suggests a higher quality for this product compared to other expensive goods. However, as previously noted, cheaper products tend to have higher ratings.

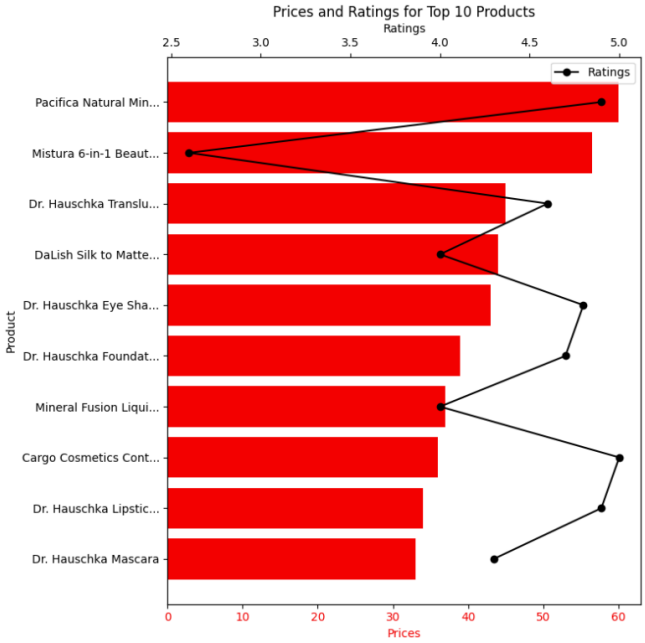

Figure 11. Top 10 products by Total price and ratings

V. MODELLING

After visualizing this dataset has been stored in PostgreSQL database form where it is fetched for future modelling. The table created in PostgreSQL, named as dap_makeup. Following this, exported dataset into CSV file named dap_makeup.csv.

For modelling, built 2 classification models such as Random forest and Decision tree to classify product type based on price, rating and id. As below shown in Table 1, Decision tree classification has better performance than Random forest classification.

Classification models	Accuracy
Random forest	0.85
Decision tree	0.94

Table 1. Accuracy for modelling

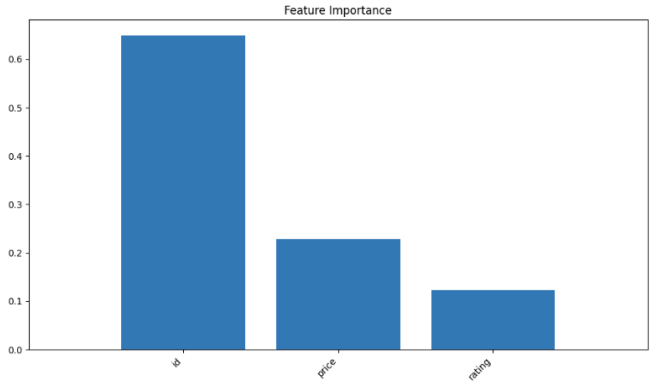

Figure 12. Random forest feature importance

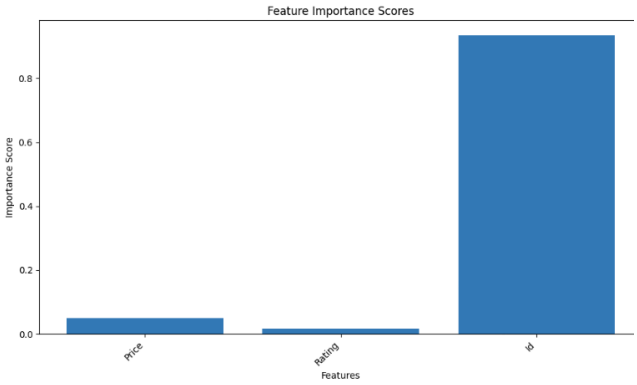

Figure 13. Decision tree feature importance

As above shown Figure 12 and 13 most vital feature of this modelling is id which provides each product details.

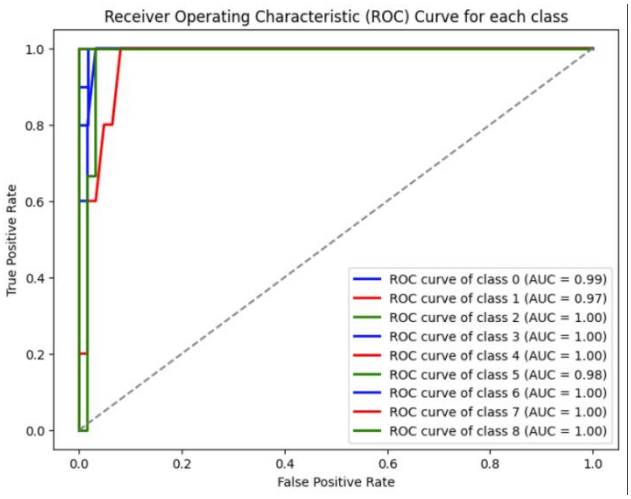

Figure 14. Random forest ROC curve

In Figure 14, the model's capacity to differentiate between classes is illustrated in this gfraphic, where the classifier's overall performance for each class is indicated by the area under the curve(AUC). In summary, having AUC values of between 0.97-1, indicates highly effective and has great cross-class discriminatory power. Nevertheless, it's essential to validate these results on unseen data to ensure the model's generalization and robustness.

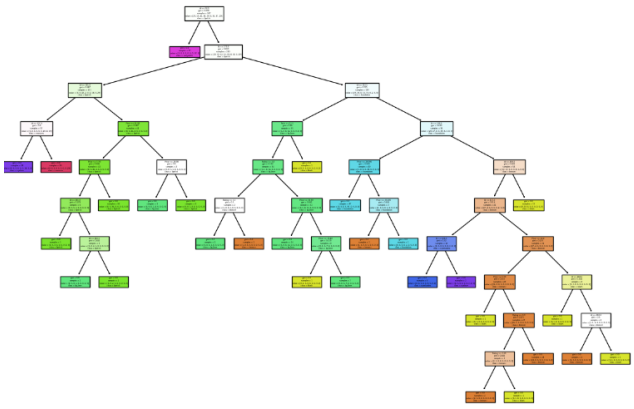

Figure 14. Decision tree unique categories

In Figure 14, it creates a visual representation of a decision tree model, showing how different features influence the prediction of these product types. The created tree shows the model's structure and provides insights into how the algorithm decides which goods to categorize into different classes based on the features.

VI. Conclusion and Future Work

From Figure 7 we can see that Moov and Mistura brands got around 3 rating level which is the least rating. Future implementation in these brands need to create samples based on skin tone or skin type, and creating AI applications for 3D make up, promotion and offer loyalty and discount.

From Figure 8, the products have 5 ratings which is the most satisfying products for customers. Furthermore, the demand is likely to increase, so prepare the stocking well.

From Figures 10 to 11, it's noticeable that pricing appears to impact the ratings, potentially indicating an inverse correlation between the two. To enhance customer satisfaction, I would recommend concentrating on a balanced approach may be helpful to increase customer satisfaction. Ensuring a quality product at a reasonable price might help maintain a positive correlation between pricing and customer ratings. Additionally, actively seeking and integrating customer feedback into product development can provide valuable insights for meeting their expectations. In future work, visualizing cosmetic sales data could be more effective for stakeholders, providing a clearer understanding of trends and patterns.

VII. Bibliography

[1] P. A. Alexis, The Big Beauty Business: A Fundamental Analysis of The Beauty Industry Institute, And Its Economic Value. Omega Publishers, 2019.

[2] L. Karchin and D. Horvath, Cosmetics Marketing: Strategy and Innovation in the Beauty Industry. Bloomsbury Publishing, 2023.

[3] A. Salvador and A. Chisvert, Analysis of Cosmetic Products. Elsevier Science, 2011.

[4] M. Abadi and S. Madden, "HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads," Proceedings of the VLDB Endowment, 2(1), pp. 922-933, 2016.

[5] D. Chen and J. Qin, "Challenges and Opportunities of Big Data Analytics in Healthcare: A Review," Sensors, 18(11), p. 4129, 2018.

[6] H. Wickham, "Tidy Data," Journal of Statistical Software, 59(10), pp. 1-23, 2014.

[7] G. Wilson et al., "Good Enough Practices in Scientific Computing," PLOS Computational Biology, 13(6), p. e1005510, 2017.

[8] W. McKinney, "Python for Data Analysis," O'Reilly Media, 2017.

[9] E. Sheng, L. Chai, and L. Wan, "Cosmetic Product Recommendation System Based on User Preferences," in 2019 IEEE International Conference on Big Data (Big Data), pp. 3525-3534, 2019.

[10] A. Sharma and S. Rai, "Market Basket Analysis for Understanding Customer Behavior: A Case Study of Cosmetic Products," in 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), pp. 242-246, 2018.

[11] A. Gupta and V. Kumar, "Predicting Customer Behavior in Cosmetic Industry Using Machine Learning Techniques," in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-6, 2017.

[12] T. Kim and S. Park, "Personalized Recommendation of Cosmetic Products Using Collaborative Filtering and Deep Learning Techniques," in Journal of Cosmetic Science, 2021.

[13] H. Li et al., "DeepBeauty: A Deep Learning Approach for Predicting Cosmetic Attributes," in 2020 IEEE International Conference on Big Data (Big Data), pp. 4676-4681, 2020.