Yuty
Data Science Group Project

Master thesis for the group project submitted for the degree of Master's in Business Analytics.

Queen Mary of University of London

Supervisors: Prof. Chris Sutton.

22nd August 2022

Submitted by,

Shalini Nayak [200695538]

Shrey Agarwal [210659076]

Neel Raut [210611096]

Anshul Basotia [200663755]

Faran Saeed [200692825]

Table of Contents

Student Declaration Form

We group 5 members,

Shalini Nayak,

Shrey Agarwal,

Neel Raut,

Anshul Basotia,

Faran Saeed,

hereby declare that the work in this Group Project Report is our original work. We have not copied from any other students' work, work of ours submitted elsewhere, or from any other sources except where due reference or acknowledgement is made explicitly in the text, nor has any part been written for us by another person. The individual sections of the report have been written by the author in the section title with no external assistance.

The referenced text has been flagged by:

- Using italic fonts, **and**
- using quotation marks " ", **and**
- explicitly mentioning the source in the text.

22nd August 2022

Date

1. Executive Summary

1.1. Introduction

Today the beauty industry is a multibillion-dollar business globally. Even though the global economy has its own shares of ups and downs, the beauty industry has held its ground strong, as per statistics, the industry was recorded to be $511B in 2021(Figure 1), and it is still increasing with a 4.75% growth rate (Roberts, 2022).

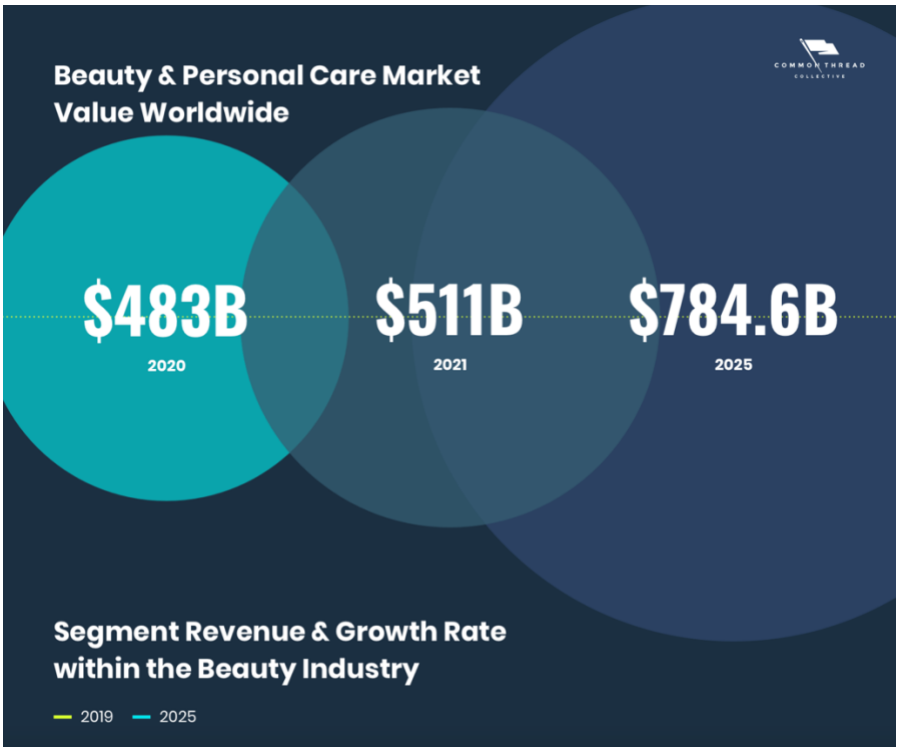


Figure 1: Beauty Industry Revenue. Source:(Roberts, 2022).

Our project Yuty is about one such company which aims at finding the right beauty product for its customers. Yuty was founded by Simi Lindgren, it was previously named YUTYBAZAR, this company aims to match the beauty products that target the niches and not the general population.

Located in the heart of London, with its head office in Covent Garden, *Yuty is strategically well connected to the landscape of London. The company's core focus is to empower all customers regardless of their skin tone, skin types and lifestyles*. The product recommendations to customers are based on their individual needs and choices. These beauty product recommendations are based on scientific methods such as machine learning, and computer vision models. The results of these models are accurate and suitable for the customers.

The goal of scientific methods such as Machine Learning in the beauty industry is done through computer vision which clocks images and processes the images to know about facial features, skin

colour and tones. When it comes to data, *artificial intelligence systems and machine learning algorithms are built to make accurate predictions and suggestions for customers.*

In today's fast pacing world, we all know time is *"time is scarce"*, and customers can't spare a lot of time in their daily life, and this system will help them to get a faster shopping experience and less wastage, which will automatically help the environment and the economy. Yuty's business goal is to tap these time constraints by providing suitable products for its customers in minimum time with the highest accuracy.

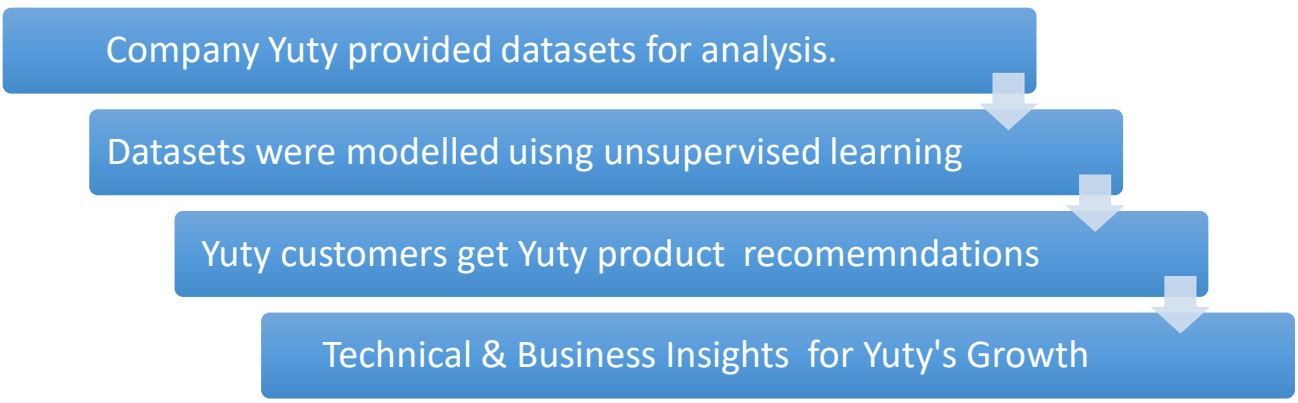
This executive summary discusses the detailed analysis of the results of four unsupervised machine learning models, their accuracy and recommendation outputs, and the overall business insights derived from each step of the analysis, which will aid Yuty in taking business decisions and providing the best products for their customers that fit perfectly and with no waste, thereby maximizing business growth. The technical analysis part has been done using Python language on Jupyter notebook.

We, as a team, were successfully able to implement four models – K-means clustering, Multinomial Naïve Bayes, Random Forest Classifier, and Cosine Similarity. After implementation of all these models, we *concluded that the cosine similarity machine learning model was the best algorithm as it gave the highest prediction accuracy*, which will be discussed in the latter part of the report in detail.

1.2. Business Problem

Our business problem is to find product recommendations for each Yuty customer based on their inputs and data samples using unsupervised machine learning models and provides business analysis on it.

The *business objective is to make a text classification model*. By analysing the dataset, we used an unsupervised learning method with the univariate variable as descriptive statistics as our analysis i.e., ingredients. Below diagram shows the entire flow of our project.



The best recommendation model, as per our modelling and analysis, once implemented by Yuty, would lead to a higher customer satisfaction rate. This is since the products that are recommended to the users won't be random as they will be personalised based on their choices. A higher customer satisfaction rate leads to a higher customer retention rate, which leads to higher sales revenue generated and Yuty building up a loyal customer base.

1.3. Exploratory Data Analysis & Data Pre-processing

Exploratory data analysis is done to understand the dataset and its quality, we were able to grasp the understanding of the functionality to which each dataset was contributing. The objective of EDA is to help analyse the pattern of the dataset with the help of visualisation. There are six CSV files and two JSON files in the dataset, it is very important to have a sense of the connections and mapping of the datasets, this was achieved by schema establishment, which is depicted in the schema diagram in Figure 2 of Point 5 in Yuty Dataset explanations.

After analysis of the variables in the dataset we successfully *implemented the bag of words technique* for information retrieval on the textual data that was provided by Yuty, and we also did an analysis of the missing values in the dataset which was tackled by data pre-processing methods. The exploratory data analysis helped us to improve the quality and strength of data samples provided by Yuty. Data pre-processing, feature engineering, and machine learning methods helped to get better accuracy for the recommendation model.

In the second phase, we dealt with the missing values and other anomalies in the data by removing them as advised by Yuty, as it affected the accuracy of our models, we were successfully able to process them *through robust data pre-processing techniques and feature engineering*.

After dealing with missing data, we used NLP to clean our data and we were able to achieve the desired data, which was suitable for unsupervised learning modelling, which is elaborately mentioned in sub-sections of 6 & 7.

1.4. Methodology Used

After a testing different machine learning models that can be implemented to build a product recommendation system for Yuty customers, we concluded that the four following machine learning techniques would work the best for our business problem:

1. **Cosine Similarity:** It's a widely acclaimed algorithm which gives the best product recommendation by calculating the shortest distance between two vectors. This model works on the concept of vectorisation. The reason for considering this model it **works very effectively**

with small datasets and gives the similarity irrespective of the data samples size. This algorithm achieved an **accuracy of 52.03%.** We do get the highest product recommendations accuracy in the model when seen product wise discussed in the Results section below.

2. **Random Forest:** In Random Forest machine learning technique, the process of voting is done for classification. Here the output of each model is looked at and the highest count of output number is observed as the output for that model. It models based on bagging and feature randomness. The reason for considering this model it is **most flexible, faster to train and works great on small datasets**. And it has provided favourable results in our analysis with an **accuracy of 36.08%.**
3. **Multinomial NB:** This model is used for both continuous and discrete data. Even Real-time applications can be predicted using this simple method. Usually, it is used for large datasets and is scalable to a large extent. The results are compared in our recommendation model using metrics like precision and F1 score. Reason for considering this model because is it a **low complexity model and gives good results when dealing with small datasets**. In this analysis, it has an **accuracy of 30.40%**.
4. **K-means:** This model works on the concept of clustering after vectorizing the text data as ML algorithms cannot process text data. K-means works by grouping data samples into clusters that share the same similarities. Clusters nearest to the mean value which is also called centroid calculates the best product suitable for users. The reason for considering this model is **because it is very popular unsupervised learning, easily adapts to new data samples**. In this analysis K-means has an estimated **accuracy of 21.45%**.

1.5. Results

The main aim of the project was to build a recommendation system that recommends top three Yuty products for each of the Yuty customers. With this aim in mind, we were able to build four different models of recommendation systems using four different machine learning techniques – K-means, Multinomial Naïve Bayes, Random Forest Classifier, and Cosine Similarity.

Yuty provided us with eight different data files of CSV and JSON origin. The first step that we took regarding our project was understanding these datasets and doing some initial exploratory data analysis on them. We identified the missing values in the datasets and after talking to our mentors, we decided to drop those missing values to build our recommendation system. After that, we cleaned the dataset and created features. These features help us to decide the efficiency of the recommendation system.

Every step that we have done and analysed concludes a business insight that can help Yuty to grow their business.

Once our features were built, we implemented the four machine learning models that are mentioned above. We concluded that the ***cosine similarity machine learning technique gave us the best recommendations as the accuracy for this model was 52.03% among the other three models*** which couldn't perform very well due to lack of richness in data quality, other than random forest which gave the second-best accuracy for product recommendation**.** Figure 2 shows the accuracy levels of all the four different recommendation models we built is given below:

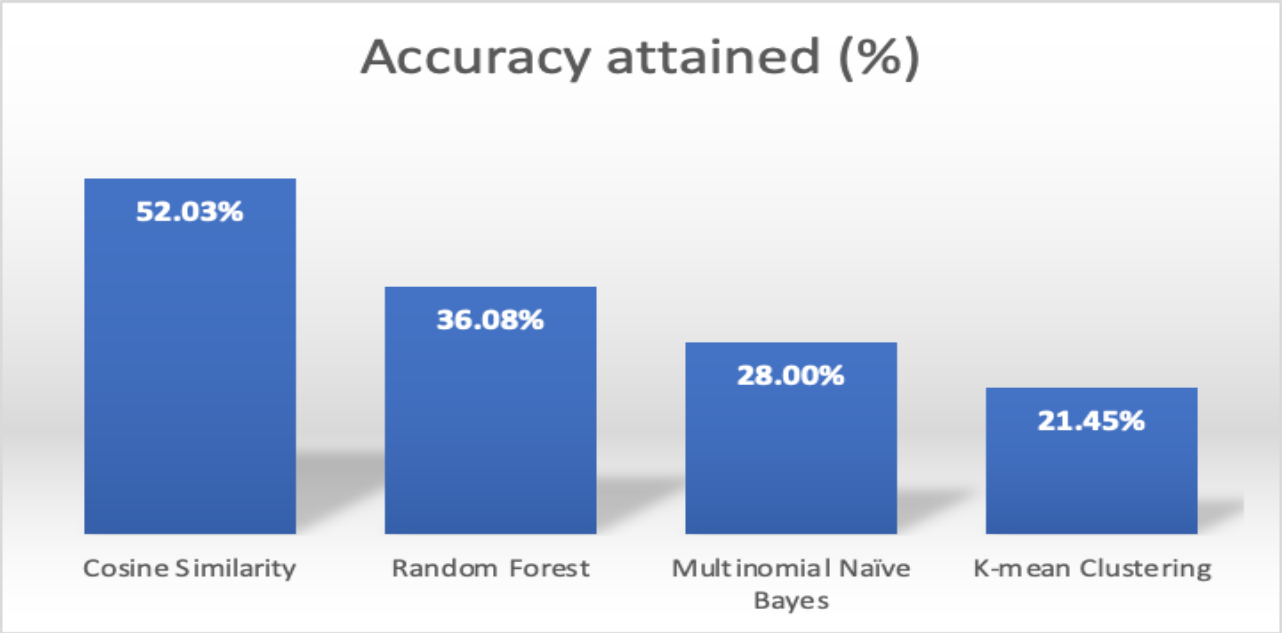


Figure 2: Accuracy Chart

A detailed explanation of all the four models that were implemented are discussed as follows:

1.5.1. Cosine Similarity - CS

Cosine similarity works on the concept of vectorisation. Text data's of "ingredients obtained from customer's quiz" & "ingredients present in the product catalogue data" were both converted into two vectors and the least distance between then was calculated to obtain most recommended product.

CS was able to give a very good accuracy in this project analysis compared to the 3 other models that was used.

```
Accuracy of Cosine Similarity Algorithm:  52.03440000000001
```

CS worked best for small algorithms and the process of vectorisation proved to give very effective solutions.

Elaborating on results, we see that customer with user id 2.0 (highlighted in red) is getting its top three recommended products according to the quiz they have answered. We see Product 1 which is, "C The Different Treatment Eye Mask" gives an accuracy of 52.94% which is very good.

Meaning that this Yuty product "C The Different Treatment Eye Mask" can solve 52.94% of the user's problem with user id 2.0. We can recommend this product to the customer for better effectiveness as well as also recommend a corresponding co-purchased product for maximum effectiveness which is "Vitamin C Prepping Tonic" with an second highest accuracy of 36.49%.

	user_id	top_3_products_recommended	top_3_products_recommended_percentage
0	2.0	(C The Difference Treatment Eye Masks, Vitamin C Prepping Tonic, DOYENNE! MIRACLE FACE SERUM)	(52.9412, 32.1678, 4.2813)

Figure: Output which shows "C The Difference Treatment Eye Masks" solves ~53% of the customers problems.

1.5.2. K-means Model

With the process of clustering our data was made into 381 clusters and a mean value was calculated which is known as centroid. And the value closest to the centroid was measured which gave the accuracy of model with product recommendations. We fit one product to one cluster which have many ingredients, while inferencing we convert user input answers to their respective ingredients and calculate the cluster with minimum distance.

Being one of the most popular models in the unsupervised machine learning realm, K-means was not able to perform too well and the accuracy of it was moderate. With our analysis we found that it was because of the data quality and quantity. We were able to achieve an accuracy of 21.4 % for product recommendations for customers.

```
Accuracy of K-means Model:  21.455369230769243
```

Even if the overall accuracy was low, it still was able to give good recommendations for Yuty customers which is 53% for user id 2.0. Figure 3: shows output of K-means model which follows the concept of clustering and could be very effective if subjected under more advanced datasets and research.

	user_id	top_3_products_recommended	top_3_products_recommended_percentage
0	2.0	(C The Difference Treatment Eye Masks, Vitamin C Prepping Tonic, DOYENNE! MIRACLE FACE SERUM)	(52.9412, 32.1678, 4.2813)

Figure 3: Output which shows "C The Difference Treatment Eye Masks" solves ~53% of the customers problems.

1.5.3. Multinomial NB

Multinomial Naive Bayes is frequently used for the classification model. We have approached this model from different angles using various NLP features. We were able to implement four different features for this technique. The following are the results that we obtained:

- The first method that we used is vectorization because it is the easiest to implement and usually gives the best results for small datasets.

Vectorization score:

```
Accuracy of Multinomial NB for Count Vectors:  0.28
Accuracy of Multinomial NB for WordLevel TF-IDF:  0.2
```

```
F1 score for Count Vectors:  0.31809523809523804
F1 score for WordLevel TF-IDF:  0.33333333333333337
```

The results show better accuracy and almost the same F1 score for count vectors as compared to TF-IDF which is surprising because usually TF-IDF vectors provide better results in these situations.

- The second method that we used is the GloVe algorithm. Below is the accuracy for it, here F1 score doesn't apply.

```
Accuracy of Multinomial NB for GloVe:  0.16
```

The accuracy here is low: 16%, which is not good enough, but we can't expect too much accuracy from this model since GloVe tries to match words with similar meanings and in our case each ingredient is unique.

- Next, is the text-based features. These features highly depend on the type of text data we have as input. After analysing the text, we create features that make the most sense. These are the results.

```
accuracy = train_model(MultinomialNB(), train_x1, train_y1, valid_x1)
print ("Accuracy of Multinomial NB for text based features: ", accuracy)
```

```
Accuracy of Multinomial NB for text based features:  0.2
```

```
F1_score of Multinomial NB for text based features:  0.25160784313725487
```

The results show that count vectors act as the best features for training a Multinomial NB model.

The accuracy attained for Multinomial NB is 28.00%.

Figure 4 depicts the four features that we have used for Multinomial NB with their F1 and Accuracy scores:

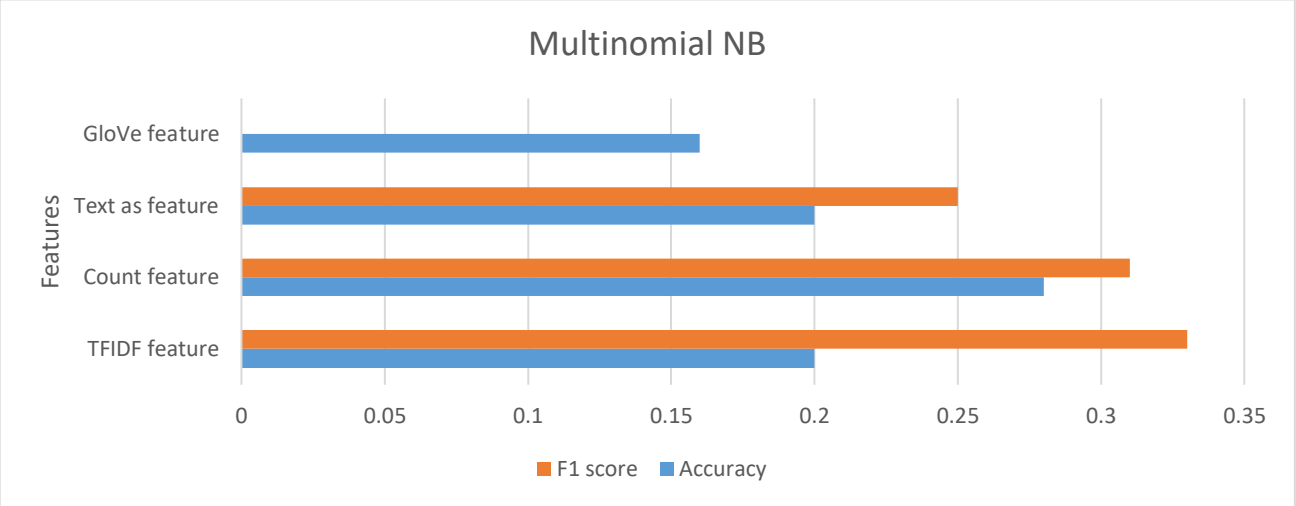


Figure 4:Results of Multinomial NB model

1.5.4. Random Forest

We have used the same features that we used in Multinomial NB for finding the best model here.

Vectorization score:

```
Accuracy of RandomForestClassifier for Count Vectors:  0.28
Accuracy of RandomForestClassifier for WordLevel TF-IDF:  0.36
```

```
F1 score for Count Vectors:  0.3928888888888889
F1 score for WordLevel TF-IDF:  0.4173109243697479
```

This time TF-IDF vectors show better results with 16% accuracy. These results are better than those that we received previously with the Multinomial model, so we are on the right track. Next is GloVe

```
Accuracy of Random Forest for GloVe:  0.16
```

Again, the accuracy is low here. So, we will move ahead with text-based features:

```
Accuracy of RandomForestClassifier for text based features:  0.12
```

```
F1 score of RandomForestClassifier for text based features:  0.14666666666666664
```

After running all the methods, TF-IDF vectors had the best accuracy. **So, the best models came down to be cosine similarity, an unsupervised learning model, and Random Forest, a supervised learning model.** Both these models take a different approach to reach our goal i.e., to match the customers to the products that are effective on their skin. Refer to Figure 2 for the accuracy of all the four models that we have implemented.

The accuracy attained with the TF-IDF feature was the best. ***The accuracy attained for Random Forest is 36.08%***. Figure 5 depicts the four features that we have used for Random Forest with their F1 and Accuracy scores:

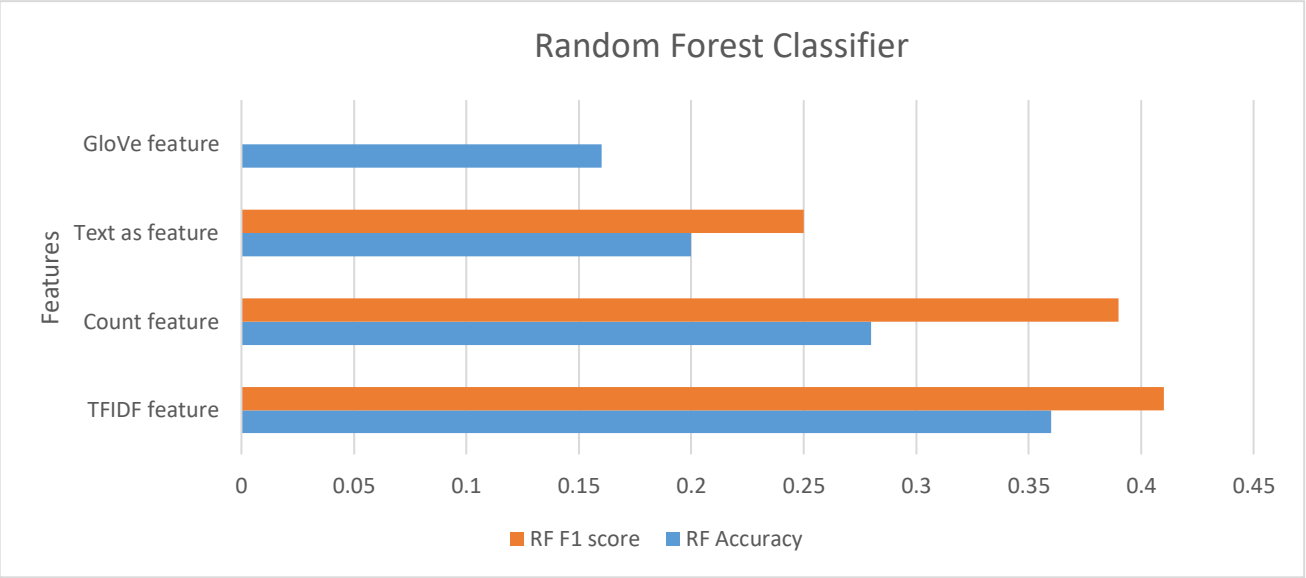


Figure 5: Results of Random Forest model

1.6. Limitations

As the project progressed, we found some challenges while implementing that would hinder the performance of our four models. These limitations are mostly related to the dataset that is provided to us. Some of the limitations and challenges we faced are:

- The datasets were small, which limited us to the use of machine learning methodologies and options. As advanced machine learning model become inefficient when the data is very small.

- No labelling was provided for the products resulting in inaccurate labelling since we had to manually label the data which affected accuracy.
- Too many anomalies in the textual data since the data was extracted from web sources.
- Less user data. Information taken from customers were not too elaborate as there were only fifteen questions Figure 6, if there was more data it could have broadened the scope of better learning by ML models.

question_id	question_text
21	How does your skin feel after cleansing?
23	What are your 3 main skin concerns?
24	Which of the following skincare products do you use on a daily basis?
25	How many hours do you spend each day in direct sunlight?
26	Based on the colour of your skin and how it reacts to sun exposure, how would you define your skin tone?
28	Are you aware of any allergies?
31	Skin is affected by external lifestyle factors, select those that apply
32	How many glasses of water do you drink on a daily basis?
33	How many hours of sleep do you get on average every night?
37	In order to recommend products individual to you, are you going through one of these key life stages?
38	What are your personal skincare preferences
44	Do you use any products containing retinol?
45	When you wear makeup, how frequently do you remove it after wearing?
46	Your skin is affected by where you live, so which city are you in?
47	How many hours do you spend each day in front of a digital device?

Figure 6: Customer quizz by Yuty.

Even though there were some limitations, we were able to achieve an effective model by our Cosine Similarity algorithm which will help serve Yuty customers in product recommendations. Also, there is a lot of scope for future improvements which we will talk about them in the upcoming section.

1.7. Business Insights, Future Scope, & Considerations

Yuty has a huge prospect in the beauty industry and its venture into product recommendation using Artificial Intelligence and Machine Learning has already set it apart in terms of uniqueness from the other companies. We know ***beauty industry is a Read Ocean with many competitors 8igure 7***. But Yuty using Data Science and ML in their recommendation system has very effectively make the ***Red Ocean into a Blue Ocean***. Here are some insights from our business and technical analysis which could help Yuty grow:

RED OCEAN STRATEGY	BLUE OCEAN STRATEGY
Compete in **existing** market space	Create **uncontested** market space
Beat the competition	Make the competition **irrelevant**
Exploit **existing** demand	Create and capture **new** demand
Make the value-cost trade-off	**Break** the value-cost trade-off
Align the whole system of a firm's activities with its **strategic choice of differentiation or low cost**	Align the whole system of a firm's activities in **pursuit of differentiation and low cost**

Figure 7: Red & Blue Ocean Strategies.

Source: (Kim and Mauborgne, 2015)

- **Business Insights:**

  Yuty goal is to reduce waste so implementing our analysis (Cosine Similarity & Random Forest) which gives product recommendation in real-time solves the ***sampling waste problem*** - reducing manufacturing and inventory cost thereby reducing wastage. Secondly, ***machine learning works great on more data***, and our analysis is a testimony that we can get even better results if ***Yuty improves the data collection quality***. Thirdly, customers like the feeling of personalised product recommendation, Yuty along with our technical and business analysis should work on ***marketing the personalised product recommendation strategies*** which will boost customer purchases significantly also recommending complementary products. Finally, and most important Yuty should try to ***improve the quality of data collection by collecting data like active and inactive data is very helpful to get right features trained.*** A ***100% data understanding is needed to iterate and improve data collection which Yuty should work on*** for best accuracy and better recommendations for its customers.

- **Future Business Enhancements:**

  This could include better ***responsive UI for both web and app for better customer experience***. ***Live product suggestions*** along with the main products works very well in revenue and trust making. Loyalty points and freebies like free scans and discounts, are very effective when

implemented with digital marketing and advertisements. This could be achieved by proper customer research and ***advanced advertising strategies like from Facebook and Instagram*** targeted ads which has billions of users and their data which can be massively useful to ***understanding their customers***. Yuty should also think to ***capture big markets*** and improve its ***logistics service*** for good customer experience along with good product recommendations. This has been discussed in detail in point 14.2 .

- **Benefits Machine Leaning:**

  Implement machine learning algorithms in business can help Yuty reduce manual work, automate many processes which will give time to focus on lucrative decision making. ML when implemented with proper and high-quality data will give best recommendations along with better insights and cognitive capabilities. Working with Machine Learning will have far less error rate than humans which will increase speed, efficiency and save money. This has been discussed in detail in point 14.3.

- **Technical Research Advancements:**

  Yuty should also try ***supervised learning and labelled data***. For this the best suited models could be XGboost, Random Forest etc. ***Deep Learning and Computer vision*** can be used to aid in high quality data collection. **Neural networks** like CNN and BERT are highly recommended for advanced results. This has been discussed in detail in point 14.4.

- **Commercial Considerations:**

  Implementation of a good recommendation system could lead to a higher ***customer satisfaction rate and higher customer retention rate.*** This could lead to ***tapping into new geographical markets and higher profit margins due to low production costs.*** This has been discussed in detail in point 15.

1.8. Conclusion

The motivation behind this report is to increase the efficiency of predictions through Machine Learning models for the customers landing on Yuty's website. The purpose of this is just not to increase customer base by suggesting right products to customers, but also to use the data when doing research and development of new products which would in return reduce wastage and help the environment. At the end of the paper, after extensive research on the industry and consideration of the accuracy of the used Machine Learning models, some suggestions are presented to the company which can benefit the company by increasing their turnover and adding to the efficiency in their processes. The paper is beautifully structured, and the learnings can be applied in other similar industries as well.

1.9. Project Distribution & Team Delivery

The project followed the proper structure and standards of a business project. We have tried to follow an AGILE methodology as we proceeded with the project. The feedback and inputs from each step were iterated and done again.

Faran Saeed and **Anshul Basotia**, who have a vast background in business and economics, played the part of business analysts focusing on industry analysis of the beauty industry, how machine learning is a better approach than the traditional method and its success rate. They have also explored and analysed the datasets provided by Yuty for useful data insights.

Anshul Basotia has covered part of project description including competitor analysis and company description of Yuty. Limitations done by him has greatly helped for analysis.

Faran Saeed have given great insights on how we can recommend Yuty for improving their business and consider new approaches to increase client base.

Neel Raut, with a background in Computer Science, has worked in the database environment. He is a median between tech and business and has overseen the project synchronisation. Hence, he has worked with the team to help understand the data in its entirety and how the different data files relate to each other. He has also worked on Overall result analysis and getting business conclusions regarding our implementation.

Shalini Nayak and **Shrey Agarwal** handled the technical part, which was more of a Data Scientist role, as well as guiding the whole project as a subject matter expert (SME) owing to their experience in Data Science projects.

Shrey Agarwal, with his background in Data Science, had worked on unsupervised learning before and has worked on two unsupervised models that are Navier Bayer's and Random Forest along with feature engineering in this project. And has also contributed to data pre-processing techniques for improving data quality.

Shalini Nayak, with her educational background in Data Science, has worked on Cosine Similarity and K-means model, full literature review as well as on the business analysis, suggestions on future analysis on how research on data analysis and machine learning can bring to the company Yuty.

Shalini Nayak her previous work experience in analytics and IT, played the role of **Team Lead**.

The whole project was executed in a proper structure with proper business meetings, discussions and brainstorming of analysis of all phrases. Everyone pitched their ideas, and with all proper

consideration, tasks were allotted for writing. With each progress in writing and analysis, there was a proper quality check done. Each team member has revised and cross-validated each other's work under the supervision of the team leader. Good quality and proper time management were followed by the whole team.

The table below shows the individual written contributions of our team members for an individual report.

Name	Table of contents	Topics
Anshul Basotia	3 8, 8.1, 8.2 11 15	Project Description Exploratory Data Analysis Limitations & Challenges Commercial Considerations
Neel Raut	5 10	Dataset Explanation Results – Data Analysis
Faran Saeed	2 8, 8.3 12 13	Introduction Exploratory Data Analysis Conclusion Recommendations to Yuty
Shalini Nayak	4 9, 9.1, 9.2 14	Literature Review Methodology – K-means, Cosine Similarity Business Insights, Suggestions on Further Analysis & Future Business Enhancements
Shrey Agarwal	6 7 9, 9.2, 9.3	Data Pre-processing Feature Engineering Methodology – NB, Random Forest

2. Introduction (By Faran Saeed)

It may be surprising, but makeup has been around for over 5000 years now. What started as mineral pigments used to follow certain rituals, has been transformed into an everyday essential for us today. Most of the modern makeup that we see today has got its fame in the early 20th century. "*The type of lipstick tube we are familiar with today was invented in 1915. Pancake makeup was invented in 1935 to help movie stars look more attractive on film"* (Evergreen Beauty College, 2012)

2.1. Industry Analysis

Thus, just a century later, makeup has emerged into being one of the biggest industries in the modern world. The country with the biggest makeup industry and purchasing power, United States of America has experienced a rise of almost 50% spending in the overall industry just over the past decade. "*In 2020, the average annual expenditure on cosmetics, perfume, and bath preparation products amounted to approximately 199 dollars per consumer unit in the United States"* (Statista Research Department, 2022). The bar chart in figure 8 as follows visualises the average annual expenditures for cosmetics for the US in the past 10 years.

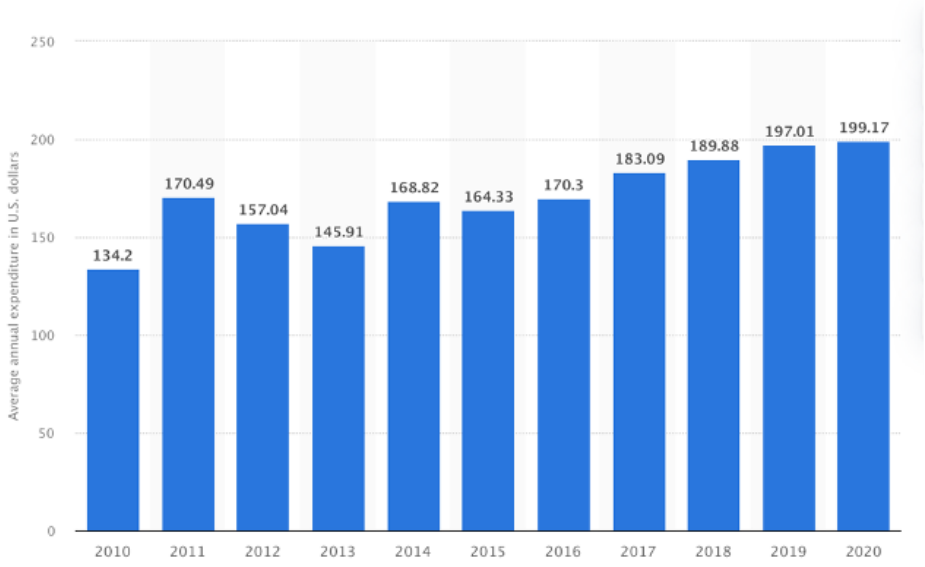


Figure 8: Source:(Statista Research Department, 2022a).

The above visualisation explains how the overall basket size for the makeup industry has increased over the years and it is safe assume that other countries are following the same trend. United Kingdom is also one of the largest cosmetic consumers in the world. *"In 2019, spending on personal care reached approximately 33 billion British pounds"* (Statista Research Department, 2022a). The industry which started from a few products is now divided into numerous categories with products niche down to serve a diverse client base. We now see makeup for all skin tones, shampoo for all hair types and

thus with such a variety of options, it is sometimes difficult to choose the perfect type of product according to personal requirements.

2.2. ML Influence

With the increasing demand and complexities in the beauty industry, it is just natural to introduce Artificial Intelligence and Machine Learning into the industry to optimise and make the decision making more efficient. Integrating ML into the makeup industry does a lot more than we may think. Using the right technology, it is quite easy to get enough statistical data based on which it is extremely easy *to help people look more attractive by making the right product suggestions.*

The main use of ML in the industry is through computer vision which captures images and processes them to learn about facial features and then on. On the basis of the data, artificial systems can be built that can in return make accurate predictions and suggestions to customers. This will not only save a lot of time but will also aids towards less wastage helping the economy as well as the environment.

Integrating Machine Learning into the industry can also aid in the development of new formulae based on extensive data and feedback collected over time. User experience can tell if or not a specific product is working and then certain tweaks in the formation can ensure that the product is refined for better performance.

Figure 9: How computer vision works. Source (EnterpriseAsia, 2019)

The image in figure 9 is a perfect depiction of how the computer vision would collect data from a facial image and then give the closest recommendations. Matching the correct skin tone with the foundation, and lip colour to the best lipsticks available and accordingly finding the perfect fit for the customers.

2.3. The traditional vs ML method

Innovation always stems from a need to improve a process to solve a problem. One needs to improve a product or a service to make it efficient and add more value to it. It can be saving of time, energy, and money.

A research paper by Hema Sekhar and her co-authors talks about how ML methods differentiate from the traditional methods when predicting certain medical factors like patient-reported outcomes, response to a specific treatment or survival rate. The paper even though explains the significance of these algorithms, it also highlights the fact that both, traditional and ML methods have their own significance and can perform better according to the situation (Rajula et al., 2020). Henrik Nyman in his article beautifully compares the two methods and talks about the significance of each model as follows:

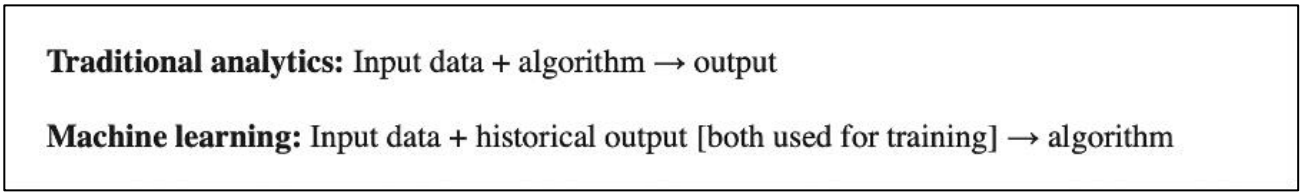


Figure 10: Henrik's Comparison of approaches. Source (Nyman, n.d.)

In figure 10 he says that one key factor that sets the ML methods apart from the traditional ones is its ability to learn from examples. Thus, in line with the previous article mentioned, Henrik also claims that Machine Learning methods are more useful when there are very complex relationships in the data or when the data is in a non-standardised format i.e., images, videos or textual data (Nyman, n.d.).

Similarly, in the makeup industry, products are reshaped so that they can maximise the results and at the same time, save costs and be environment friendly. Apart from these generic goals, there always exists a thin line between how a brand improves a product and what the general target market wants from the product. This is one of the problems which can be eliminated by introducing ML and AI algorithms as these can be used in a variety of ways to filter customer data and give concise and effective suggestions.

Avon’s mascara was one of the developments made by analysing the requirements and needs of the customers by their reviews on social media platforms. The Machine Learning Algorithm then pinpointed and ranked customers comments by various stages of filtration and analysed what exactly were the customers looking for in a mascara (Analytics Steps, n.d.). This is a classic example of how these ML and AI models can be used and integrated smartly to develop just the right products which in return can scale any business and increase profits.

2.4. Summary of the project

This paper aims to provide the best recommendations to Yuty customers based on the answers to a quiz they have taken, using unsupervised models. We have used four models on the Yuty datasets provided. The main purpose of choosing an unsupervised model is to be able to give the best recommendation to customers. In this project, we have recommended the top 3 products for each user who have taken the quiz. Upon rigorous modelling and analysis, we have concluded that the Cosine Similarity Algorithm provides the most suitable algorithm which provides the best accuracy within the boundary of values the Yuty datasets had to offer.

Our analysis is also supported by a very acclaimed research papers, it has proven to be the best algorithm. The purpose of the group report is to analyse and discuss unsupervised learning algorithms used for product recommendations. The report firstly talks about the makeup industry and then the Yuty brand history. Then we dwell into the data analysis and exploration. Following this, the discussion of the methodologies used, in increasing order of accuracy are discussed, and finally, a detailed discussion on how our analysis can help the company Yuty by delivering meaningful business insights to help, aim and attract customers by recommending them products from which benefits thereby attracting more sales and a loyal customer base.

3. Project Description (By Anshul Basotia)

3.1. Aim of the project

The main goal of this project is to get the precise recommended list of product id from the product catalogue for each customer based on their responses when they attempt the quiz on the company's website and get recommended products based on their answers. The project delivers different technical insights into the choice of solution alongside the code, it also considers the limitations of the solutions from the insights. This project also gives a detailed report regarding the eight data files given to us by Yuty, the report also includes data pre-processing, data cleaning, natural language processing, merging, and creating structure. The project deliverables help to produce results of the process which is being followed by us as a group.

3.2. Company Description (Yuty)

YUTY was founded in the year 2020 by Simi Lindgren. Yuty is an artificial intelligence-driven conscious beauty brand. The product recommendations to the customer are based on science. Yuty focuses (Figure 11) to match the products which are effective for one individual and not the mass population. *The main aim of Yuty is to enable the customers to get products that are effective for them regardless of their race, genetics, environment, and preferences.* The artificial intelligence technology built by Yuty is exclusive to help the customers recognize the products which suit them perfectly and there is no trial and error in trying different products which will result in zero wastage by product sampling. (www.yuty.me,n.d.)

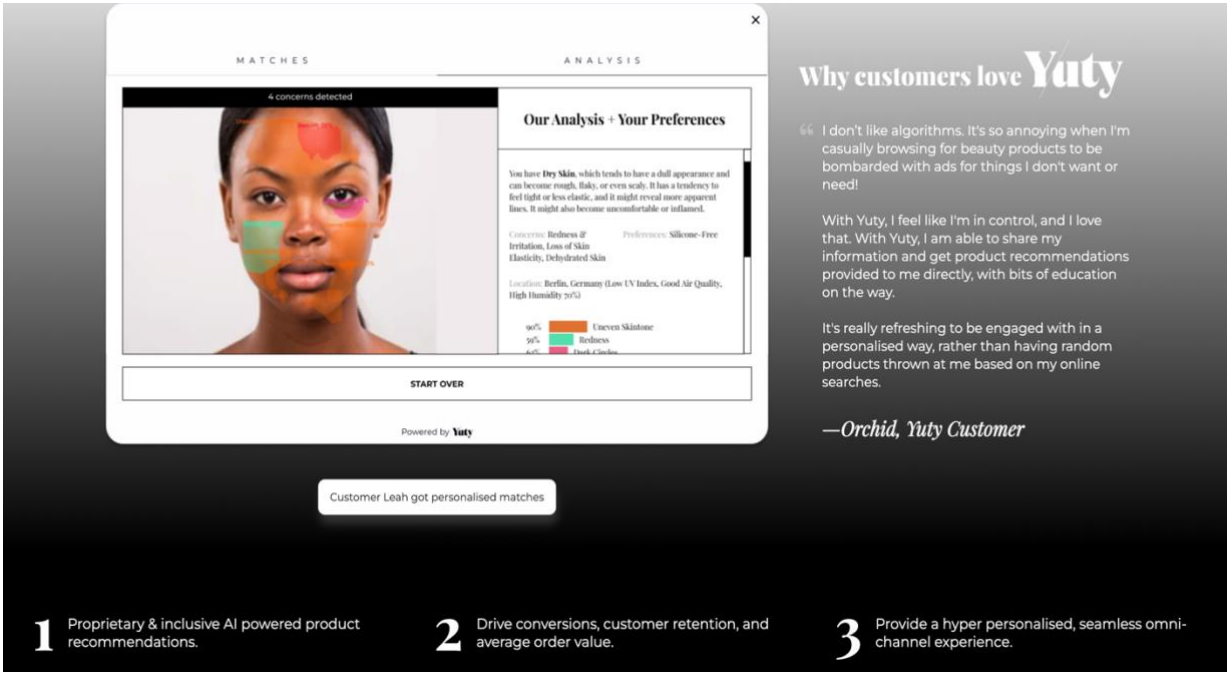


Figure 11:Yuty quiz page and website. Source: (www.yuty.me,n.d.)

The company has three steps to help the customer get the best product, which is required for them, everyone gets recommended various products based on how they answer the quiz, which is the first step for the customer to follow when they want the product from the company's website. The products recommended to the customers in the second step will be of two types i.e., one will be the best product irrespective of the price, and the second product recommended would be a cheaper one, then the customers can select the products based on their likings. The last step would be selecting the product and checkout. The customers can also add the other products to their wish list if they want.

Yuty uses machine learning, computer vision models, deep learning, datasets of diverse photo types, and different skin types to get precise accuracy of suitable products to the customers.

3.3. Competitors using Machine Learning & Data Analytics

Amidst the global pandemic, people do not always prefer to step out of their homes specially to shop for beauty products in store, since there is always a fear of different people trying on the same testers for makeup products. Therefore, they usually prefer shopping online and sticking to the same products they use, making recurring purchases.

However, today, with the advent of the latest technologies like AI and Augmented Reality, the beauty industry has flourished exponentially. ***They have been able to create a massive buzz around a complete personalised approach for everyone***, helping them buy their favourite makeup products with ease. In this way, even loyal customers are willing to try on new products and brands altogether, without having to step out of their comfort zones.

Selfridges is a well-known departmental designer store which offers many segments of shopping for customers, beauty skin care being one of them. There are world-class brands like Charlotte Tilbury, Yves Saint Laurent, Hermes Beauty and so on. Selfridges is one of the stores which have started creating a personalised approach for every customer by taking advantage of technologies like augmented reality and artificial intelligence. (www.selfridges.com, n.d.)

The technology used by multiple companies is to give a more personalised experience on par as it is a gamble to order online but the Covid-19 pandemic changed the methods which companies followed.

Perfect Corp is a company for AI and AR beauty technological solutions(Figure 12) and their main objective is assisting brands in their virtual try-on and restricting over-consumption with the help of customized beauty experiences for each individual and giving them precise product suggestions. Perfect Corp also focuses on discovering new ways for customers to interact and simultaneously lowering the carbon footprint by rethinking the consumer's shopping journey. (Sophie Smith, 2022)

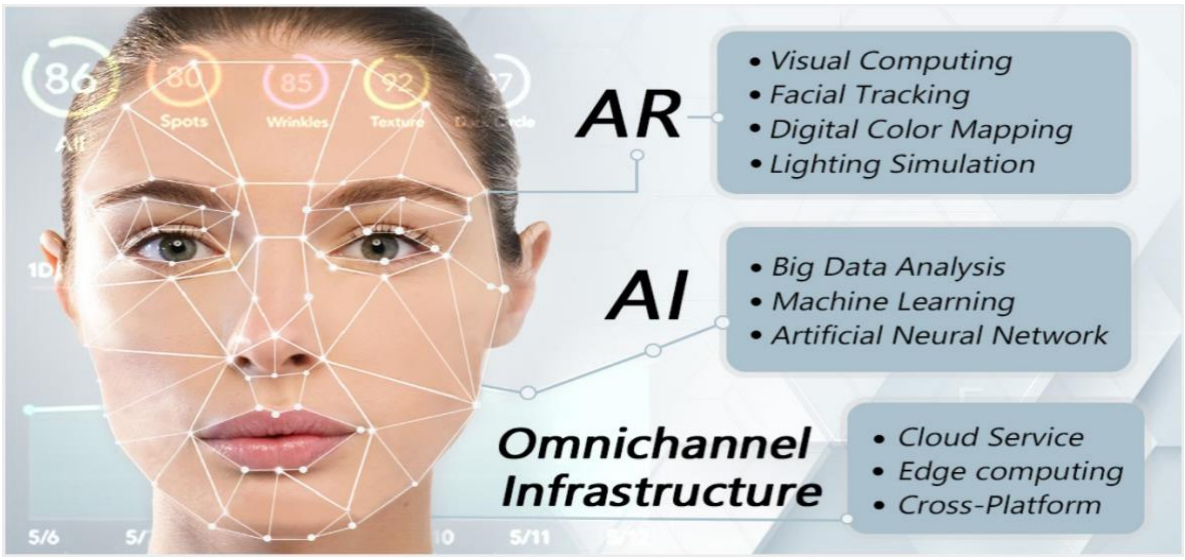


Figure 12: How AI detects facial features. Source: (www.perfectcorp.com, n.d.)

Estée Lauder has collaborated with Perfect Corp for its AI and AR technologies; Perfect Corp is considered one of the world's biggest beauty tech solutions suppliers having more than 800 million downloads globally. They have supplied L'Oréal Paris and Estée Lauder with their technologies of Augmented reality and Artificial intelligence. (GreyB,2020)

In today's world, there are major optimisations taking place by using Artificial intelligence. Companies like Estée Lauder have utilised AI and AR technologies for virtual makeup solutions. They provide this technology to the customer in-store as well as online. Estée Lauder noticed an increase in conversion rate with the help of its technology of Lip virtual Try-on.

This is a particularly good example of our competitor analysis for "product recommendations" on how they are using Artificial Intelligence for business growth and providing personalised solutions to customers providing value, service and quality although machine learning and artificial intelligence techniques. Studying their technological and business advancements can be a profound knowledge repository for Yuty to excel in providing effective as well as personalised solutions to its customers.

4. Literature Review (By Shalini Nayak)

4.1. Background

As our business problem is finding the right products for customers by data analysis and machine learning by their inputs which are in the form of text answers (data samples) from a quiz response from the Yuty website.

Recommendation systems are systems which uses machine learning algorithms, supervised or unsupervised for text analysis and product recommendation to customers based in the current input data or past data.

So, what we are attempting to create in this project is a personalised recommendation model for Yuty which provides best accuracy. The recommendation model will thereby be used by Yuty to recommend products to their customers which can solve their maximum problems. To achieve this, we have used unsupervised machine learning algorithms for product recommendations as we have unlabelled and unique data.

4.2. Product Recommendation Models

Definition: Recommendation models are information processing and filtering algorithms which learns and processes customer's or user's likes, dislikes and preferences and predicts products for them using machine learning and artificial intelligence Figure 13.

Usage: Our business aim is to process the data provided by company Yuty of its customers and predict the best recommendations for them using unsupervised machine learning models. There are many popular machine learning models, and after our analysis, we have used four unsupervised machine learning models that are Cosine Similarity, K-means, Random Forest, and Navier Bayes which were used in this project and business analysis which are discussed section **4.4**.

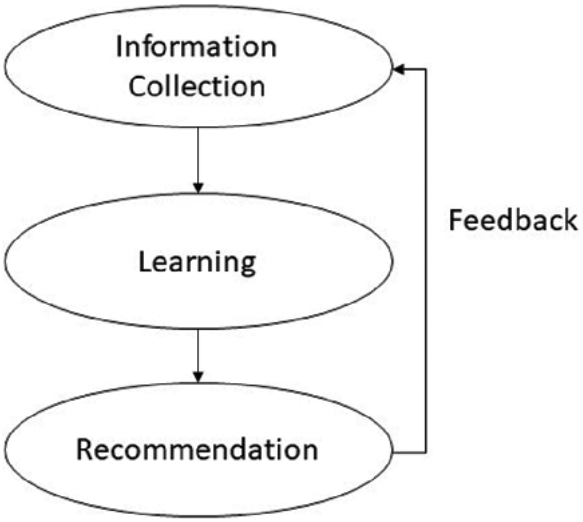


Figure 13: How recommendation models work. Source: (Research Gate,2011)

Popularity: Recommendation systems and algorithms play a very important part in many businesses huge success and popularity. Top tech, entertainment, and e-commerce giants such as Netflix, Shopify, Amazon, YouTube, and Facebook all used recommendation algorithms for personalised customer experience resulting in popularity, customer retention and revenue generation for their companies, thereby saving billions of dollars in manual work and analysis. (Leapfrog Technology, 2021)

In shorter words, "***business lessen the distance between customer needs and satisfaction"*** by implementing product recommendation algorithms.

4.3. Unsupervised learning

Grouping data samples together that are similar index is called unsupervised learning Figure 14. Works by looking for patterns inside data samples without any labels on its own with the least human supervision. It is widely known to perform more robust and complex tasks.

Figure 14&15: Working of unsupervised learning & supervised model. Source: (kaggle.com, n.d.)

The reason for choosing unsupervised machine learning is because:

- It can perform more complex tasks than its counterpart i.e., supervised learning. (SearchEnterpriseAI, n.d.)
- Only the input data or variables (customer quiz answers) will be given.
- Owing to the nature and contents of the datasets provided, it will help us to determine unknown hidden patterns in it.
- It is a reliable method as the learning method in unsupervised machine learning happens in real-time than supervised learning. (Johnson, n.d.)
- According to IBM Cloud Education, it is a very popular method and nowadays very commonly used to improve customers' experience with the business data groupings and patterns in the data. (IBM Cloud Education, 2020)

4.4. Unsupervised Machine Learning Models & Reasons

After researching and trying out many models we have used four unsupervised machine learning models K-means for Yuty product recommendation for customers due to various reasons and factors like data quality, popularity of models, robustness in training and learning and accuracy predictions.

5. Cosine Similarity

This model works on the concept of vectorisation. It's a widely acclaimed algorithm which gives the best product recommendation by calculating the shortest distance between two vectors. Our text data

was converted into vectors and CS algorithm calculated Euclidian distance between them which gave best product recommendation for each user.

The **reason for considering** this model it **works very effectively with small datasets and gives the similarity irrespective of the data samples size**. This algorithm achieved an **accuracy of 52.03%.** We do get the highest product recommendations accuracy in the model when seen product wise.

6. **Random Forest**

Classification in Random Forest is done on the voting process. Every decision tree output is considered, and from the output, the count which is the highest is considered as the result of the model. It models based on bagging and feature randomness.

The **reason for considering** this model it is **most flexible, faster to train and works great on small datasets**. And it has provided favourable results in our analysis with an **accuracy of 36.08%.**

7. **Multinomial NB**

This model is used for both continuous and discrete data. Even Real-time applications can be predicted using this simple method. Usually, it is used for large datasets and is scalable to a large extent. The results are compared in our recommendation model using metrics like precision and F1 score.

The **reason for considering** this model because is it a **low complexity model and gives good results when dealing with small datasets**. In this analysis it has an **accuracy of 30.40%**.

8. **K-means**

This model works on the concept of clustering. K-means works by grouping data samples into clusters that share the same similarities. Clusters nearest to the mean value which is also called centroid calculates the best product suitable for users.

The **reason for considering** this model is **because it is very popular unsupervised learning, easily adapts to new data samples**. In this analysis K-means has an estimated **accuracy of 21.45%**.

We can summarize from our modelling and analysis that Cosine Similarity algorithm performed very well among other three other unsupervised machine learning algorithm Yuty data samples and also gave very effective predictions for each of the 65 Yuty customers which has been very elaborately discussed in Result Analysis section of point 9.1.

5. Yuty Dataset Explanation (By Neel Raut)

The dataset given by Yuty consists of eight data files. Two of them are of JSON origin while the rest are of CSV origin. The dataset gives us information regarding a questionnaire (quiz) that users completed. The 8 data files are as follows – Answer_options.csv, Questions.csv, Question_answers.csv, Quizzes.csv, ing_functions.csv, product_catalogue_data.csv, skin.json, and skin_removals.json.

The in-depth explanation of all the eight data files that we have used for this project is as follows:

1. **QUIZZES.CSV:** This dataset gives us information regarding the quizzes that the users took. The dataset contains 328 rows and has two attributes defining the rows. The two attributes explanation is as follows, Figures 16 and 17 show a pictorial representation of all the essential aspects of the variables:
    - **QUIZ_ID:** This attribute indicates the unique ID given to each quiz. Quiz_id values are in the integer format.

Figure 16:Quiz_ID variable information

    - **USER_ID:** This attribute indicates the unique ID given to each user. User_id values are in the integer format.

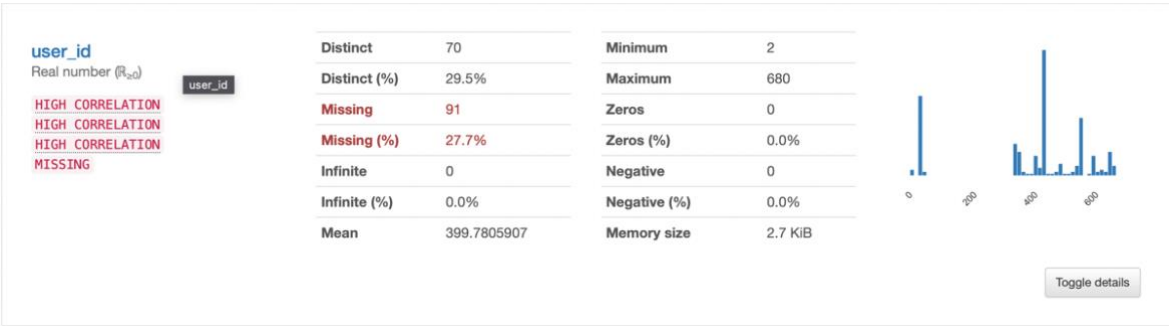


Figure 17: User_ID variable information

2. **QUESTIONS.CSV:** The questions.csv dataset gives us information regarding the questions presented to users in the quizzes. This dataset contains two attributes and has 15 rows. The

attributes are explained below, Figures 18 and 19 show a pictorial representation of all the essential aspects of the variables:

- **QUESTION_ID:** This attribute indicates the unique ID given to each question. Question_id values are in the integer format.

Figure 18: Question_ID variable information

- **QUESTION_TEXT:** Indicates the actual text presented to the users while they were completing the quiz. Question_text values are in the string format.

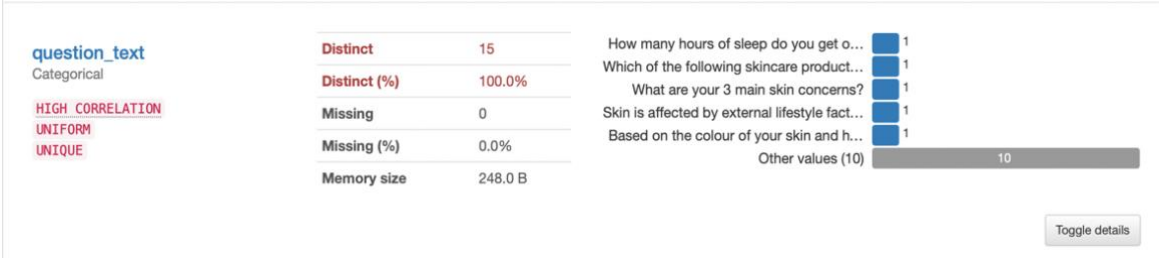


Figure 19: Question_text variable information

3. **ANSWER_OPTIONS.CSV:** This dataset tells us about the different answer options that were presented to the user for each question while completing the quiz. This CSV file contains 97 rows and has three attributes. The attributes are as follows; Figures 20 to 22 show a pictorial representation of all the essential aspects of the variables:

- **ANSWER_ID:** This attribute indicates the unique ID given to each answer option presented to the user while they were given the quiz. Answer_id values are in the integer format.

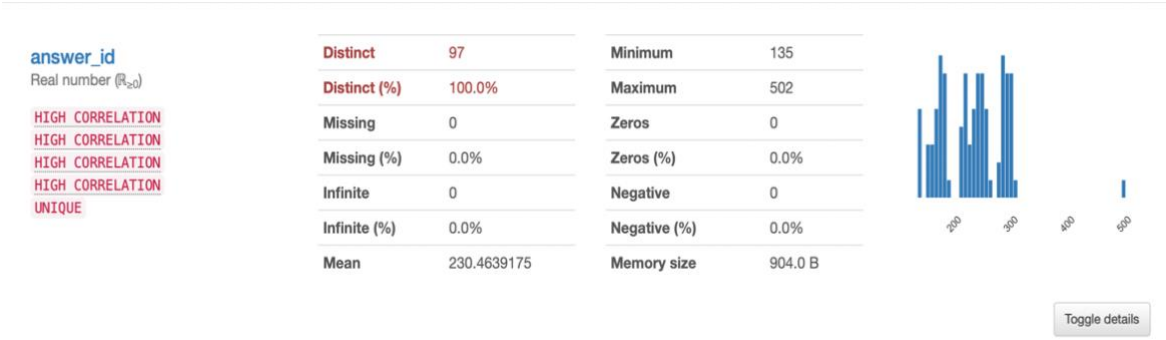


Figure 20: Answer_ID variable information

- **ANSWER_TEXT:** Indicates the text associated with each answer_id. This was shown to the user while they were giving the quiz. Answer_text values are in the string format.

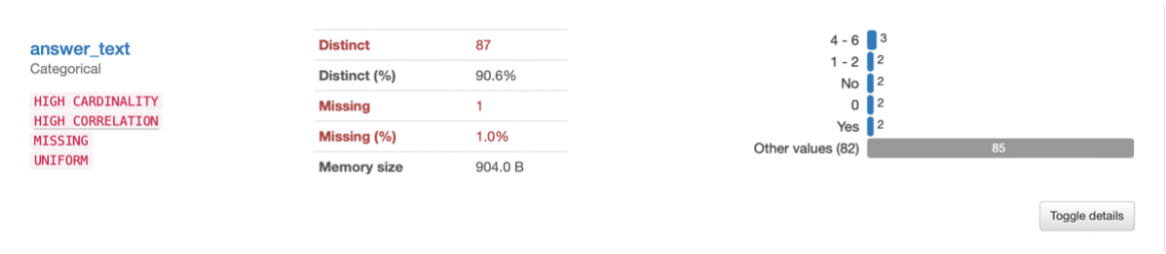


Figure 21: Answer_text variable information

- **QUESTION_ID:** Indicates the question_id that is related to each of the answer_id. Question_id values are in the integer format.

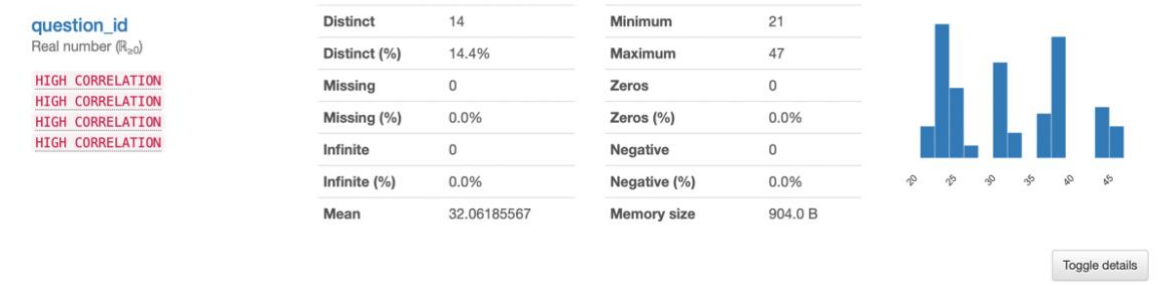


Figure 22: Question_id variable information

4. **QUESTION_ANSWERS.CSV:** The Question_answers.csv dataset tells us about each of the quizzes that were ever taken, the questions that are in that quiz, and the options the user chose while giving the quiz. There are 3026 rows, and four attributes define these rows. The attributes are explained below, Figures 23 to 25 show a pictorial representation of all the essential aspects of the variables:

    - **QUIZ_ID:** Indicates the unique ID for each quiz. Quiz_id values are in the integer format.

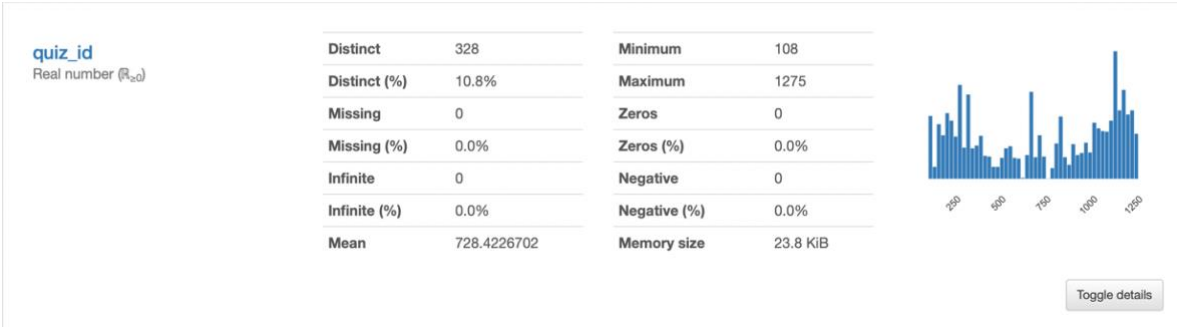


Figure 23: Quiz_ID variable information

- **SPREE_QUESTION_ID:** This attribute tells us about the question id presented to the users for each quiz. Spree_Question_id values are in the integer format.

Figure 24: Spree_Question_id variable information

- **ANSWER_OPTION_IDS:** This variable tells us about the answer options that were selected by the user while giving the quiz. All the answer options are presented in a curly bracket. Answer_option_ids values are in the integer format.

- **CREATED_AT:** This variable gives us information regarding when the users submitted the quiz. Created_at values are in date and time format (YYYY-MM-DD HH:MM:SS).

Figure 25: Created_at variable information

5. **ING_FUNCTIONS.CSV:** The Ing_functions.csv dataset tells us about the different ingredients in the different products that Yuty has to offer. This dataset gives us the ingredient name and whether the ingredients are active or inactive. This dataset has 1377 rows and has two variables defining these rows. The two variables are explained below, Figures 26 and 27 show a pictorial representation of all the essential aspects of the variables:

    - **CUSTOMER_ING:** This variable gives us the different names of all the ingredients in the Yuty products. Customer_ing values are in the string format.

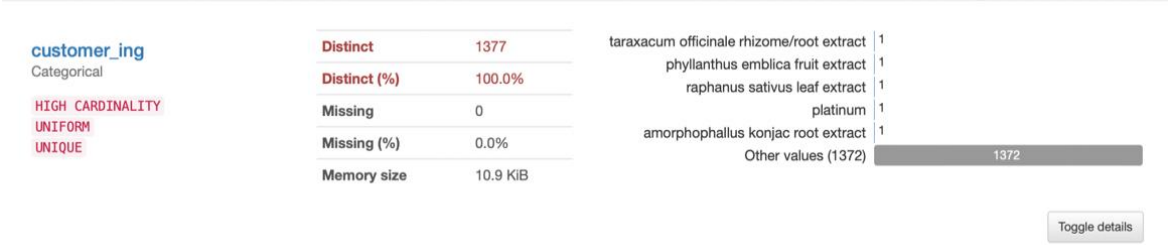


Figure 26: Customer_ing variable information

- **ACTIVE_INACTIVE:** This attribute tells us whether the corresponding ingredient is active or inactive. Active_Inactive values are in the string format.

Figure 27: Active_inactive variable information

6. **PRODUCT_CATALOGUE_DATA.CSV:** The product_catalogue_data.csv file gives us different information regarding the different products that Yuty offers to their customers. This dataset has 350 rows and eight attributes defining these rows. The eight variables explanation is as follows, Figures 28 to 35 show a pictorial representation of all the essential aspects of the variables:

- **PRODUCT_ID:** This attribute indicates the unique value given to each product that Yuty has to offer. Product_id values are in the integer format.

Figure 28: Product_id variable information

- **PRODUCT_NAME:** This variable gives us the actual name of the product that Yuty has. Product_name values are in the string format.

Figure 29: Product_name variable information

- **DESCRIPTION:** This variable briefly describes the product, what the product is essentially made of, and how the product works when used. Description values are in the string format.

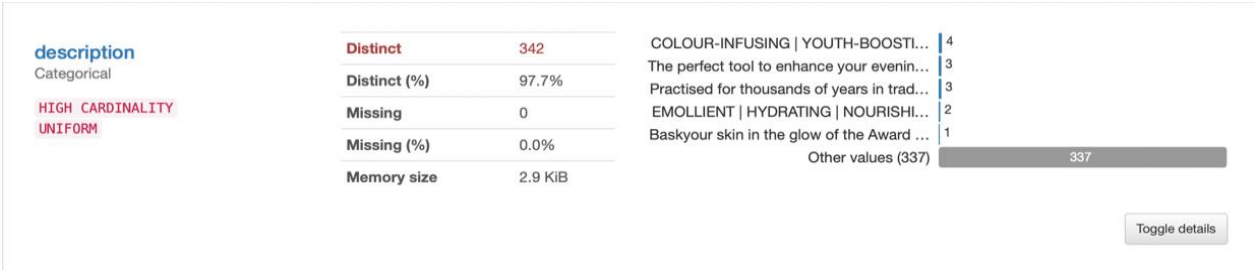


Figure 30: Description variable information

- **SHORT_DESCRIPTION:** This variable gives us a summary of the description that was shown in the Description variable. Short_Description values are in the string format.

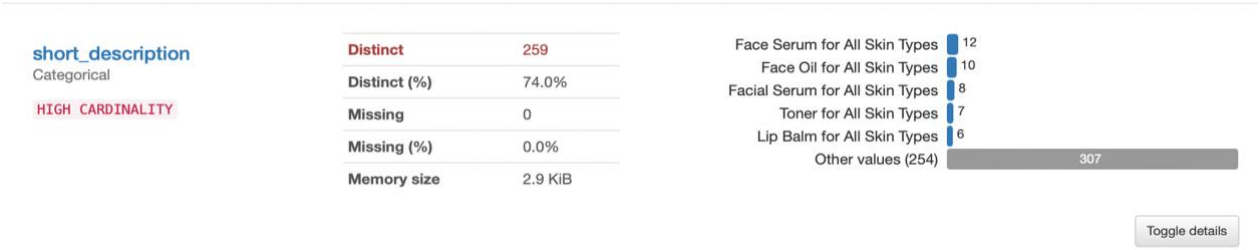


Figure 31: Short_description variable information

- **HOW_TO_USE:** This variable gives us information regarding how the product should be ideally used to get the maximum results. How_to_use values are in the string format.

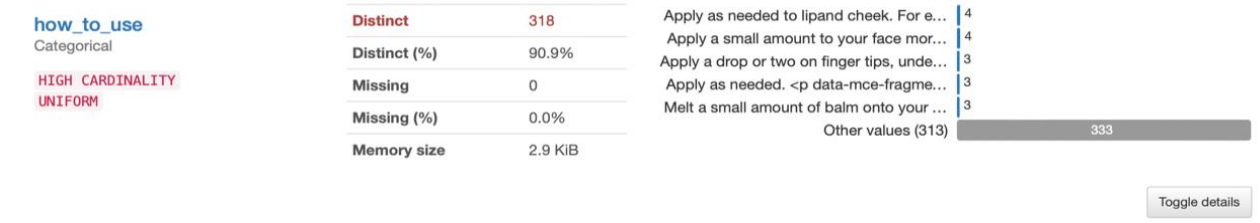


Figure 32: How_to_use variable information

- **INGREDIENTS:** This variable gives us a list of all the ingredients used to make the Yuty product. Ingredients values are in the string format.

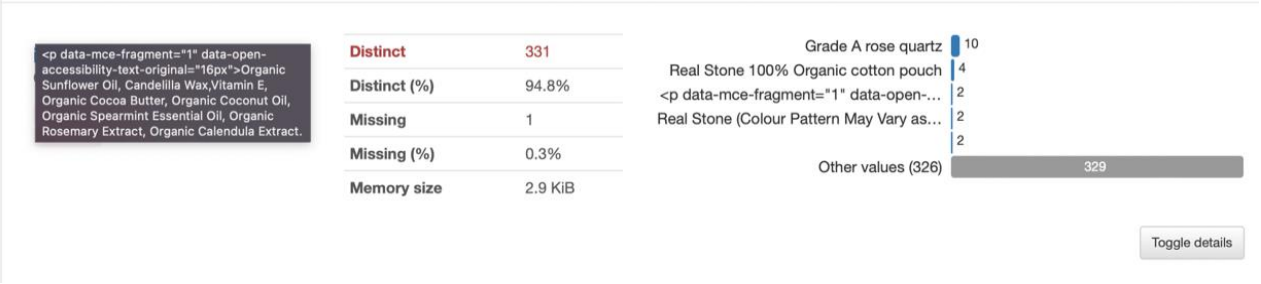


Figure 33: Ingredients variable information

- **WHY_WE_LOVE_IT:** This attribute explains why this product is good and why Yuty loves it. This attribute lists the advantages of using that product. Why_we_love_it values are in the string format.

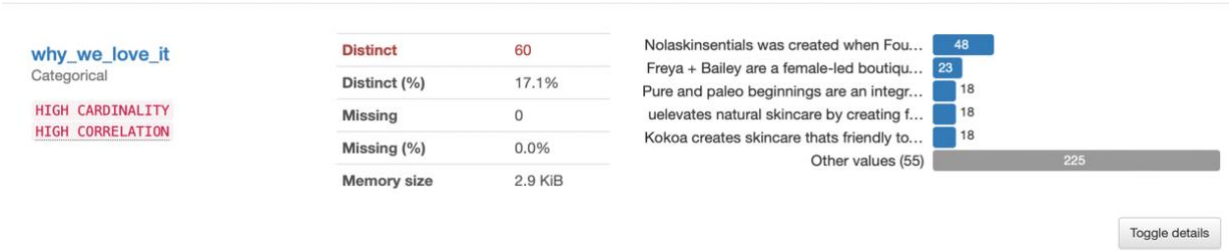


Figure 34: Why_we_love_it variable information

- **GOOD_TO_KNOW:** This variable gives us a detailed explanation of different things about the product. This information could be related to how the product is made, how many times to use the product, what kind of chemicals are not used to make the product, whether the product is cruelty-free, and if the product has passed any certification. Good_to_know values are in the string format.

Figure 35: Good_to_know variable information

7. **SKIN.JSON:** The skin.json file is a critical data file as it gives us information regarding the ingredients associated with the different answers selected by the user while giving the quiz. This file is an index-oriented JSON file. The first index in this file is regarding the question_id. The second index gives us information regarding the different answer options presented for the question. Each answer option also gives us information regarding the different ingredients associated with that answer option. All the ingredients in this file are separated by a tilde (~). This file is essential because this file is used while making recommendations.
8. **SKIN_REMOVALS.JSON:** The skin_removals.json file is precisely like the skin.json file. The only difference between those two is that the ingredients listed in the skin.json file should be present in the recommended products, and the ingredients listed in the skin_removals.json file should not be present in the products that are recommended to the users.

Figure 36 depicts a database ***schema diagram*** showing a pictorial representation of all eight datasets and how they are related to each other.

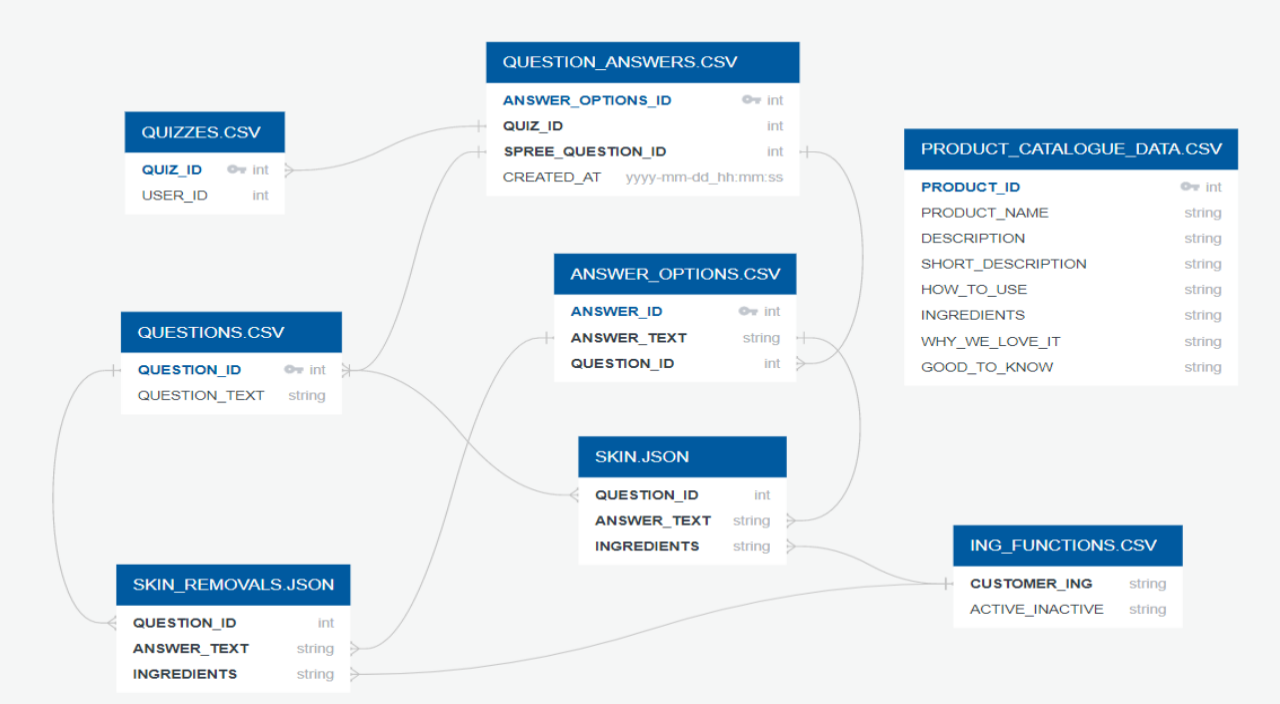


Figure 36: Database Schema Diagram explaining the eight datasets and how they relate.

6. Data Pre-processing (By Shrey Agarwal)

Our goal is to make a text classification model which accurately assigns a sentence or a document to an appropriate category. This will help us to organise and structure our data, i.e., ingredients data, and therefore help us categorise the right products for the Yuty customers.

We usually use supervised learning to classify text, but the products and ingredients are not labelled in our product catalogue dataset. Usually, data annotation requires a lot of manual effort and high expenses. Therefore, we have chosen two methods to solve this problem.

- The First method is to use the unsupervised learning approach to classify the text and
- The second method is to manually label a sub-sample and train the model to predict the labels for the rest of the data and then use Supervised learning to classify text.
- We need to pre-process the data for modelling both the methods that we are going to use.

6.1. Data Manipulation & Data Formatting

After observing all the data in the given data files, we found some missing data in the Questions_answers and Quizzes files. After realising it we talked to the people at Yuty about it, and they told us that these missing values originate when a customer filling out the survey has an internet issue and cannot answer the question. Realising it we concluded that these rows are of no use to us and therefore need to be dropped. We then removed these values and then merged all the customer response data. This is how the merged files looked –

	quiz_id	user_id	spree_question_id	answer_option_ids	created_at
0	108	361.0	80	{}	2021-08-06 15:26:11.936119
1	108	361.0	44	{293}	2021-08-06 15:26:35.230364
2	108	361.0	28	{}	2021-08-06 15:26:35.231366
3	108	361.0	45	{296}	2021-08-06 15:26:35.232891
4	108	361.0	46	{}	2021-08-06 15:26:35.233819

Now we come to the product_catalogue_data file, which needs the most cleaning. This file contains seven columns and 384 rows. The seven columns are as follows - **product_id, product_name, description, short_description, ingredients, why_we_love_it and good_to_know.** Since we only need to match the products' ingredients to the ingredients required by the user to improve their skin, we only need product_id, product_name and ingredients column; the rest is dropped from the file. Now we come to the main part, which is the ingredients column.

Since the data of products are taken from different sources, this data is not in readable form for the computer to match it with the user ingredients. Therefore, we need to clean the ingredient's text so that it can be matched. This is done using a famous automatic computational processing of human languages called Natural language Processing.

Natural Language Processing, or NLP, is a part of AI that deals with transforming human language into a language that is easily understood by computers. It has various techniques that can help us interpret human languages. We will use NLP to clean our data and create features for our supervised and unsupervised learning models.

NLP is usually used for tokenization, language detection, and identification of semantic relationships.

In the section, we used NLP to:

- Removing punctuation, symbols – This removes everything that is not an alphabetic or numeric letter i.e., /';- <> Etc.
- Lowercase – To match words like Play and play, we need to lowercase all the letters in the text.
- Stop Words – We removed all the connecting words like 'the', 'and', 'at' since they are unnecessary while building our model.
- We have also used Excel to clean our text files. Both these techniques have helped us remove punctuation, unnecessary words and much more.

Now we are left with labelling the data. Since our data is unsupervised, we are not given any labels, but after talking to Yuty, they provided us with their different product categories –

- Cleanser
- Day Moisturiser
- Toner
- Serum / Oil
- Eye Cream
- Masks / Treatments
- Exfoliant (Physical/Chemical/Enzymatic)
- Lip Care
- Spf
- Night Moisturiser

We used these categories to manually label a sub-sample which we will use for training our data. This will help us in measuring accuracy among models and these accuracy measures can be used to compare the models and choose the best one for our recommendations. This labelled data can be used by Yuty too to structure their models.

We don’t need any pre-processing on the skin. JSON, skin_removal. JSON and ing_functions.csv files as they are already structured.

6.2. Splitting of Dataset

The data is then split into testing and training data for each type of algorithm to train our model and check for accuracy.

```
from sklearn.model_selection import train_test_split
train_x, valid_x, train_y, valid_y = train_test_split(trainDF['newtext'].values, trainDF['label'])

from sklearn.model_selection import train_test_split
train_x1, valid_x1, train_y1, valid_y1 = train_test_split(trainDF.iloc[:,3: 6].values, trainDF['label'])
```

6.3. Result Summary

In this section, the data has been structured our data and pipelines have been created for the models to run. We have come to a step further in making our classification model and then using that model to recommend the right products to the customers of Yuty. The next step is feature modelling which is also covered by me.

Text Before Cleaning:

```
'<p>Prunus Armeniaca (Apricot) Kernel Oil, Cera alba, Butyrospermum parkii (Organic Shea) Butter, Vitis vinifera (Grapeseed) Se
ed Oil, Rosa canina (Rosehip) Fruit Oil, Helianthus annuus (Sunflower) Seed Oil, Simmondsia chinensis (Jojoba) Seed Oil, B
orago officianalis (Borage) Seed Oil, Hippohae rhamnoides (Sea buckthorn) Pulp Oil, Boswellia carterii (Frankincense)
Oil, Tocopherol (Vitamin E), Cedrus atlantica (Cedarwood) Bark Oil, Citrus aurantium dulcis (Orange) Peel Oil, Pogost
emon cablin (Patchouli) Leaf Oil, Cananga odorata (Ylang Ylang) Flower Oil, Propylene glycol, Styrax benzoin (Benzoi
n) Gum Oil, Zingiber officinale (Ginger) Root Oil, Commiphora myrrha (Myrrh) Resin<br />\nOil, Santalum spicata (Sandalwoo
d) Wood Oil, *Limonene,*Benzyl Salicylate, *Farnesol,*BenzylBenzoate* Naturally occuring in Essential oil</p>\n'
```

Text After Cleaning:

```
'Prunus Armeniaca  Kernel Oil, Cera alba, Butyrospermum parkii  Butter, Vitis vinifera  Seed Oil, Rosa canina  Fruit Oil, Helia
nthus annuus Seed Oil, Simmondsia chinensis  Seed Oil, Boragoofficianalis  Seed Oil, Hippohae rhamnoides  Pulp Oil, Boswellia c
arterii  Oil,Tocopherol , Cedrus atlantica  Bark Oil,Citrus aurantium dulcis  Peel Oil, Pogostemon cablin Leaf Oil, Cananga odo
rata  Flower Oil,Propylene glycol, Styrax benzoin  Gum Oil, Zingiberofficinale  Root Oil, Commiphora myrrha  ResinOil, Santalum
spicata  Wood Oil, Limonene,Benzyl Salicylate, Farnesol,BenzylBenzoate Naturally occuringin Essential oil'
```

7. Feature Engineering (By Shrey Agarwal)

After pre-processing the data comes feature engineering in which we use the raw dataset and use it to create flat features which help us to train our ML models.

In this step, raw text data will be transformed into feature vectors and new features will be created using the existing dataset. We will implement the following different ideas in order to obtain relevant features from our dataset.

This part of the project is important because if we find the right features to train our model, then we will have very good accuracy, which will help us match the customers of Yuty to the right product solving the issue that the company is targeting.

Here we are presenting five of the most commonly used modelling techniques: the Model based on the TF-IDF Algorithm, the Model based on the Count vector Algorithm, GloVe Models, LDA Algorithm and text as features.

7.1. Vectorisation

7.1.1. Count Vectors as features

We use the Count Vectorization function to count vectors. This is a technique used to convert a given text into a vector based on how many times each word appears across our raw data Figure 37. We use this for text analysis and when we have the same words occurring at different times in our text. It returns a matrix in which each unique word is represented by a column of the matrix, and each text sample from the document is a row in the matrix. (Analytics Vidhya, 2019)

Count Vectorizer

	also	love	programming
love programming	0	1	1
programming also love	1	1	1

TF-IDF Vectorizer

	also	love	programming
love programming	0.000000	0.707107	0.707107
programming also love	0.704909	0.501549	0.501549

Figure 37 (www.linkedin.com, n.d)

7.1.2. TF-IDF Vectors as features

Another technique to vectorize the data is transforming the text into TF-IDF vectors. TF-IDF means Term Frequency - Inverse Document Frequency. This is different from Count vectors in terms that it is based on the frequency of a word in the corpus, but it also provides a numerical representation of how important a word is for statistical analysis. (Analytics Vidhya, 2019)

Since it provides importance to words, we can now eliminate less important words for analysis, which makes our model less complex hereby making it easy for the models to run.

Vectorization is done because when classifying, NLP models can't understand textual data they only accept numbers, so this textual data needs to be vectorized. This is also a disadvantage of it.

To provide meaning to the words and similarity with other words, we need to use embedding techniques. These provide actual linguistic relationships of the words to train our model.

We have used both these vectorization techniques to transform our model and results have been compared to see which has performed better.

7.2. Word Embedding

Word Embedding constructs a representation of words that reflects their semantic links, meanings, and various usage contexts. The two widely used Word Embedding techniques are "word2vec" and "Glove". These are also called Pre-trained Word Embedding because they are trained on large datasets saved and then used for solving other tasks.

We use these pre-trained techniques on our simple models because they capture the semantic and syntactic meaning of a word and help improve our Natural Language Processing Model.

We will talk about both techniques in the below subpoints points **7.2.1, 7.2.3, 7.2.3**.

7.2.1. Word2vec as features

One of the most well-known pre-trained word embeddings created by Google is Word2Vec. The Google News dataset is used to train Word2Vec (about 100 billion words). It has a variety of applications, including Recommendation Engines, Knowledge Discovery, and various Text Classification issues.

Word2Vec's architecture is really straightforward. A single hidden layer feeds forward neural network is what it is. As a result, the architecture is also sometimes called a shallow neural network.

7.2.2. GloVe as features

The gloVe is a word vector technique that rode the wave of word vectors after a brief silence. Just to refresh, word vectors put words in a nice vector space, where similar words cluster together and different words repel.

GloVe is better than Word2vec in the sense that GloVe incorporates word co-occurrence to obtain word vectors whereas Word2Vec only relies on local context information of words (Nasher, 2021).

Since these word embedding files are quite large, they take a lot of time to train. Therefore, we are going to use only one method of word embedding techniques. We will go with GloVe as we know this is a better technique which in turn will provide us with a better model.

These are the following steps we use to train our model -

1. A pre-trained text data is downloaded - embeddings_index is created by loading the 'glove.42B.300d.txt' file.
2. Then the labels are tokenized– A variable token is created using text.Tokenizer()
3. Text documents are then transformed - train_seq_x and valid_seq_x is created with a maximum length of 70.
4. Create a mapping of tokens and their respective embeddings - embedding_matrix[i] is created from the word index.

These are some of the results of our text data:

```
Number of unique words in dictionary= 814
Dictionary is =  {'oil': 1, 'seed': 2, 'extract': 3, 'sodium': 4, 'acid': 5, 'flower': 6, 'citrus': 7, 'glycerin': 8, 'aqua':
9, 'organic': 10, 'leaf': 11, 'gum': 12, 'tocopherol': 13, 'root': 14, 'rosa': 15, 'alcohol': 16, 'fruit': 17, 'prunus': 18,
'essential': 19, 'water': 20, 'simmondsia': 21, 'aloe': 22, 'barbadensis': 23, 'vitamin': 24, 'chinensis': 25, 'dagger': 26,
'vegetable': 27, 'powder': 28, 'helianthus': 29, 'officinalis': 30, 'potassium': 31, 'cetearyl': 32, 'xanthan': 33, 'citric':
34, 'oils': 35, 'aurantium': 36, 'kernel': 37, 'hyaluronate': 38, 'annuus': 39, 'phenoxyethanol': 40, 'ferment': 41, 'filtrat
e': 42, 'canina': 43, 'butter': 44, 'glutamate': 45, 'clay': 46, 'ascorbyl': 47, 'sativa': 48, 'and': 49, 'caprylic': 50, 'ca
pric': 51, 'limonene': 52, 'glyceryl': 53, 'tetrasodium': 54, 'cocos': 55, 'persea': 56, 'gratissima': 57, 'dulcis': 58, 'leu
conostoc': 59, 'peel': 60, 'radish': 61, 'nucifera': 62, 'parkii': 63, 'lavandula': 64, 'linalool': 65, 'melaleuca': 66, 'dia
cetate': 67, 'armeniaca': 68, 'sorbate': 69, 'butyrospermum': 70, 'angustifolia': 71, 'argania': 72, 'spinosa': 73, 'benzyl':
74, 'graveolens': 75, 'glycerine': 76, 'blend': 77, 'e': 78, 'allantoin': 79, 'polysorbate': 80, 'ethylhexylglycerin': 81, 'c
alendula': 82, 'pelargonium': 83, 'palmitate': 84, 'olivate': 85, 'stone': 86, 'commiphora': 87, 'multivitamin': 88, 'comple
x': 89, 'certified': 90, 'natural': 91, 'olea': 92, 'glycol': 93, 'alba': 94, 'squalane': 95, 'benzoate': 96, 'glucoside': 9
7, 'sativus': 98, 'wax': 99, 'in': 100, 'c': 101, 'a': 102, 'damascena': 103, 'phosphate': 104, 'avena': 105, 'beta': 106, 'h
yaluronic': 107, 'stearate': 108, 'nobilis': 109, 'sorbitan': 110, 'cucumis': 111, 'sinensis': 112, 'hamamelis': 113, 'de': 1
14, 'stressing': 115, 'urea': 116, 'alpha': 117, 'acetate': 118, 'geraniol': 119, 'rosmarinus': 120, 'caprylate': 121, 'curcu
ma': 122, 'bark': 123, 'citral': 124, 'hydrogenated': 125, 'triglyceride': 126, 'vinifera': 127, 'betaine': 128, '100': 129,
'zinc': 130, 'panthenol': 131, 'from': 132, 'stearic': 133, 'oxide': 134, 'vaccinium': 135, 'chamomilla': 136, '20': 137, 'ci
trate': 138, 'l': 139, 'amygdalus': 140, 'eugenol': 141, 'brassica': 142, 'anthemis': 143, 'olive': 144, 'sitosterol': 145,
```

7.2.3. Text / NLP-based features

We can even create some really important features for our classification model using the raw text data that we have with us (Analytics Vidhya, 2019). These could be:

1. The length of the characters in the text.
2. Number of words in the text.
3. Number of Upper-Case words in the text.
4. Number of lower-Case words in the text.
5. Frequency distribution of nouns, verbs, adverbs, pronouns, etc.

Like this, I have created 6 new features for our model.

	char_count	word_count	word_density	punctuation_count	title_word_count	upper_case_word_count
count	100.00000	100.000000	100.000000	100.000000	100.000000	100.00000
mean	342.49000	41.170000	7.978304	23.710000	31.930000	1.00000
std	212.54321	25.249084	1.272339	16.393085	23.669123	1.57634
min	11.00000	2.000000	3.666667	0.000000	1.000000	0.00000
25%	189.50000	22.000000	7.691106	13.000000	14.750000	0.00000
50%	300.00000	36.500000	8.152681	20.500000	26.000000	0.00000
75%	434.75000	52.000000	8.646077	32.000000	43.500000	1.00000
max	886.00000	109.000000	12.600000	72.000000	100.000000	9.00000

Sometimes these types of features help us to explain the dependent variables and therefore help us to predict the values.

7.3. Topic Models as features

Now we move to Topic Modelling. This is a technique to identify the groups of words (called a topic) from a collection of documents that contain the best information in the collection. It is a type of unsupervised learning that looks for similar words in a document and tries to categorize them by making clusters of them as you can see from the figure 38. Using Topic Modelling as a feature is still relatively new but it has been found effective for making classification models.

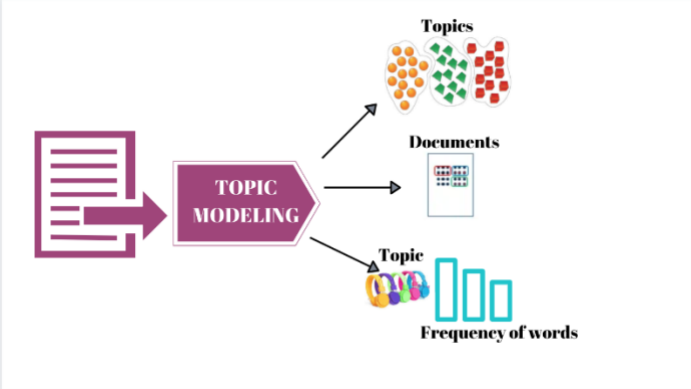


Figure 38: Topic Modelling (Khushalani, 2019)

The most famous top modelling methods are **Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA).**

7.3.1. Latent Semantic Analysis (LSA)

Latent Semantic Analysis has its basis come from top modelling but is similar to cosine similarity. In LSA, we generate a matrix from a bag of words that are presented in a document. In this matrix, the rows represent unique words found in each paragraph of the document and columns represent each paragraph. Later the model uses singular value decomposition to latent topics and does a matrix decomposition assuming that the words that can have similar meanings will occur in a similar part of the text. These text summaries can now be used as features for our text classification model.

7.3.2. Latent Dirichlet Allocation (LDA).

Another popular topic modelling algorithm is Latent Dirichlet Allocation (LDA). It is different from other models like LSA in the sense that it is a type of supervised learning which gives interpretable topics. The disadvantage is that while modelling, it ignores word order and semantic information. This algorithm works on two properties-

Every document is a mixture of topics. We speculate that certain word counts from various topics may be used in each text. In a two-topic model, for instance, we might state that "Document 1 is 90% subject A and 10% topic B, while Document 2 is 30% topic A and 70% topic B."

Every topic is a mixture of words. As an illustration, consider an American news model with two topics: "politics" and "entertainment." In a topic of politics, the most frequent words might be "President," "Congress," and "government," but in a topic of entertainment, the most frequent words might be "movies," "television," and "actor." Words can be shared between themes, which is significant because a word like "budget" may appear in both equally. (Robinson, n.d.)

In our project, we will LDA as a feature and not LSA because our text data is ingredients. Therefore, there is no need for word order and semantics while modelling.

Below is the topic summary for the document –

```
['dagger turmeric longa vanilla oil ecocert curcuma chinesesis olibanum kalonite',
 'stressing de complex multivitamin blend oil glycerine avocado nucifera vegetable',
 'armeniaca pumice sustainably ground kernel stearic lentinus phthalate africanus cinnamomum',
 'collagen ndash ingredients seaweed 85 elastin hazel helix mucopolysaccharides antimicrobials',
 'triglyceride glucoside var chinensis flour glycerin marula yangu pomegranate canina',
 'vegetal soil association silver dna origin preservative glycolic ecocert from',
 'stone cotton 100 real vary colour may is this pattern',
 'calendula meadowfoam nigella distillate limnanthes capryl cocomidopropyl veg black oilorganicingredient',
 'a eo vitamins fragrance clay powder in rich antioxidants d',
 'herb zea freereformulated hibiscus flour ubidecarenone azadirachta mays phthalate melia',
 'preservative water sodium angustifolia a certified levulinate farming flower fragment',
 'b copper vitamin zinc clay e earth citrullus as wood',
 'oil stone olivate seed acid citric titanium from extremely ternifolia',
 'oil seed extract sodium flower acid glycerin citrus aqua leaf',
 'root chestnut horse esculin liquorice glabra glycyrrhiza caffeine saponified album',
 'oil extract leaf resin coffea seed alcohol arbutin barbadensis cucumis',
 'ceramide 5 niacinamide floral b3 vitamin water polylysine willow aloe',
 'powder cacao dehydroacetate chlorphenesin propylene chondrus crispus seaweed tremella arbutin',
 'burdock hamamelis root fucus vesiculosus vegan serrulata serenoa seaweed larrea',
 '15 hyaluronic c vitamin acid kelp coenzyme q10 thioctic cinnamon']
```

7.4. Result Summary

To sum up, here, we have used feature engineering as a tool for creating features used for our ML models from the raw ingredients text data. We have created different types of features which will act

as our hyperparameters in finding the best results from our ml models. After comparing the algorithms by looking at research papers and analysing them, we are going forward with 4 different methods – TFIDF algorithm, Count vector Algorithm, GloVe model and text as features. These features will be applied on our supervised ML models to predict the products for Yuty customers.

8. Exploratory Data Analysis (By Anshul Basotia & Faran Saeed)

8.1. Data Analysis

The main aim of exploratory data analysis is to understand data distribution more efficiently and analyse the pattern of the dataset with the help of the visualisation of graphs and statistics. In this project, we have a total of eight dataset files i.e., six CSV files and two JSON files. In exploratory data analysis, we did data processing which improved the quality of the analysis. Machine learning helped to get good accuracy which was not possible before because of missing values and less data in the dataset files. Data analysis helped to give information regarding the recommendation system is less accurate because of the size of data is small.

8.2. Bags of Words

Bag of words as the name suggests is a model using the world cloud library used to represent text data when machine learning algorithms are used. The figure below is achieved during our analysis which represents occurrence of each word in this case ingredients in product catalogue. The bigger the letter, the more the number of repetitions. The model bag of words is considered as a simple portrayal used in natural language processing; it is a model where texts are being represented as bag of its words. There is no regard of grammar or word order in this model, but the font gets bigger and bigger to show the repetition of a certain word being used multiple times.

The table in Figure 39 below shows the number of times the top 10 words were repeated.

1	aloe vera callus extract	1116
2	amaranthus hypochondriacus seed extract	1116
3	3-laurylglyceryl ascorbate	1116
4	achillea millefolium flower extract	1116
5	scutellaria baicalensis root extract	726
6	bambusa vulgaris shoot extract	703
7	bis-glyceryl ascorbate	703
8	citrus unshiu fruit extract	703
9	citrus australasica seed oil	703
10	arthrospira platensis extract	703

Figure 39: Table of word(ingridients) occurances.

Leaf extract, fruit extract, ferment filtrate and seed oil are the words with the biggest font portraying that these words have been repeated the most in the datasets as you can see below in Figure 40.

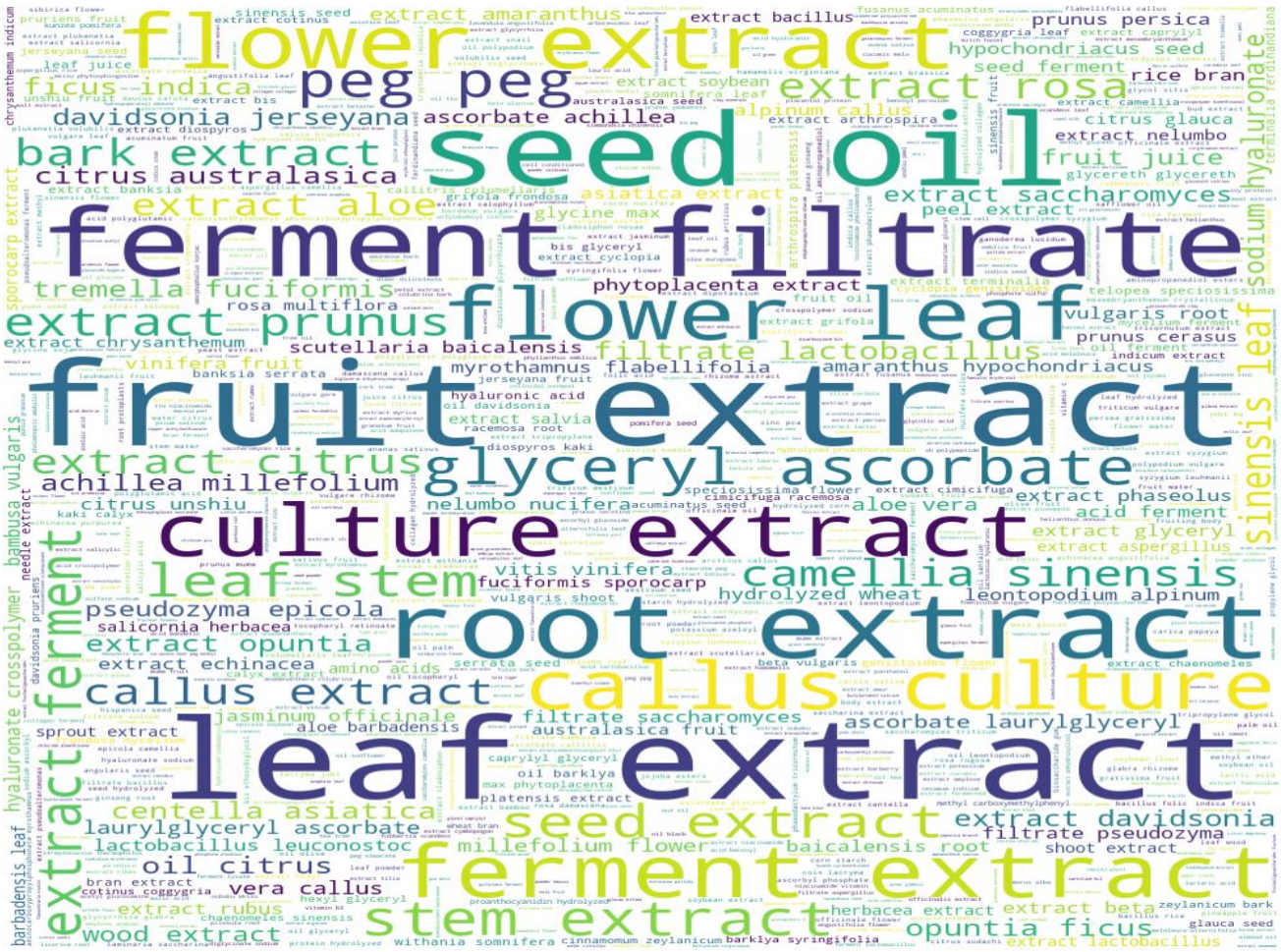


Figure 40: Bag of Words achived by wordcloud library in Python.

8.3. Analysis of Missing values in Datasets

Here is a detailed exploratory data analysis of the data samples provided by Yuty. By performing this analysis, the quality and strength of data samples and progress further with data pre-processing, manipulation, feature engineering and machine learning modelling we had to do in order to get good accuracy with our unsupervised learning algorithms to recommend the product with highest accuracy to Yuty customers.

QUIZZES.CSV:

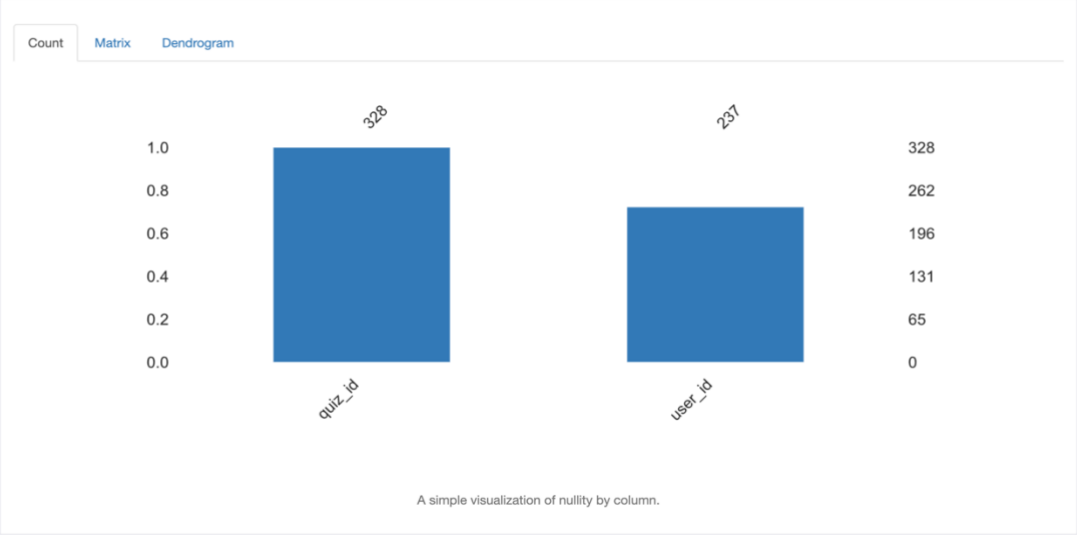


Figure 41: Quizzes.CSV

In QUIZZES.CSV dataset there are two attributes which are QUIZ_ID and USER_ID. The maximum of 328 need to be in this dataset.

QUIZ_ID – The values in this attribute are in integer format and there are no missing values in it, maximum of 328 values are there.

USER_ID- In this particular dataset there are 237 rows. The missing values in this attribute are 91, each user is given a unique ID.

QUESTIONS.CSV:

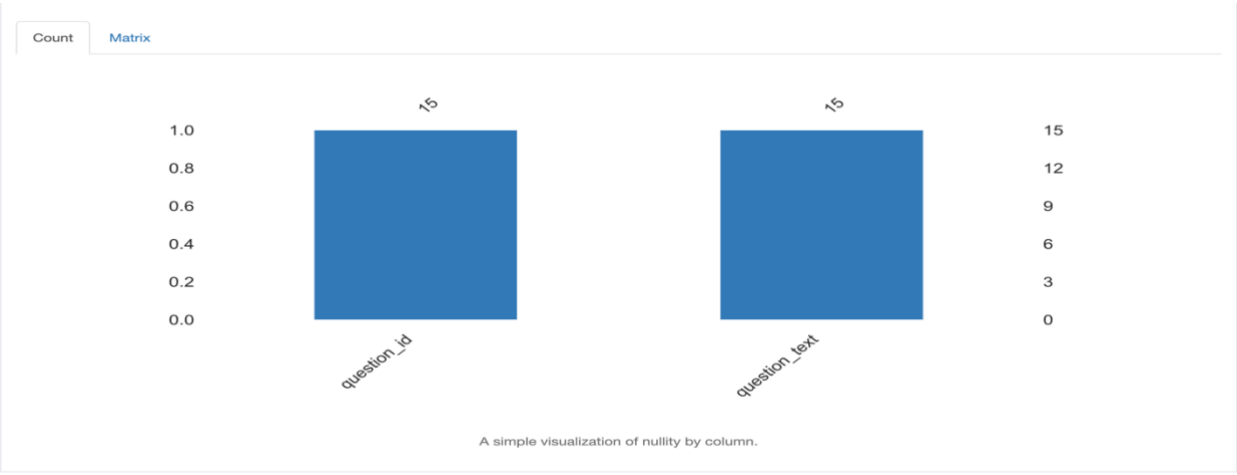


Figure 42: QUESTION.CSV

In QUESTIONS.CSV dataset there are two attributes i.e. QUESTION_ID and QUESTION_TEXT. The dataset has maximum of 15 rows. There are no missing values in the attributes of this dataset.

QUESTION_ID attribute has values in integer format while the attribute of QUESTION_TEXT has values in string format.

ANSWER_OPTIONS.CSV:

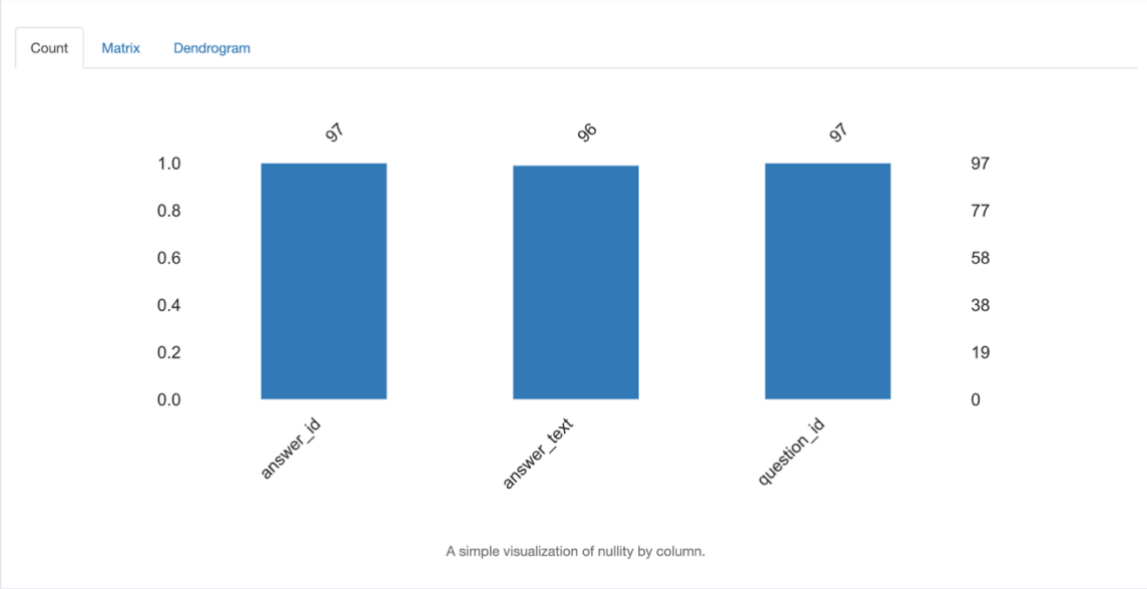


Figure 43: ANSWER_OPTIONS.CSV

ANSWER_OPTIONS.CSV dataset has total of three attributes which are ANSWER_ID, ANSWER_TEXT and QUESTION_ID. The values in this dataset have the maximum of 97 rows. ANSWER_ID and QUESTON_ID are the attributes in this dataset with no missing values but the attribute of ANSWER_TEXT has one missing value in it. This attribute is in string format, it also indicates the text associated with each attribute of ANSWER_ID.

QUESTION_ANSWERS.CSV:

In QUESTION_ANSWER.CSV we can see that there are three attributes which are quiz_id, question_id and answer_id. From the visualisation, we can see that there are no missing values in either of the attributes as there are a maximum of 3026 entries.

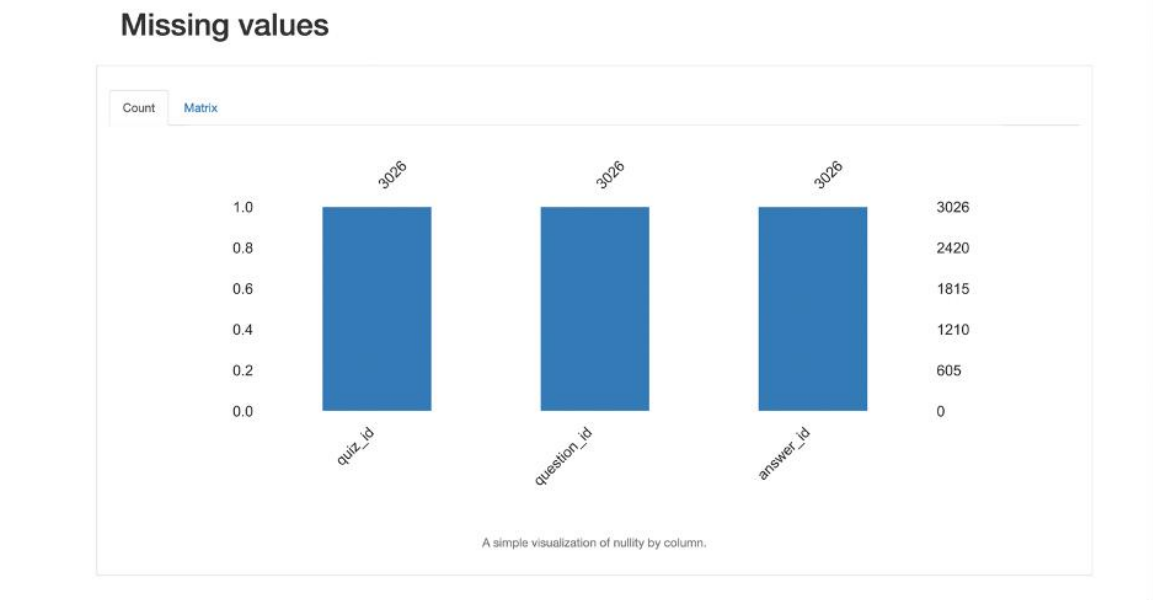


Figure 44: QUESTIONS_ANSWERS.CSV

ING_FUNCTIONS.CSV:

In ING_Function.CSV we can see that there are two attributes which are customer_ing, and active_inactive. From the visualisation, we can see that there are no missing values in either of the attributes as there are a maximum of 1377 entries.

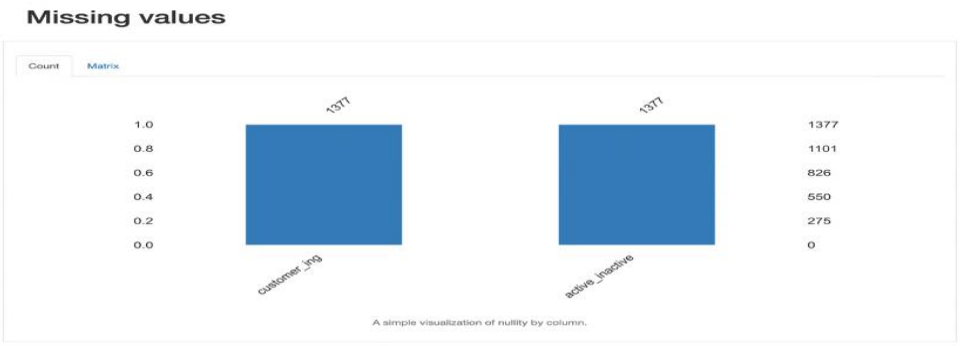


Figure 45: ING_FUNCTIONS.CSV

PRODUCT_CATALOGUE_DATA.CSV:

In PRODCUT_CATALOGUE_DATA.CSV we can see that there are nine attributes which are product_id, product_name, description, short_description, how_to_use, ingredients, why_we_love_it, good_to_know and pred . From the visualisation, we can see that there is one missing value in ingredients as there are a maximum of 350 entries.

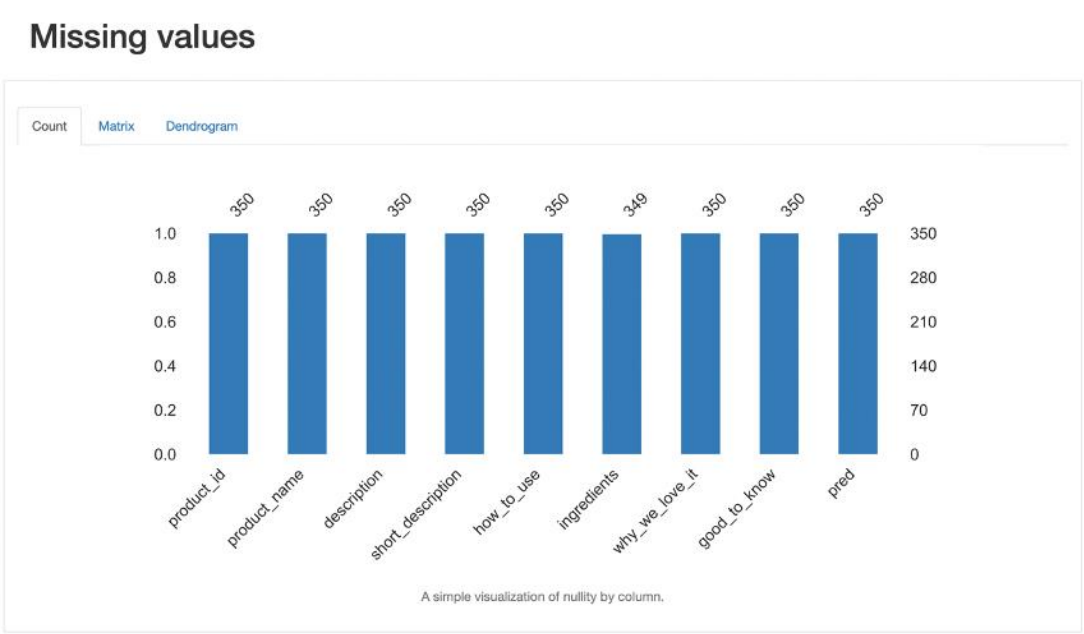


Figure 46: PRODUCT_CATALOGUE_DATA.CSV

9. Machine Learning Methodology (By Shalini Nayak & Shrey Agarwal)

9.1. Cosine Similarity (By Shalini Nayak) (Best Model)

9.1.1. Background

Cosine Similarity is a widely used algorithm which is majorly used in text mining and product recommendation. Cosine similarity uses the concept of vectorization in our text analysis. Figure 47 below shows the working of CS, which depicts the "smaller the angle between two vectors, higher the similarity".

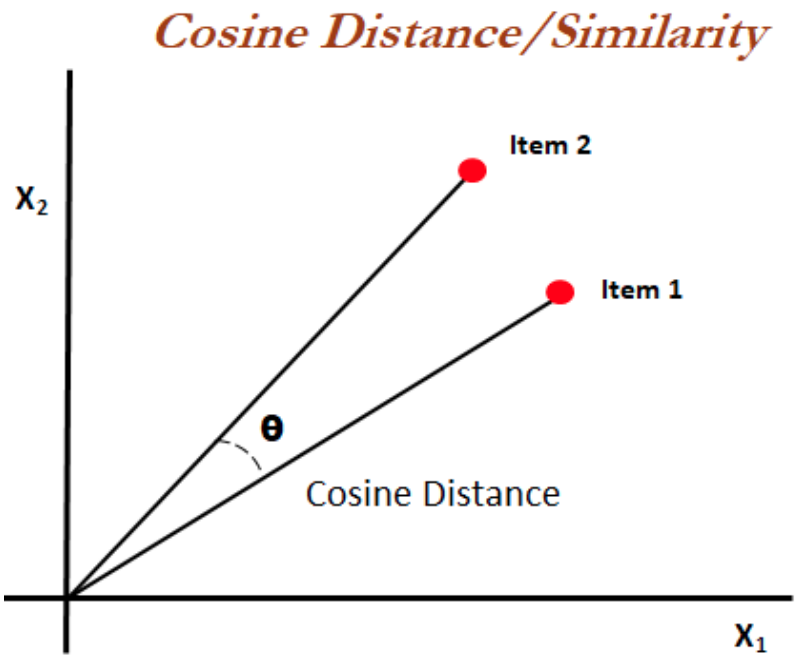


Figure 47: Cosine Similarity. Source: (Prabhu, 2019)

Vectorisation: It is a process of converting text into vectors, as text can't be fitter to a machine learning model, we need a numeric value hence we have converted the ingredients present in Yuty products into a vector. It's a very important process in text analysis because only by vectorization - text data can be readily consumed by algorithms like clustering and search etc. Vectorisation has been explained ha in the subpoint of Feature Engineering in 8.1. Figure 48 process shows the working of vectorization.

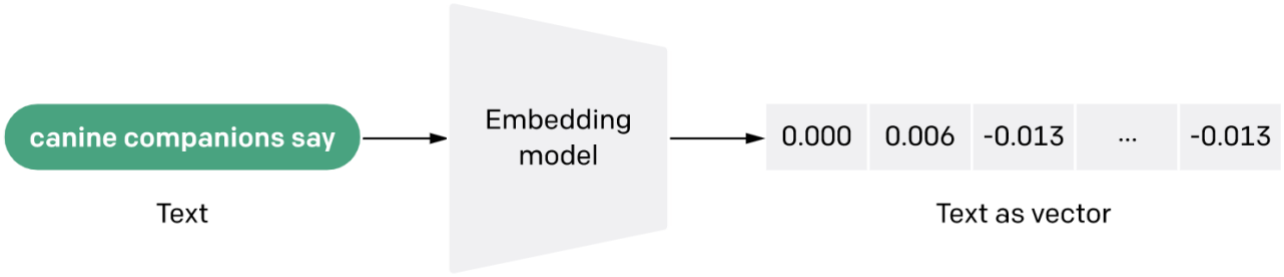


Figure 48: Text to Vector. Source: (OpenAI, 2022)

Reason for choosing cosine similarity is because ***it is low complexity and works well with small data which and sparse data and gives the nearest distance between 2 vectors,*** was very suitable for our analysis with Yuty datasets. Also, CS has a reputation of giving good results in text mining, product recommendation and information retrieval analysis.

We are getting a good accuracy of ~53% on the model taken on the prediction analysis for the total number of users.

```
Accuracy of Cosine Similarity Algorithm:  52.03440000000001
```

Figure 49: Accuracy of Cosine Similarity.

But we do get the highest product recommendations accuracy for Yuty customer where top three products are recommended for each user which solves percentages of their problems product wise which is discussed in point 9.1.3 below.

9.1.2. Steps/Procedure

"Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. "(Han, Kamber and Pei, 2012)

Step 1: Firstly, we have converted ingredients from the product catalogue and answers from customer's quiz which are text into two different vectors.

Step 2: And then, using the cosine similarity algorithm we have compared the similarity between then and used for predicting the best recommendation for Yuty customers.

9.1.3. Result Analysis

In this part, we have analysed the result and accuracy through the Cosine Similarity Algorithm. This is the best model far among the three other unsupervised machine learning models.

Every quiz answer is mapped to a set of ingredients, so these ingredients are mapped to the ingredients which are contained in the different products made by Yuty through an unsupervised machine learning algorithm calculating the closest distance. In our analysis, it does not have ***hyperparameters*** as it is a distance measure.

Elaborating on the results in Figure 50, we see that customer with user id 2.0 (highlighted in red) is getting their top three recommended products according to the quiz they have answered. We see Product 1, which is, "C The Different Treatment Eye Mask", gives an **accuracy of 52.94%, which is very good.**

Meaning that this Yuty product "C The Different Treatment Eye Mask" can solve 52.94% of the user's problem with user id 2.0. We can recommend this product (Product 1) to the customer for better effectiveness as well as also recommend a corresponding co-purchased product for maximum effectiveness which is "Vitamin C Prepping Tonic" (Product 2) with a second highest accuracy of 36.49% selling the concept of "maximum effectiveness".

	user_id	top_3_products_recommended	top_3_products_recommended_percentage
0	2.0	(C The Difference Treatment Eye Masks, Vitamin C Prepping Tonic, DOYENNE! MIRACLE FACE SERUM)	(52.9412, 32.1678, 4.2813)
1	34.0	(Vitamin A Night Shift Repair Cream, Vitamin B Miracle Serum, Vitamin C Prepping Tonic)	(40.0, 36.4964, 32.1678)
2	51.0	(Enzyme Gel Cleanser, Treat, HYDRATING GEL CREAM)	(1.3874, 0.9906, 0.073)
3	342.0	(Vitamin A Night Shift Repair Cream, Vitamin C Prepping Tonic, Vitamin C 20% Super Serum)	(40.0, 32.1678, 28.0543)
4	359.0	(Vitamin A Night Shift Repair Cream, Treat, HYDRATING GEL CREAM)	(40.0, 0.9906, 0.0716)
5	361.0	(DOYENNE! MIRACLE FACE SERUM , GLYCOLIC INTENSIVE MASQUE, Cleanse)	(3.5883, 1.1917, 1.0005)
6	362.0	(Vitamin C Prepping Tonic, Treat, HYDRATING GEL CREAM)	(32.1678, 0.9906, 0.0716)
7	363.0	(Vitamin C Prepping Tonic, GREEN TEA ANTIOXIDANT FACE MASK, Treat)	(32.1678, 5.3631, 0.9906)
8	371.0	(Vitamin C Prepping Tonic, Treat)	(32.1678, 1.4649)
9	380.0	(Vitamin C Prepping Tonic, Treat, HYDRATING GEL CREAM)	(32.1678, 1.4586, 0.0716)
10	386.0	(Vitamin C Prepping Tonic, Treat, HYDRATING GEL CREAM)	(32.1678, 0.9906, 0.0721)
11	407.0	(Vitamin B Miracle Serum, Treat, HYDRATING GEL CREAM)	(36.4964, 0.9906, 0.073)
12	409.0	(Vitamin A Night Shift Repair Cream, Vitamin C Prepping Tonic, DOYENNE! MIRACLE FACE SERUM)	(40.0, 32.1678, 3.5883)
13	412.0	(Vitamin A Night Shift Repair Cream, Treat, Enzyme Gel Cleanser)	(40.0, 1.4649, 1.3733)
14	413.0	(Vitamin A Night Shift Repair Cream, Treat, HYDRATING GEL CREAM)	(40.0, 1.3548, 0.0734)
15	414.0	(Vitamin C Prepping Tonic, Treat)	(32.1678, 1.3548)

Figure 50: Cosine Similarity results, Top three recommendations for the first 15 users who have answered the quizzes by Yuty company.

Extra Research & Output:

There is also another interesting output Figure 51 where we get all the top recommended products for Yuty customers out of curiosity for understanding the flow and ingredients working. For example user with id 668(Highlighted in red) has 4 best product recommendation depicted in the 2nd column

and its percentage in 3rd column and our top 3 products among them in the 4th column and its percentage in 5th column. ***This is a great insight that can help Yuty improve its product quality and work on its ingredients for better recommendations.***

	user_id	all_products_recommended	all_products_recommended_percentage	top_3_products_recommended	top_3_products_recommended_percentage	Total Accuracy
1	34.0	(SLAY! REVEAL FACE SCRUB , Everyday Glow Serum, AFTER GLOW MIST , Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, BRIGHTER DAYS DAILY CLEANSER, Hydrating Serum, Facial Cleansing Gel, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(0.3816, 0.0236, 2.6971, 33.3333, 1.2407, 21.3439, 0.5328, 0.1493)	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, AFTER GLOW MIST)	(33.3333, 21.3439, 2.6971)	57.3743
61	668.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, BRIGHTER DAYS DAILY CLEANSER, Hydrating Serum)	(33.3333, 1.2407, 21.3439)	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER)	(33.3333, 21.3439, 1.2407)	55.9179
44	562.0	(Eye Mask, Everyday Glow Serum, AFTER GLOW MIST , Vitamin C Granular Exfoliator, Eye Serum, SLAY! REVEAL FACE SCRUB , Facial Cleansing Gel, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream, Hydrating Serum, Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner)	(8.9552, 0.0238, 2.6971, 2.6259, 2.4099, 0.3816, 0.5504, 1.2407, 0.1493, 21.3439, 25.6158)	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, Eye Mask)	(25.6158, 21.3439, 8.9552)	55.9149
52	640.0	(Real Rose Quartz Pink Facial Roller, BRIGHTER DAYS DAILY CLEANSER, Hydrating Serum, Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner)	(1.1417, 1.2407, 18.6813, 33.3333)	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER)	(33.3333, 18.6813, 1.2407)	53.2553
21	439.0	(DIVINE Cleansing Crème, Everyday Glow Serum, NEVAEH Hydrating Mist, SLAY! REVEAL FACE SCRUB , AFTER GLOW MIST , Eye Serum, 3 in 1 Intense Nutrition | Anti-Aging Cream, BRIGHTER DAYS DAILY CLEANSER, Facial Cleansing Gel, Hydrating Serum, Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner)	(4.1348, 0.0236, 0.1755, 0.3521, 2.6971, 2.4557, 0.1493, 1.2407, 0.5504, 21.3439, 25.6158)	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, DIVINE Cleansing Crème)	(25.6158, 21.3439, 4.1348)	51.0945
0	2.0	(Everyday Glow Serum, DIVINE Cleansing Crème, BRIGHTER DAYS DAILY CLEANSER, Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, SLAY! REVEAL FACE SCRUB , Hydrating Serum)	(0.0233, 4.1348, 1.2407, 25.6158, 0.3816, 21.3439)	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, DIVINE Cleansing Crème)	(25.6158, 21.3439, 4.1348)	51.0945
25	463.0	(Eye Mask, DIVINE Cleansing Crème, BRIGHTER DAYS DAILY CLEANSER, Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner)	(8.9552, 4.1348, 1.2407, 25.6158)	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Eye Mask, DIVINE Cleansing Crème)	(25.6158, 8.9552, 4.1348)	38.7058

Figure 51: Cosine Similarity results, <u>All best recommendations</u> for the first 6 users who have answered the quizzes by Yuty company.

9.2. K-means (By Shalini Nayak)

9.2.1. Background

K-means clustering is a simple and widely acclaimed approach for partitioning a data sample into K unique and non-overlapping clusters. This model follows a process called clustering, after vectorization because text can't be fitter to a machine learning model, we need a numeric value hence we are first converting it into vector then proceed with clustering. Clustering can be defined as an algorithm which groups data points or samples based on a specific feature or attribute and forms several" clusters" known as k clusters.

The ***reason for choosing this model is because***, K-means ***is a very popular and effectively adapts to smaller data samples, and it gives an average value of the feature over the cluster which is very useful in calculating the best product recommendations for users.*** (Al-Masri, 2019)

We are getting accuracy of ~53% on the model taken on the prediction analysis for the total number of users which is not very good compared to the other models.

```
Accuracy of K-means Model:  21.455369230769243
```

Figure 52: Accuracy of K-means model.

9.2.2. Steps/Procedure

The working of K-means is, firstly it takes the data from the 8 Yuty datasets including two JSON files, secondly, it undergoes a training phrase where the data are grouped into clusters, and finally, the result is calculated by K-means by utilizing the distance from different data points to calculate the similarity ratio, based on the mean value.

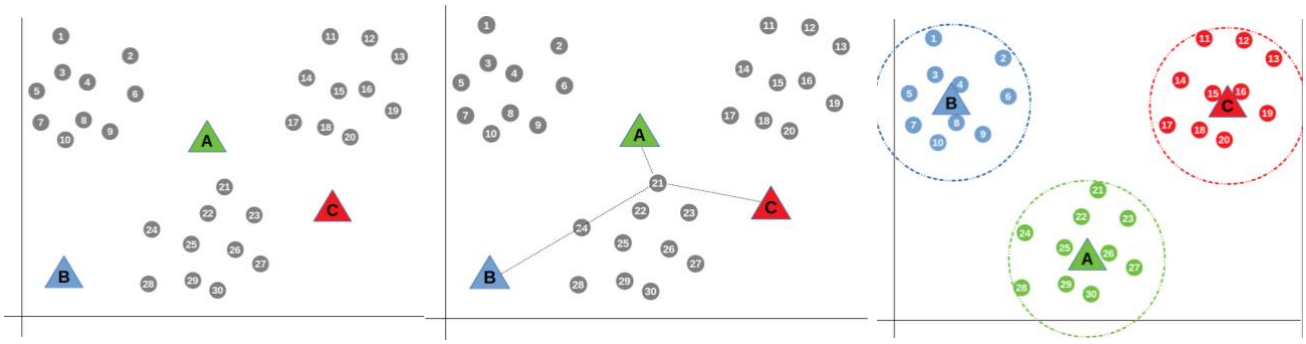


Figure 53: Working of the k-means algorithm with data samples.

Source: Towards Data Science(Al-Masri, 2019)

Step 1: K-means assign different average values which are also called means(centroids) to the data point. Now the algorithm goes through all the data samples, data points, calculating its nearest Euclidean distance or Manhattan distance from the centroids then group with its closest centroid measuring closest in distance making an association or groups called as K – group figure 53.

Step 2: Once the association has been made between the data points and the centroid, the machine learning algorithms runs again and re-calculates the mean value of the centroid.

Step 3: The new centroid's mean value is calculated by adding the values of all data points associated with it and dividing it with number of data points in that association belong to that group.

In our machine learning model, we have performed K-means clustering Figure 54 by:

- Firstly we have specified the desired number of clusters K in our case it is 381 clusters because Yuty has 381 products in its catalogue; then the K-means algorithm will assign each data sample acquired from customers quiz to exactly one of the 381 clusters.

- The K-means algorithm then clusters data by trying to separate samples in n number of groups of equal variance and "minimizing a criterion known as the inertia or within-cluster sum-of-squares". (Loukas, 2020)
- The K-means algorithm aims to choose centroid that minimise the inertia, or within-cluster sum-of-squares criterion. (Loukas, 2020)

```
(0, 1200)      1
(0, 110)       1
(0, 95)        1
(0, 810)       1
(0, 1036)      3
(0, 375)       1
(0, 1014)      1
(0, 374)       1
(0, 1206)      1
(0, 1430)      1
(0, 1122)      1
(0, 666)       1
(0, 141)       1
(0, 1555)      1
(0, 643)       1
(0, 673)       1
(0, 54)        1
(0, 982)       1
(0, 397)       1
(0, 1432)      1
(0, 536)       1
(0, 1040)      1
(0, 208)       1
(1, 817)       1
(1, 632)       1
:      :
(331, 1036)    3
(331, 375)     1
(331, 1014)    1
(331, 374)     1
(331, 560)     1
(331, 587)     1
(331, 316)     1
(331, 245)     1
(331, 1095)    1
(331, 1362)    1
(331, 239)     2
(331, 1344)    1
(331, 1549)    1
(331, 604)     1
(331, 918)     1
(331, 757)     1
(331, 267)     1
(331, 1030)    1
(331, 1609)    2
(331, 735)     1
(331, 174)     1
(331, 1143)    1
(331, 919)     1
(331, 1398)    1
(331, 1401)    1
```

Figure 54; Python output of how clustering looks with your Yuty data. 381 clusters as there are 381 products.

9.2.3. Result Analysis

In this part, we have analysed the result and accuracy though K-means algorithm. K-means s we know calculates accuracy by training the data points to nearest mean or centroid. In our analysis it focusses on result accuracy of recommendation.

From the table below we you can observe that we were able to achieve the top three recommendations for the users who have answers the quizzes by Yuty company. Every quiz answer is mapped to set of ingredients, so these ingredients are mapped to the ingredients which are contained in the different products made by Yuty through unsupervised machine learning algorithm calculating the closest distance.

In our analysis ***Hyperparameter*** used are n_clusters and random_state.

Elaborating on results in Figure 15, we see that customer with user id 34.0 (highlighted in red) is getting its top three recommended products according to the quiz they have answered. We see Product 1 which is, ***"Kokoa's Orange & Juniper Berry Brightening and Skin clearing toner"*** gives an accuracy of 26.19% which is moderate.

Meaning that this ***""Kokoa's Orange & Juniper Berry Brightening and Skin clearing toner"" will solve 26.19% of the user's problem.*** We can recommend this product to the customer for better effectiveness as well as also show the corresponding co-purchased product with second best accuracy, which is "Hyderating Serum" which has an accuracy of 24.66%.

	user_id	top_3_products_recommended	top_3_products_recommended_percentage
0	2.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, DIVINE Cleansing Crème)	(24.6305, 21.3439, 2.9096)
1	34.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, AFTER GLOW MIST)	(26.1905, 24.6657, 3.112)
2	51.0	(BRIGHTER DAYS DAILY CLEANSER, Spring Essential Facial Oil, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(1.861, 0.2798, 0.1991)
3	342.0	(Hydrating Serum, Eye Mask, AFTER GLOW MIST)	(21.3439, 9.9502, 3.112)
4	359.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(18.6813, 1.861, 0.0236)
5	361.0	(AFTER GLOW MIST , BRIGHTER DAYS DAILY CLEANSER, SLAY! REVEAL FACE SCRUB)	(3.112, 1.861, 0.2726)
6	362.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.861, 0.0236)
7	363.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(24.6657, 1.861, 0.0236)
8	371.0	(Hydrating Serum, Facial Cleansing Gel, Everyday Glow Serum)	(21.3439, 0.6604, 0.0232)
9	380.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 2.2869, 0.0236)
10	386.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.861, 0.0238)
11	407.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(26.1905, 1.861, 0.1991)
12	409.0	(Hydrating Serum, AFTER GLOW MIST , BRIGHTER DAYS DAILY CLEANSER)	(21.3439, 3.112, 1.861)
13	412.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Eye Mask)	(18.6813, 1.861, 1.2137)
14	413.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(18.6813, 1.861, 0.2001)
15	414.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.861, 0.0208)

Figure 55: K-means result, top three recommendations for the first 15 users who have answered the quizzes by Yuty company.

9.3. Random Forest Classifier (By Shrey Agarwal) (2nd Best Model)

Random Forest Classifier is a famous classification model which usually deals with high dimensional noisy text data. It keeps Bootstrap Aggregation and features randomness as the fundamental basis for choosing which tree provides the best results for classification. An RF model comprises a set of decision trees each of which is trained using random subsets of features. These decision trees are very sensitive to the training data and even a slight difference in the pre-processing can really change the way these. The researchers suggest that Random Forest integrated with Semantics demonstrates the superior performance of the proposed approach in textual information retrieval and initiates a new direction of research to utilize the interpretability of the classifier. (Islam et al, 2019). Figure 56 shows an example of how a decision tree works.

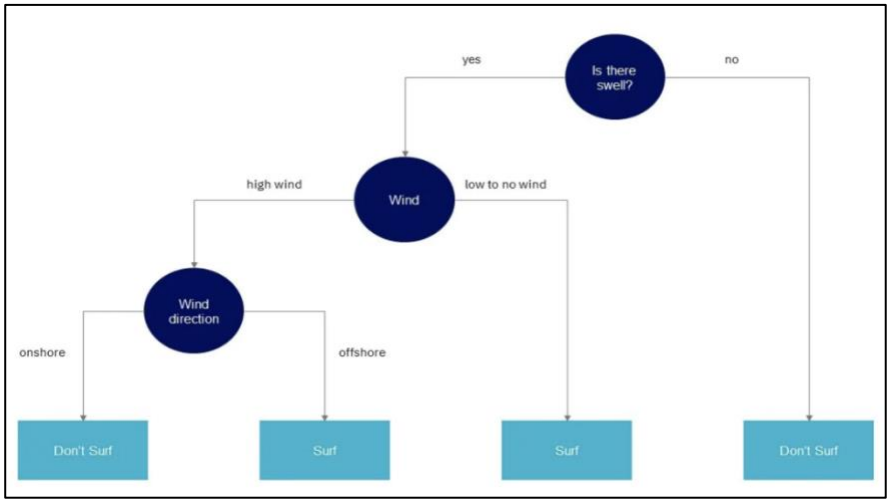


Figure 56 : An example of a RF classification model, with labels 'surf' or 'don't surf' (IBM Cloud Education, 2020)

9.4. Multinomial Naïve Bayes (By Shrey Agarwal)

Popular in Natural Language Processing is the Bayesian learning method known as the Multinomial Naive Bayes algorithm (NLP). Using the Bayes principle, the computer makes an educated prediction about the label of a text, which in our case is the product category. It determines the likelihood of each label for a particular sample and outputs the label with the highest likelihood.

We determine the chance of class A when predictor B is available. Its foundation is the following equation: **P(A|B) = P(A) * P(B|A)/P (B)**. (upGrad blog, 2021)

9.5. Procedure - Random Forest Classifies & Multinomial Naïve Bayes

We have imported both Random Forest Classifier and Multinomial Naive Bayes from the sklearn package, which is a machine-learning library for Python that provides simple and efficient tools for data analysis and data mining, with a focus on machine learning.

This model is fit and transformed with our training features and dependent variable and then used to predict the values on our testing data.

At first, all the independent variables were used in the modelling. Metrics like accuracy, F1 score, and confusion matrix are used to evaluate the results and compare them to choose the best combination of features.

$$F_1\text{−score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})}$$

9.6. Result Analysis - Random Forest Classifies & Multinomial Naïve Bayes

This part focuses on the model performance. We have used four different methods for creating features. These models are implemented and iterated as we learn to better understand the text data. These are the results:

RF	TFIDF feature	Count feature	Text as feature	GloVe feature
Accuracy	0.3	0.28	0.2	0.16
F1 score	0.36	0.34	0.23	0.0

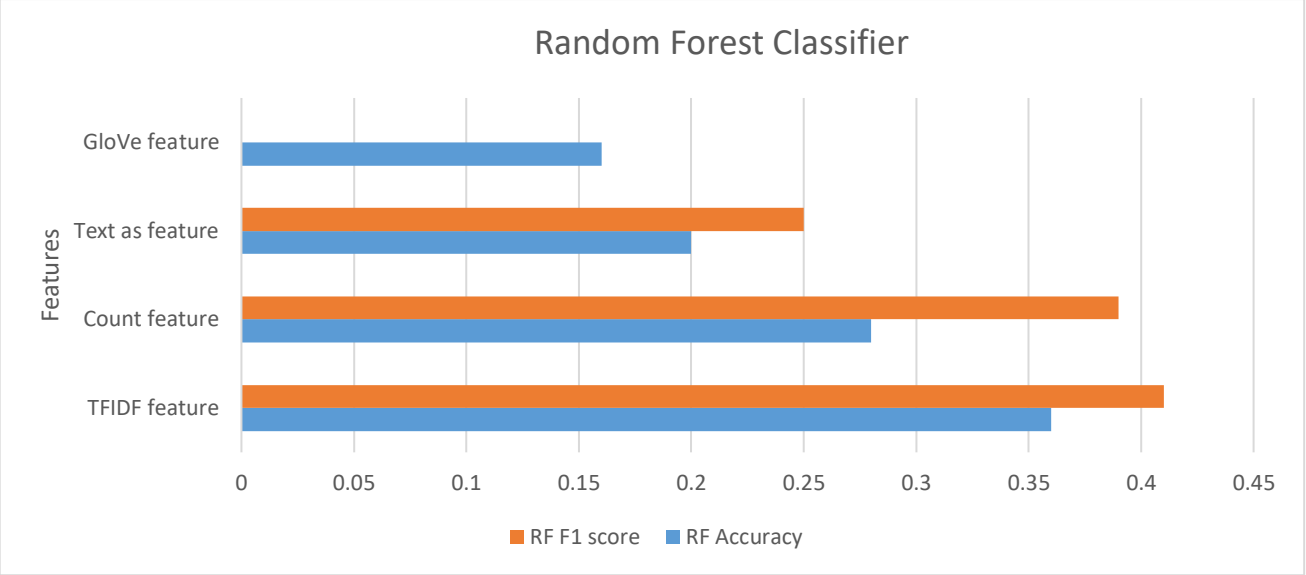


Figure 57: Results of Random Forest model

After tuning different parameters, these are the best results for each type of feature method.

We can see that the TFIDF feature provides the best accuracy and F1 score as compared to other methods. One biggest limitation is that these models are trained in sub-sampled labelled data and models like RF require a lot of data to provide a good prediction.

NB	TFIDF feature	Count feature	Text as feature	GloVe feature
Accuracy	0.2	0.28	0.2	0.16
F1 score	0.33	0.31	0.25	0

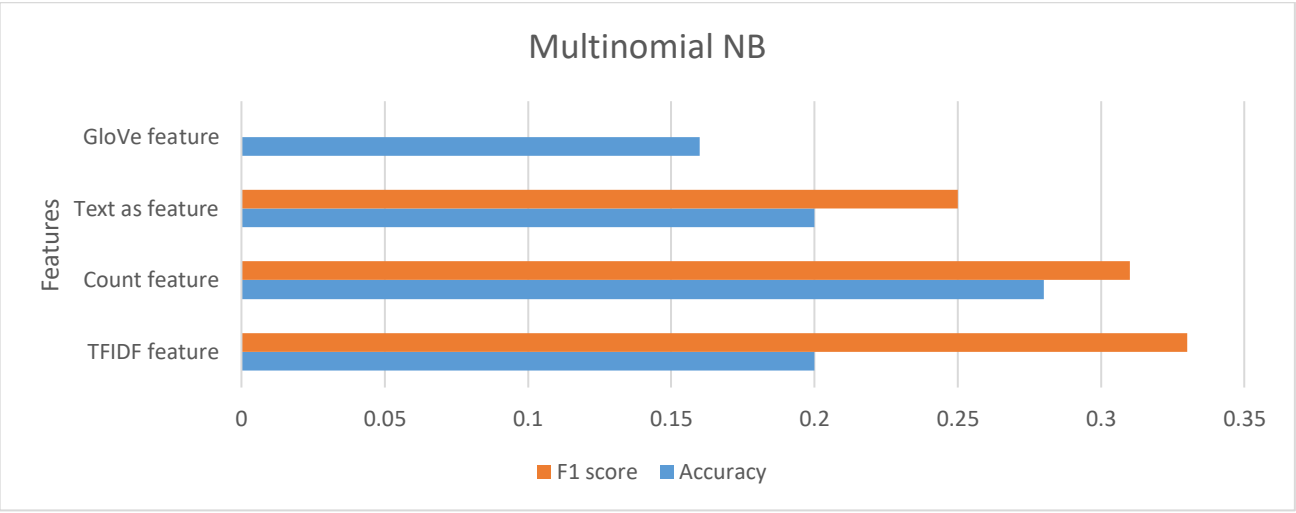


Figure 58:Results of Multinomial NB model

Here, the Count feature showed the best results for the Multinomial model.

One reason for the low accuracy in this model is that it assumes that all predictors (or features) are independent, which rarely happens in real life.

After analysing all 4 ML models, we concluded that Cosine Similarity and Random Forest Classifier showed the best results. Therefore, we will go further with these 2 models for making our recommendation model to match customers of Yuty to the right products based on the survey.

10. Results: Possible Findings? (By Neel Raut)

10.1. Result Overview, Result Table & Best Model

In this project, which was given to us by Yuty, we worked on the eight datasets and were able to build four models of a recommendation system that recommended a list of products for each of the Yuty customers.

The first steps that we took for our project were understanding the dataset given to us by Yuty and to do some exploratory data analysis regarding the data. We firstly identified the number of missing values in each of the CSV data files and presented them using graphs. Secondly, we used the bag of words method to identify the most recurring words. We applied this method to the ingredient variable of the product_catalogue_data.CSV dataset. Due to this, we were able to identify the most repeated ingredients in all the products that are offered by Yuty. Refer to points 5.0 and 8.0 for a detailed explanation.

The next part of the project to build a recommendation system was to clean and pre-process the data provided by Yuty. Firstly, we dropped the rows which contained missing values. We also used Natural Language Processing to clean the most important CSV file for our project – product_catalogue_data.csv. We dropped a few columns that were not useful to us. We used NLP to remove punctuations, lowercase a few words, and remove stop words. We also used MS Excel to help clean the data further. Refer to point 6.0 for a detailed explanation.

After this, we used feature engineering to create features for our ML models. These features help to decide how efficient our recommendation system is. The better the features, the better the accuracy. Refer to point 7.0 for a detailed explanation.

Once our features were decided and our data was clean, the next step that we took was the implementation of the 4 machine learning models to build our recommender. The results gained from these are explained below:

K-means: The first machine learning model that we have implemented for our project is the K-means clustering technique. K-means clustering makes use of unsupervised learning to make clusters. Clustering is a machine learning task that is used to group data that are similar to each other. We were able to use the K-means algorithm in the dataset provided by Yuty for product recommendation. Using this algorithm, we were able to attain an accuracy of 18.53% for the recommendations that were provided to the Yuty customers. K-means clustering algorithm works best when the dataset is huge. Since the dataset given to us by Yuty was small, we couldn't fully utilise the capability of K-means

and hence the accuracy attained was a bit underwhelming (Garbade, 2018). Refer to point 9.1 for a detailed explanation regarding K-means clustering.

Random Forest Classifier: The second machine learning technique we used to recommend products to the users was the Random Forest Classification method. Random forest uses classification and regression to build a different range of decision trees when the data is being trained (Wikipedia Contributors, 2019b). "*A decision tree is a predictive model which is used to conclude a set of data or observation"* (Wikipedia Contributors, 2019a). With this machine learning technique, we attained an accuracy of 28.30%. Although random forest is one of the best machine learning techniques for recommendation systems, it still needs a large number of resources to provide a good recommendation. Since the dataset was small and didn't have many hyperparameters, we couldn't get higher accuracy levels (Mbaabu, 2020). Refer to point 9.3 for a detailed explanation regarding Random Forest.

Multinomial Naïve Bayes: The penultimate technique that we have implemented is the Multinomial Naïve Bayes machine learning method. "*This NLP technique makes use of the Bayes theorem, which deals with probability as the likelihood that data belongs to a specific class*" (Ratz, 2021). Using this method, we were able to get an accuracy score of 28.70. A larger dataset of textual data would have helped this algorithm to gain a higher level of efficiency. Refer to point 9.4 for a detailed explanation regarding Multinomial Naïve Bayes.

Cosine Similarity: The final machine learning method we implemented was the Cosine Similarity method. For our recommendation system, this method yielded the best results. Cosine similarity works on the concept of finding the smallest distance between two vectors. In our case, vectors are generated using text values of product ingredients. The lesser the distance between two vectors, the more similar the vectors are to each other. Subsequently, using this method, we attained an accuracy of 52.03%. Refer to point 9.2 for a detailed explanation regarding cosine similarity (Sciencedirect.com, 2019).

The following bar chart (Figure 60) and table (Figure 59) give a visual representation of all the models that we have implemented alongside their accuracy. They are in descending order of accuracy.

Model Number	Model Name	Accuracy Attained
Model 1(Best Model)	Cosine Similarity	52.03%
Model 2	Random Forest	36.08%
Model 3	Multinomial Naïve Bayes	28.00%
Model 4	K-means Clustering	21.45%

Figure 59: List of tables alongside their accuracy

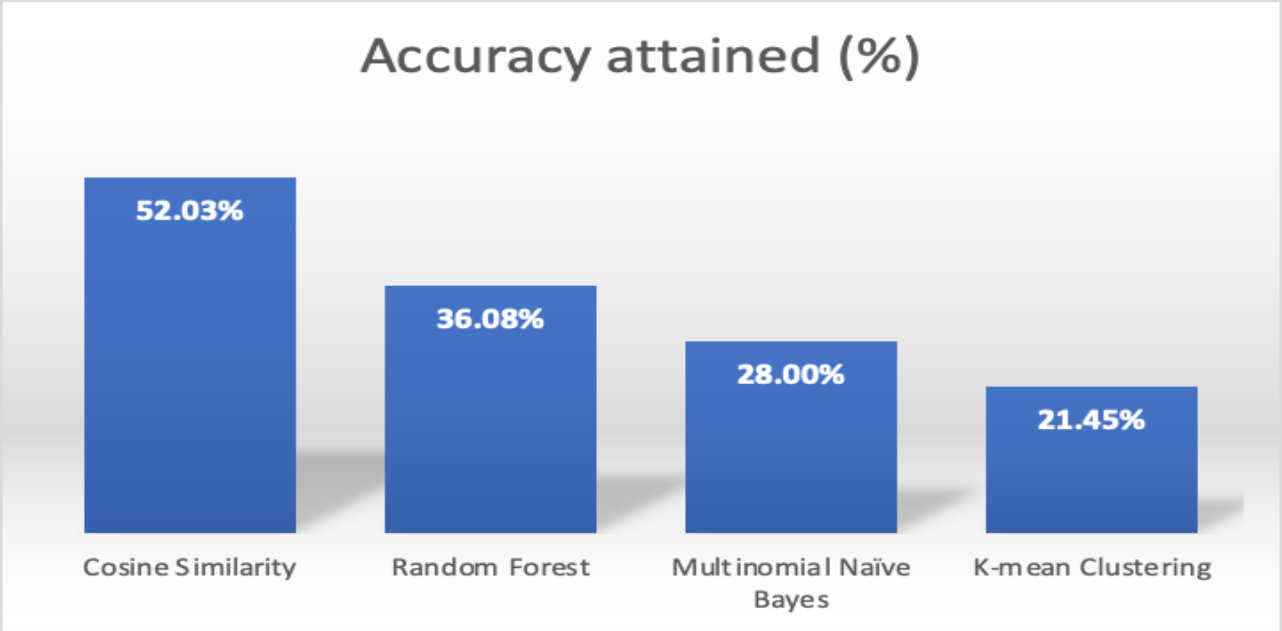


Figure 60: Bar chart showing the accuracies of all the four models

The accuracy we got from the cosine similarity method was the highest among all the machine learning techniques we implemented. Hence this model is a clear winner and will be used to predict customer's product recommendations. Finally, our final recommendation for all the 65 Yuty customers using our best machine learning model is depicted in Output 1:

	user_id	top_3_products_recommended	top_3_products_recommended_percentage
0	2.0	(C The Difference Treatment Eye Masks, Vitamin C Prepping Tonic, DOYENNE! MIRACLE FACE SERUM)	(52.9412, 32.1678, 4.2813)
1	34.0	(Vitamin A Night Shift Repair Cream, Vitamin B Miracle Serum, Vitamin C Prepping Tonic)	(40.0, 36.4964, 32.1678)
2	51.0	(Enzyme Gel Cleanser, Treat, HYDRATING GEL CREAM)	(1.3874, 0.9906, 0.073)
3	342.0	(Vitamin A Night Shift Repair Cream, Vitamin C Prepping Tonic, Vitamin C 20% Super Serum)	(40.0, 32.1678, 28.0543)
4	359.0	(Vitamin A Night Shift Repair Cream, Treat, HYDRATING GEL CREAM)	(40.0, 0.9906, 0.0716)
5	361.0	(DOYENNE! MIRACLE FACE SERUM , GLYCOLIC INTENSIVE MASQUE, Cleanse)	(3.5883, 1.1917, 1.0005)
6	362.0	(Vitamin C Prepping Tonic, Treat, HYDRATING GEL CREAM)	(32.1678, 0.9906, 0.0716)
7	363.0	(Vitamin C Prepping Tonic, GREEN TEA ANTIOXIDANT FACE MASK, Treat)	(32.1678, 5.3631, 0.9906)
8	371.0	(Vitamin C Prepping Tonic, Treat)	(32.1678, 1.4649)
9	380.0	(Vitamin C Prepping Tonic, Treat, HYDRATING GEL CREAM)	(32.1678, 1.4586, 0.0716)
10	386.0	(Vitamin C Prepping Tonic, Treat, HYDRATING GEL CREAM)	(32.1678, 0.9906, 0.0721)
11	407.0	(Vitamin B Miracle Serum, Treat, HYDRATING GEL CREAM)	(36.4964, 0.9906, 0.073)
12	409.0	(Vitamin A Night Shift Repair Cream, Vitamin C Prepping Tonic, DOYENNE! MIRACLE FACE SERUM)	(40.0, 32.1678, 3.5883)
13	412.0	(Vitamin A Night Shift Repair Cream, Treat, Enzyme Gel Cleanser)	(40.0, 1.4649, 1.3733)
14	413.0	(Vitamin A Night Shift Repair Cream, Treat, HYDRATING GEL CREAM)	(40.0, 1.3548, 0.0734)
15	414.0	(Vitamin C Prepping Tonic, Treat)	(32.1678, 1.3548)

16	420.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0208)	22.6054
17	426.0	(Hydrating Serum, DIVINE Cleansing Crème, BRIGHTER DAYS DAILY CLEANSER)	(21.3439, 4.1348, 1.2407)	26.7194
18	429.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, SLAY! REVEAL FACE SCRUB)	(21.3439, 1.2407, 0.3704)	22.9550
19	437.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0233)	22.6079
20	438.0	(BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream, Everyday Glow Serum)	(1.2407, 0.1493, 0.0211)	1.4111
21	439.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, DIVINE Cleansing Crème)	(25.6158, 21.3439, 4.1348)	51.0945
22	445.0	(BRIGHTER DAYS DAILY CLEANSER, SLAY! REVEAL FACE SCRUB , Everyday Glow Serum)	(1.2407, 0.3816, 0.0238)	1.6461
23	448.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, SLAY! REVEAL FACE SCRUB)	(21.3439, 0.8184, 0.3464)	22.5087
24	449.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0233)	22.6079
25	463.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Eye Mask, DIVINE Cleansing Crème)	(25.6158, 8.9552, 4.1348)	38.7058
26	469.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, AFTER GLOW MIST)	(18.6813, 1.2407, 0.6775)	20.5995
27	483.0	(DIVINE Cleansing Crème, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(4.1348, 1.2407, 0.1493)	5.5248
28	485.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, DIVINE Cleansing Crème)	(18.6813, 1.2407, 0.6912)	20.6132
29	497.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0205)	22.6051
30	499.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(18.6813, 1.2407, 0.0232)	19.9452
31	500.0	(AFTER GLOW MIST , BRIGHTER DAYS DAILY CLEANSER, SLAY! REVEAL FACE SCRUB)	(2.6971, 1.2407, 0.3816)	4.3194
32	503.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0233)	22.6079
33	507.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0233)	22.6079
34	522.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, DIVINE Cleansing Crème)	(21.3439, 1.2407, 0.6912)	23.2758
35	542.0	(AFTER GLOW MIST , BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(2.6971, 1.2407, 0.1493)	4.0871
36	543.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(33.3333, 1.2407, 0.15)	34.7240
37	547.0	(Eye Serum, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(2.4263, 1.2407, 0.1493)	3.8163
38	548.0	(Hydrating Serum, AFTER GLOW MIST , BRIGHTER DAYS DAILY CLEANSER)	(18.6813, 2.6971, 1.2407)	22.6191
39	550.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, AFTER GLOW MIST , BRIGHTER DAYS DAILY CLEANSER)	(25.6158, 2.6971, 1.2407)	29.5536
40	552.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0208)	22.6054
41	554.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Spring Essential Facial Oil)	(18.6813, 1.2407, 0.2798)	20.2018
42	558.0	(Hydrating Serum, AFTER GLOW MIST , BRIGHTER DAYS DAILY CLEANSER)	(21.3439, 2.6971, 1.2407)	25.2817
43	561.0	(BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream, Everyday Glow Serum)	(1.2407, 0.1493, 0.0233)	1.4133
44	562.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, Eye Mask)	(25.6158, 21.3439, 8.9552)	55.9149
45	568.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, BRIGHTER DAYS DAILY CLEANSER, Eye Mask)	(25.6158, 1.2407, 1.0767)	27.9332
46	597.0	(BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream, Everyday Glow Serum)	(1.2407, 0.1493, 0.0212)	1.4112
47	601.0	(Hydrating Serum, DIVINE Cleansing Crème, BRIGHTER DAYS DAILY CLEANSER)	(21.3439, 4.1348, 1.2407)	26.7194
48	610.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(25.6158, 1.2407, 0.0233)	26.8798
49	615.0	(BRIGHTER DAYS DAILY CLEANSER, DIVINE Cleansing Crème, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(1.2407, 0.6912, 0.1493)	2.0812
50	636.0	(DIVINE Cleansing Crème, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(4.1348, 1.2407, 0.1493)	5.5248
51	639.0	(BRIGHTER DAYS DAILY CLEANSER, Facial Cleansing Gel, SLAY! REVEAL FACE SCRUB)	(1.2407, 0.5356, 0.3816)	2.1579
52	640.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER)	(33.3333, 18.6813, 1.2407)	53.2553
53	651.0	(Hydrating Serum, DIVINE Cleansing Crème, BRIGHTER DAYS DAILY CLEANSER)	(18.6813, 4.1348, 1.2407)	24.0568
54	654.0	(BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream, Everyday Glow Serum)	(1.2407, 0.1493, 0.0212)	1.4112
55	657.0	(BRIGHTER DAYS DAILY CLEANSER, DIVINE Cleansing Crème, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(1.2407, 0.6912, 0.1493)	2.0812
56	658.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0233)	22.6079
57	660.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(12.523, 1.2407, 0.1493)	13.9130
58	661.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(18.6813, 1.2407, 0.0209)	19.9429
59	663.0	(Hydrating Serum, Eye Mask, AFTER GLOW MIST)	(18.6813, 8.9552, 2.6971)	30.3336
60	664.0	(Hydrating Serum, AFTER GLOW MIST , Eye Serum)	(21.3439, 2.6971, 2.446)	26.4870
61	668.0	(Kokoa's Orange & Juniper Berry Brightening & Skin Clearing Toner, Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER)	(33.3333, 21.3439, 1.2407)	55.9179
62	670.0	(Eye Serum, BRIGHTER DAYS DAILY CLEANSER, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(2.4446, 1.2407, 0.1493)	3.8346
63	671.0	(BRIGHTER DAYS DAILY CLEANSER, DIVINE Cleansing Crème, 3 in 1 Intense Nutrition | Anti-Aging Cream)	(1.2407, 0.6912, 0.1493)	2.0812
64	680.0	(Hydrating Serum, BRIGHTER DAYS DAILY CLEANSER, Everyday Glow Serum)	(21.3439, 1.2407, 0.0212)	22.6058

Output 1: Result of Cosine Similarity, which was used for a product recommendation for Yuty customers.

As you can see in the above image, we see the user_id, the names of the top three recommended products for that user, and the corresponding percentage of the problem solved by that particular Yuty product. For example, User number 2 is recommended three products – C The Difference Treatment Eye Masks (52.9414), Vitamin C Prepping Tonic (32.1678), and Doyenne Miracle Face Serum (4.2813). Hence, we can confidently say that according to the recommendations of our cosine similarity model, for user 2.0, 52.9414% of the problems regarding their skin treatment are solved by "C The Difference Treatment Eye Masks". Similarly, only 32.1678% of user 2.0's problem is solved by Vitamin C prepping Tonic.

10.2. Overall Result Summary & Business Insights

In the modern world, people have many options when it comes to anything and everything. The beauty products market is no exception. To stay ahead of their competitors, companies must employ certain strategies to differentiate them from their competitors. Hence, for Yuty to stay ahead of its competitors, they need to target increasing customer satisfaction.

Customer satisfaction leads to higher customer retention rates (gocardless.com, n.d.). "*Customer retention rate can be defined as the percentage of customers who will stay your customer after a given period."* (Salesforce.com, n.d.). It's difficult for any business to acquire new customers as this process depletes a lot of companies' resources, such as time and money. However, it's much easier for businesses to hold on to their existing customers as fewer resources are depleted in this process (Salesforce.com, n.d.).

Companies employ different strategies to increase their customer retention rates. One of the strategies that companies employ is building a recommendation system that recommends relevant products to their customer. The best example of a company that has successfully used a recommendation system to ensure higher sales, higher customer satisfaction, and higher customer retention rate is Amazon. Amazon's recommendation system in their online store recommend personalised products for all their customers. These recommended products construct up to 35% of the revenue generated by Amazon (Miquido, 2020).

The recommendation system we have built for Yuty will help increase the customer retention rate for Yuty. A higher customer retention rate will lead to higher sales revenue, higher profit margins, and a higher number of new customers coming to Yuty to use their services.

However, the data provided by Yuty wasn't rich enough to build the best version of our recommendation system. Richer data would have led to richer accuracies for the product recommended by the recommendation system for Yuty customers.

To enhance the richness of the dataset, Yuty can offer their customers coupon codes or hampers for completing a questionnaire that will help increase the recommendation system's efficiency. Yuty can also build a mobile app as most customers prefer using their smartphones for beauty product shopping.

Suppose all these recommendations are followed by Yuty. In that case, we can confidently say that Yuty will have a very high number of loyal customers, leading to higher sales revenue and profit margins for Yuty.

11. Limitations & Challenges (By: Anshul Basotia)

The project has certain limitations which make a difference in results, accuracy, and time allocation.

- A recommendation model has few limitations when it comes to recommending a product to a customer, recommendation system has the limitation of scalability of algorithms with data sets which are being used in the real world, as user interactions keep on changing in form of reviews which creates a huge problem for the recommendation model.
- There are possible outcomes that customers wouldn't agree with the product which is recommended by the model. From the dataset, we can also figure out that the structure is not conventional, which automatically results in too many anomalies. The structure is important for a pre-defined model, text files.
- The limitation of human preference needs to be considered as well, whenever a product or artificial application is recommended to a consumer or user, there is always an error of human preferences, there are many people who consider emotions over logic, and that results in the data models not being accurate even if the models are.
- The sub-sample data in the project had 100 rows which is insufficient to get accurate results. In this case, *a sufficient or large amount of data can help train the particular models* which are available for this project.
- The *dataset has less data* which also creates machine learning issues, dataset also had no features which had to be created by us as a group which made it time-consuming.
- The limitation of having certain words in the dataset not matching from the product catalogue to skin.json is showing data inconsistency as well.
- Consumers when deciding on a product recommended to them cannot be accurate because of the psychological need to think, feel and use the products which unfortunately cannot be done on the online platform.

12. Conclusions (By Faran Saeed)

The core of this project was to improve the recommendations of product IDs for each customer based on their answers and requirements. This would lead to customers being recommended a product which fits them the most efficient way considering their skin tones, types, and requirements.

Why would we want this? Better recommendations mean positive customer feedback which in return would directly lead to an increased pool of happy clients increasing the overall turnover for the company and increasing the profits. The company's vision goes beyond just having a satisfied client base. The addition of ML and AI algorithms also helps the company work on the products which are needed and in demand thus saving a lot of wastage and proving healthy for the environment.

We started with understanding the datasets provided by the company. Then followed data pre-processing where our goal was to create an accurate text classification model. Then we moved on to feature engineering where we used the raw datasets to create flat structures to train our Machine Learning Models. Then followed the Exploratory Data Analysis where we highlighted and then handled the missing values in order to run ML models.

Referring to machine learning methodology section in point 9. We were able to achieve the result for product recommendation to Yuty customers by the Cosine Similarity algorithm (Point 9.1.1) which gave an accuracy of ~52% using the concept of vectorisation which is the highest among all four models used in this project.

From Figure below you can see that customer for example with user id 2.0 were able to achieve good solutions from the top recommended product or could also be recommended combined solution of top two Yuty products for better effectiveness.

	user_id	top_3_products_recommended	top_3_products_recommended_percentage
0	2.0	(C The Difference Treatment Eye Masks, Vitamin C Prepping Tonic, DOYENNE! MIRACLE FACE SERUM)	(52.9412, 32.1678, 4.2813)

And the accuracy was good for almost ¾ of the 65 Yuty customers who took the test which depicts huge positivity in customer recommendation and great future scope if Yuty decides to take its research to higher level which could be done with bigger datasets and more advanced machine learning algorithms.

Random Forest also has proved promising in terms of accuracy and recommendation elaborated in point 9.3. As being a decision tree Random Forest has also good chances of improvement when subjected to good data quality.

Overall, the project was exceptionally good learning where we not only got to know about the Machine Learning algorithms and their functions, but we also learned how to work within a team, forming a synergy and growing together by adopting a sense of consideration within a team.

13. Recommendations to Yuty (By Faran Saeed)

From our extensive research, understanding of the industry and analysis within the company, we would have the following recommendations for the company:

- Update and enhance the Machine Learning models with time as more data would lead to more accurate results and thus better predictions for the customers. We were still able to get a decent accuracy on our models, but it could greatly improve if we were provided with a larger dataset.
- Update the questionnaire for customers from time to time as well as a lot of times, a fad product might not stay in the market for a long time and thus it would over-shadow the latest trending products. This will also ensure that relevant data is obtained which can be used for a longer time period. A good example is Netflix updating its home page constantly with new content specifically tailored to its users.
- Company should also develop a mobile application for the business as most of the traffic today on the internet is through smartphones. An addition of an app will hugely increase the target market for the business.
- There should be a contingency plan in place for customers who are not satisfied with the prediction of the ML algorithm and thus a user-friendly feedback form shall be in place to understand why the prediction was not accurate. ***Most of the subscription-based companies like Shopify or Pret have feedback forms once user cancels the subscription.***
- To counter the uncertainty of not having a physical inspection of the products, the company can introduce small samples of the products recommended to the customers on low or no costs in order to increase customer satisfaction.
- With the increasing trend of skincare in men, company should also focus on hygiene and cosmetic products for men as this is an emerging and a huge market and can increase the total revenue of the company.

14. Business Insights, Benefits, Further Analysis & Future Business Enhancements (By Shalini Nayak)

14.1. Business Insights

In our analysis and results conducted by our machine learning algorithms on Yuty data, we discovered many important business insights that were inferred from them, which Yuty can implement in their business, they are:

- ***Tons of products go to waste in sampling*** when customers try them on themselves with little or no guarantee that will suit them, or they will purchase them. This directly affects cost and production, which the company must bear to retain customers and provide them with a personalised solution. This has effectively ***been solved by our analysis and modelling, Yuty uses our Cosine Similarity algorithm,*** which uses data samples or "live data" to test-train and learn about their customers and provide the best solutions for them. Thereby it could cut down sampling costs and reduce wastage.
- With our exploratory data analysis and data pre-processing, we have come to know that "***more data" is better for analysis in this type of project***. Apart from asking only 15 questions to customers in a quiz, Yuty should focus on a customised and interactive user interface to know their customers even better, to have better recommendations for them.
- Use ***recommendation strategies along with product recommendations Figure 61*** with machine learning. Yuty can use concepts like "suggested products", which makes customers 4.5x likely to checkout. (Anon,2020)

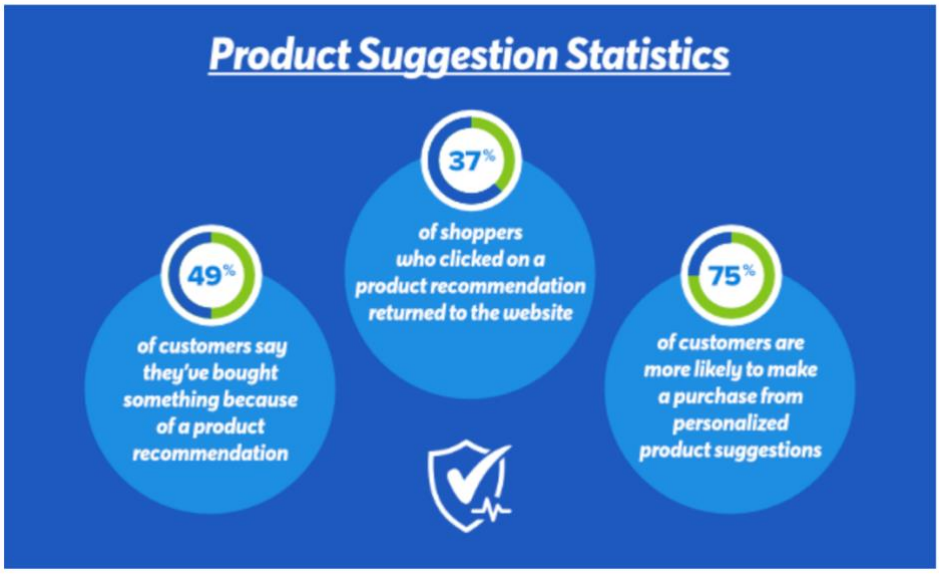


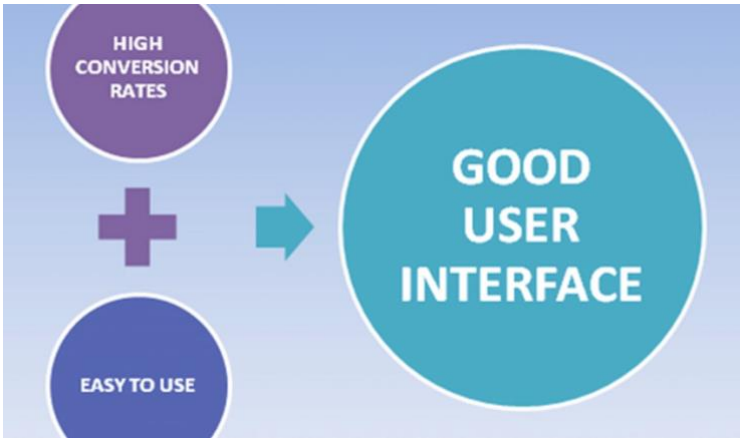
Figure 61: Product suggestion statistics.(Anon, 2020)

- Yuty can work on giving a good, personalised experience, it is reported that more than 80 % of customers go for that product or purchase when they get personalised feedback and suggestions.

14.2. Future Business Enhancements Guide

Yuty has a very good prospect for future scope of advancements- **if it targets offering personalised solutions to customers and focusses on more advanced data science and analysis research while focus on research planning, project goals and target customers**. Using our results and analysis, some of the implementations Yuty can make in their business to increase customer retention and maximize profit are:

- **Good UI – Web & App experience:** Having an app with a ***good user interface will encourage users to explore more and will increase conversion***, where customers can interact and get familiarised with Yuty products and see testimonials, is a very good way to build relationships and a loyal following because. In today's world, there are about 6.4 billion smartphone users. (Turner, 2021) so having an app is very crucial and it will also increase the brand name and loyalty.



- **Live product suggestions** using advanced machine learning and AI techniques. So that it becomes easy for customers to get recommendations, choose and purchase. These features should be embedded in both in-app and web applications.

- **Discounts, Gift cards & Loyalty points:** This is the most efficient way to attract customers. Yuty could lure leads, or potential customers into a sales funnel by first offering them a freebie,

discount or offers like ***"Answer a few questions, and get the best product recommendations for your skin with 20% off* on your first purchase ".*** And for old customers, email about new products, schemes and loyalty points could be a perfect approach for customer retention.

- **Targeted Marketing:** Yuty should venture into social media marketing and use Facebook and Instagram targeted advertisements where most of the young crowds interact. Also, according to world market share, the biggest market is in North America and Asia in Figure 62 below; Yuty should also try to take their products global by growing their manufacturing and warehouses.

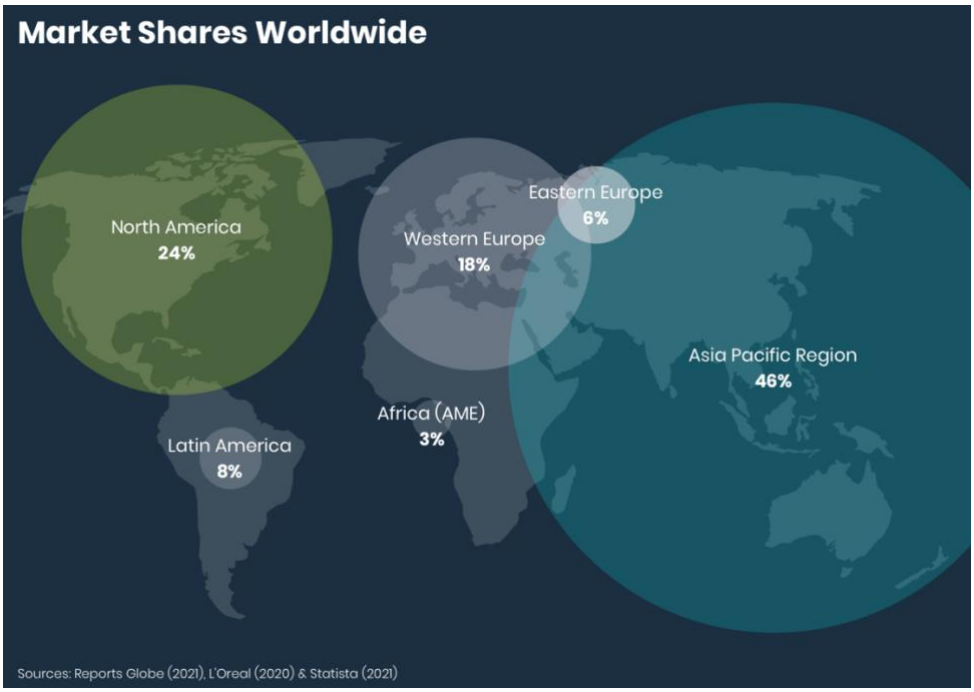


Figure 62: Beauty Industry Revenue. Website:(Roberts, 2022)

- **Logistics:** Yuty should have a full logistics service for manufacturing, inventories, order fulfilment, customer delivery, and return.

14.3. Ways Yuty Can Benefit by using Machine Learning.

1. **Sales Forecast and Product Marketing:** Machine learning will help in recognition of market patterns and trends rom large, small or diverse datasets effortlessly and gives excellent solutions which can help in Yuty decisions giving it a better edge while entering such a huge and saturated market.
2. **Better Customer Segmentation by Data Insights:** Datacoms from various sources like from leads, quizzes, emails, surveys, and even if it is relevant, they are unstructured data, ML is used to eliminate unnecessary data which helps to engage with potential customers better which can be later converted into leads.

3. **Recommending the right product to buys/customers:** ML for Product recommendation is widely used by various business worldwide for their business growth. More technical advancements and research can help grow Yuty very significantly and help it recommend better product to customers, some of them are discussed in 14.4 section.
4. **Better insights and cognitive capabilities:** Learning through the data, drawing conclusions from the data with ML algorithms leads to better and faster business insights.
5. **Speed and Efficiency:** Machine Learning increases computational abilities with data processing and offers a low error rate compared to humans. ML also decreases the chances of failures there by increase efficiency in business insights and decision making.
6. **Automation:** Yuty can automate its processes and analysis, which were done by humans before which can empower companies to focus on lucrative business decisions.

14.4. Technical Advancements Suggestions

- Keeping in mind Yuty goal and our data analysis, modelling and trials, in our opinion **XG Boost with labelled data is among the best approaches for recommendations**.
- Yuty should keep on improving **data quality and collection** and should stick to machine learning models for product predictions.
- Supervised learning is obviously better for deployment prediction but huge challenge is to collect labels for training and collected data shouldn't be biased to one particular sub community. So if there was a possibility to connect data labels **Yuty should will go with supervised learning method after collecting a good amount of data.**
- **Deep Learning and Computer Vision** – They can be useful for live data collection for cosmetics or makeup product as it uses live facial feature detection of customers for recommendation.
- Also, we can go further and study the implementation of **Neural Network like CNN or BERT which are very effective**, if Yuty manages to acquire huge amount of customer data in future.

15. Commercial Considerations (By: Anshul Basotia)

Commercial consideration generally means a gist of overall business costs and opportunities. Businesses need to consider price of their products, supply chain, how the product can be marketed and benefit the customers when they purchase or use the services.

Our analysis aims to increase the customer base to get more revenue and have an intangible asset of being reputable which automatically leads to customer being loyal to the company and multiple sales. Customer retention will make the sales more recurring which will help the company to build trust and loyalty in the hearts of customers.

For any business, the idea behind automation and increasing efficiency is to directly increase the turnover and then the profits for the company. To do so, a company would need to constantly grow its target market. While expanding into new geographical regions sounds like the easiest option, increasing the customer base within a specific territory is a more efficient way to increase profits. Thus, using our AI and ML analysis, the company would be able to make better predictions for its customers and will increase customer retention as more customers would be satisfied with their recommended products.

The company can also use this analysis to further strengthen their business processes. Doing so will not only save time and manpower but will also lead to saving huge costs in the long run. This analysis can be interpreted in several ways by the company executives to solve other complex problems which have either not surfaced yet or might have been faced in the past in a more effective manner.

Brands like Coty which is a multi-brand beauty company launched an application called “Let’s Get Ready” in 2018, this application can help customers create their own personalised look for the events they are attending. This application was created for Amazon’s first Echo devices with screens. Examples of events can be marriages, funerals, or cocktail parties. The figure 63 below shows Amazon Echo with screens used for the application ‘Let' Get Read’.

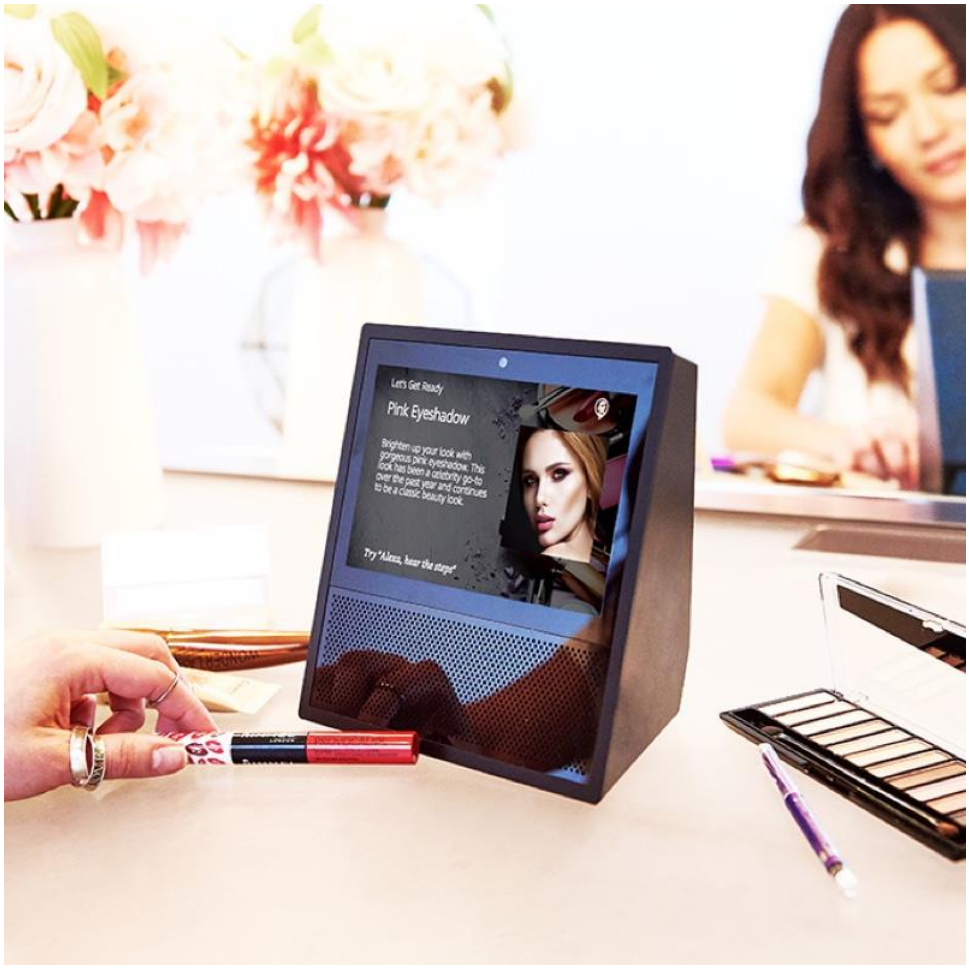


Figure 63: Coty application (www.cosmeticsbusiness.com, n.d.)

The application Let's get ready asks about the preferences of the looks the customers want, such as hair colour, skin colour and the event they are attending. The application used Alexa by Amazon to recommend products from the brand Coty's consumer portfolio. The application offers a personalised look having over 1000 combinations using the brand's recommended products. (theappsolutions.com, n.d.)

16. References

kaggle.com. (n.d.). *Recommender System using Un-supervised Learning*. [online] Available at: https://www.kaggle.com/code/basu369victor/recommender-system-using-un-supervised-learning/notebook

Kim, C. and Mauborgne, R. (2015). Red Ocean vs Blue Ocean. [online] Blue Ocean Strategy. Available at: https://www.blueoceanstrategy.com/tools/red-ocean-vs-blue-ocean-strategy/.

OpenAI. (2022). *Introducing Text and Code Embeddings in the OpenAI API*. [online] Available at: https://openai.com/blog/introducing-text-and-code-embeddings/ [Accessed 22 Aug. 2022].

SearchEnterpriseAI. (n.d.). *What is Unsupervised Learning?* [online] Available at: https://www.techtarget.com/searchenterpriseai/definition/unsupervised-learning.

Leapfrog Technology. (2021). *5 Companies Making the Most of Recommendation Systems*. [online] Available at: https://www.lftechnology.com/blog/recommendation-systems/.

Han, J., Kamber, M. and Pei, J. (2012). Getting to Know Your Data. *Data Mining*, [online] pp.39–82. doi:10.1016/b978-0-12-381479-1.00002-2.

Loukas, S. (2020). *K-Means Clustering: How It Works & Finding The Optimum Number Of Clusters In The Data*. [online] Medium. Available at: https://towardsdatascience.com/k-means-clustering-how-it-works-finding-the-optimum-number-of-clusters-in-the-data-13d18739255c.

Al-Masri, A. (2019). *How Does k-Means Clustering in Machine Learning Work?* [online] Medium. Available at: https://towardsdatascience.com/how-does-k-means-clustering-in-machine-learning-work-fdaaaf5acfa0.

Prabhu (2019). *Understanding NLP Word Embeddings — Text Vectorization*. [online] Medium. Available at: https://towardsdatascience.com/understanding-nlp-word-embeddings-text-vectorization-1a23744f7223.

Johnson, D. (n.d.). *Supervised vs Unsupervised Learning: Key Differences*. [online] www.guru99.com. Available at: https://www.guru99.com/supervised-vs-unsupervised-learning.html#:~:text=well%20%E2%80%9Clabeled.%E2%80%9D-.
IBM Cloud Education (2020). *What is Unsupervised Learning?* [online] www.ibm.com. Available at: https://www.ibm.com/cloud/learn/unsupervised-learning.

Turner, A. (2021). *How Many Smartphones Are In The World?* [online] BankMyCell. Available at: https://www.bankmycell.com/blog/how-many-phones-are-in-the-world.

Roberts, R. (2022). *2022 Beauty Industry Trends & Cosmetics Marketing: Statistics and Strategies for your Ecommerce Growth*. [online] Common Thread Collective. Available at: https://commonthreadco.com/blogs/coachs-corner/beauty-industry-cosmetics-marketing-ecommerce#statistics.

Anon, (2020). *7 Easy Product Recommendation Strategies You Need to Use Today - TrustPulse*. [online] Available at: https://trustpulse.com/product-recommendation-strategies/.

Evergreen Beauty College. (2012). *How Makeup Has Changed Over Time*. [online] Available at: https://www.evergreenbeauty.edu/blog/bloghow-makeup-has-changed-over-time/.

Statista Research Department (2022). *U.S. expenditure cosmetics, perfume, bath preparation products 2018*. [online] Statista. Available at: https://www.statista.com/statistics/304996/us-expenditure-on-cosmetics-perfume-and-bath-preparation/.

Statista Research Department (2022a). *Topic: Cosmetics market in the United Kingdom (UK)*. [online] Statista. Available at: https://www.statista.com/topics/5760/cosmetics-market-in-the-united-kingdom-uk/#dossierContents__outerWrapper.

EnterpriseAsia. (2019). *Beauty Redefined: How a Company Leveraged on Technology to Revolutionise the Beauty Industry | Enterprise Asia*. [online] Available at: https://www.enterpriseasia.org/portfolio-item/beauty-redefined-how-a-company-leveraged-on-technology-to-revolutionise-the-beauty-industry/.

Rajula, H.S.R., Verlato, G., Manchia, M., Antonucci, N. and Fanos, V. (2020). Comparison of Conventional Statistical Methods with Machine Learning in Medicine: Diagnosis, Drug Development, and Treatment. *Medicina*, 56(9), p.455. doi:10.3390/medicina56090455.

Nyman, H. (n.d.). *Machine learning, and how it differs from traditional analytics*. [online] www.linkedin.com. Available at: https://www.linkedin.com/pulse/machine-learning-how-differs-from-traditional-analytics-henrik-nyman [Accessed 20 Aug. 2022].

Analytics Vidhya (2019). *A Comprehensive Guide to Understand and Implement Text Classification in Python*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/.

Khushalani, A.S. (2019). *Topic Modeling with Amazon Reviews*. [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/topic-modeling-with-amazon-reviews-8dcb40ffc97d [Accessed 20 Aug. 2022].

IBM Cloud Education (2020). *What is Random Forest?* [online] www.ibm.com. Available at: https://www.ibm.com/cloud/learn/random-forest.

www.linkedin.com. (n.d.). *Chapter 9.1 : NLP - Word vectors.* [online] Available at: https://www.linkedin.com/pulse/chapter-91-nlp-word-vectors-madhu-sanjeevi-mady-/ [Accessed 20 Aug. 2022].

Nasher, K. (2021). Sentiment Analysis using GloVe. [online] Medium. Available at: https://khuloodnasher.medium.com/sentiment-analysis-using-glove-92d72e6489e8 [Accessed 18 Aug. 2022].

Robinson, J.S. and D. (n.d.). *6 Topic modeling | Text Mining with R.* [online] *www.tidytextmining.com*. Available at: https://www.tidytextmining.com/topicmodeling.html#:~:text=Every%20document%20is%20a%20mixture [Accessed 20 Aug. 2022].

upGrad blog. (2021). *Multinomial Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2022*. [online] Available at: https://www.upgrad.com/blog/multinomial-naive-bayes-explained/#:~:text=The%20Multinomial%20Naive%20Bayes%20algorithm%20is%20a%20Bayesian%20learning%20approach.

Islam, Md. Zahidul & Liu, Jixue & Li, Jiuyong & Liu, Lin & Kang, Wei. (2019). A Semantics Aware Random Forest for Text Classification. 1061-1070. 10.1145/3357384.3357891.

Garbade, M. (2018). *Understanding K-means Clustering in Machine Learning*. [online] Towards Data Science. Available at: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1.

Wikipedia Contributors (2019b). *Random forest*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Random_forest.

Wikipedia Contributors (2019a). *Decision tree learning*. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Decision_tree_learning.

Mbaabu, O. (2020). *Introduction to Random Forest in Machine Learning*. [online] Engineering Education (EngEd) Program | Section. Available at: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

Ratz, A. (2021). *Classification and Natural Language Processing (NLP)*. https://towardsdatascience.com/multinomial-naïve-bayes-for-documents-classification-and-natural-language-processing-nlp-e08cc848ce6.

Sciencedirect.com. (2019). *Cosine Similarity - an overview | ScienceDirect Topics*. [online] Available at: https://www.sciencedirect.com/topics/computer-science/cosine-similarity.

gocardless.com. (n.d.). *Why Is Customer Satisfaction So Important?* [online] Available at: https://gocardless.com/guides/posts/customer-satisfaction/.

Salesforce.com. (n.d.). *How to Improve Your Customer Retention Rate*. [online] Available at: https://www.salesforce.com/resources/articles/customer-retention-rate/.

Miquido (2020). *Recommendation Systems: Benefits, Types & Examples - Miquido Blog*. [online] Miquido. Available at: https://www.miquido.com/blog/perks-of-recommendation-systems-in-business/ [Accessed 21 Aug. 2022].

www.yuty.me. (n.d.). *Putting You in Beauty - Yuty*. [online] Available at: https://www.yuty.me/.

www.perfectcorp.com. (n.d.). *Beauty AR Company and Makeup AR Technology Platform*. [online] Available at: https://www.perfectcorp.com/business/blog/general/how-ai-and-ar-innovation-are-changing-the-beauty-tech-industry.

https://theindustry.beauty/. (n.d.). *How AI and AR are revolutionising the beauty industry | The Industry Beauty*. [online] Available at: https://theindustry.beauty/the-role-of-ai-and-ar-in-the-beauty-industry/.

GreyB, T. (2020). *AI in Beauty Industry: Companies & Startups Research*. [online] GreyB. Available at: https://www.greyb.com/ai-in-beauty-industry/.

www.selfridges.com. (n.d.). *Beauty in Question: how is AR transforming beauty? | |*. [online] Available at: https://www.selfridges.com/GB/en/features/articles/beauty/beauty-in-question/ar-technology-masha-batsii-geraldine-wharry/

theappsolutions.com. (n.d.). *Top 5 Applications of AI in the Beauty Industry to Restrain COVID Disruptions*. [online] Available at: https://theappsolutions.com/blog/how-to/how-to-use-ai-in-the-beauty-industry/.

www.cosmeticsbusiness.com. (n.d.). *Coty develops personal beauty assistant for Amazon Echo Show*. [online] Available at: https://www.cosmeticsbusiness.com/news/article_page/Coty_develops_personal_beauty_assistant_for_Amazon_Echo_Show/138585

Code References:

QMplus: Masterclass in Business Analytics module's lectures, lab practicals and videos.

https://towardsdatascience.com/

https://www.youtube.com/

https://www.analyticsvidhya.com/

https://www.simplilearn.com/