# **Fashion Product Image Classification**

# Team 19

# Gaurav Gunjal NUID: 002820367

# Niraj Malpani NUID: 002849800

## **Introduction**

### **Background:**

### Within the dynamic realm of e-commerce, the fashion industry stands as a beacon of innovation and creativity, constantly shaping and reshaping consumer trends and preferences. In this digital age, online platforms have revolutionized how fashion products are marketed, sold, and consumed, offering consumers unparalleled access to a vast array of styles and trends. The abundance of fashion product data, comprising high-resolution images, detailed descriptions, and categorical labels, provides a unique opportunity to delve deep into the intricacies of consumer behavior, market dynamics, and industry trends within the e-commerce fashion ecosystem.

### **Motivation:**

# Our motivation stems from the technical challenge of leveraging image-based machine learning models to enhance user experience in e-commerce fashion. By enabling users to classify products based on images, we aim to provide a more intuitive and efficient way for users to explore fashion items. Our focus is on developing robust algorithms that can accurately analyze visual features and classify products in real-time, ultimately improving user engagement and satisfaction. Through this technical innovation, we strive to push the boundaries of online fashion retailing and set new standards for user interaction and convenience in the industry.

### **Goal:**

### Our primary objective is to leverage fashion product data to derive actionable insights, inform strategic decisions, and foster growth within the e-commerce fashion landscape. We aim to develop robust image classification models for accurate categorization and enhance customer experiences by recommending similar products based on the initial classification. Through this model e-commerce websites can show relevant products from the database based on the classification made by the model.

**Methodology**

### **Data Preprocessing:**

To prepare the data for analysis, the initial steps include data preparation and cleaning. Prepare data by loading the images and corresponding metadata from the datasheet. Data cleaning will involve identifying and addressing missing values, anomalies, and inconsistencies in the dataset to ensure its suitability for analysis. Additionally, features will be refined to differentiate between various fashion product categories and optimize classification accuracy.

### **Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) will be conducted to uncover insights into the characteristics and relationships within the fashion product dataset. Utilizing Univariate and Bivariate analysis techniques, including correlation analysis, we will unveil associations between product attributes, customer preferences, and market trends. Data visualization methods such as bar graphs, distribution plots, and box plots will be employed to identify patterns and anomalies, facilitating a deeper understanding of the dataset.

### **Feature Extraction:**

Drawing upon image processing techniques and domain expertise, we will engineer new features from the images to enhance classification accuracy and predictive capabilities. Image features such as color histograms, texture descriptors, and shape-based features will be extracted to capture unique characteristics of fashion products.

### **Model Building and Evaluation:**

The dataset will be partitioned into training and testing sets to facilitate model development and evaluation. We will employ a variety of machine learning algorithms, including Support Vector Machine (SVM), Random Forest, Logistic Regression, and Decision Tree classifiers, to classify fashion products based on their image features. Each classifier will be trained using the training dataset and fine-tuned using cross-validation techniques to optimize model parameters and improve classification accuracy. Model performance will be rigorously evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1 score to identify the most effective solution for fashion product classification.

### **Prediction:**

Once trained, the classifiers will make predictions on the testing dataset.

# **Description of the Dataset**

The dataset is sourced by running a script on the e-commerce website name [www.myntra.com](http://www.myntra.com). This script extracts the image and other data of various products sequentially. The dataset size is 572 MB. It consists of 44441 rows and 10 columns. Each row has a product ID which is mapped to the product image. The description of columns from the dataset are:

1. **id:** Unique identifier for each entry in the dataset.
2. **gender:** Gender category of the clothing item. Can be "Men", "Women", or possibly other categories.
3. **masterCategory:** High-level category of the clothing item. Examples include "Apparel", "Accessories", "Footwear", etc.
4. **subCategory:** Sub-category of the clothing item, providing more specific information about the type of clothing. Examples include "Topwear", "Bottomwear", "Watches", "Shoes", etc.
5. **articleType:** Specific type of clothing item. Examples include "Shirts", "Jeans", "T-shirts", "Casual Shoes", "Perfume and Body Mist", etc.
6. **baseColour:** The primary color of the clothing item.
7. **season:** Season during which the clothing item is typically worn, such as "Fall", "Summer", "Winter", "Spring", etc.
8. **year:** Year of the clothing item's release or production.
9. **usage:** Intended usage scenario for the clothing item, such as "Casual", "Party", "Formal", etc.
10. **productDisplayName:** Name or description of the clothing item.

# 

# **Data Source**

<https://www.kaggle.com/datasets/bhavikjikadara/e-commerce-products-images>

**Results and Analysis**

**Data Preprocessing**

The dataset for our project was sourced from a CSV file, encompassing 10 columns and approximately 44,500 rows. This extensive data set provides a rich foundation for our analysis and insights.

A screenshot of a computer

Description automatically generated

For our project, we load random images from a directory. The directory contains a diverse set of images, spanning various categories.

A collage of a person's clothes

Description automatically generated

This dataset boasts a treasure trove of random images showcasing a myriad of online shopping products.

Delving into our dataset, we diligently check for any missing values across all columns, ensuring the integrity and completeness of our data.

A screenshot of a computer

Description automatically generated

Venturing into the depths of our dataset, we meticulously uncover the unique values within the 'usage' column, despite its notable prevalence of missing data. This exploration offers valuable insights into usage patterns, enriching our understanding of the dataset  
  
A screenshot of a computer

Description automatically generated

Examining 'productDisplayName' patterns for null values in 'usage', we seek correlations despite the missing data.

A screenshot of a computer screen

Description automatically generated

Identified a pattern: all null values in the 'usage' column correspond to cosmetic products. Thus, replaced null values with 'Cosmetic'.

**Exploratory Data Analysis (EDA)**

Performed EDA to unveil the dataset's structure, distributions, and relationships, enriching our understanding for further analysis.

The 'productDisplayName' combines brand name, gender, base color, and article type, encapsulating various product attributes within a single field.

Explored the gender column to understand its relationship with 'productDisplayName', unveiling nuanced connections between gender and product attributes.

Displaying unique gender

A screen shot of a computer

Description automatically generated

Calculated the count of rows where 'gender' values are absent from the 'productDisplayName' column, revealing instances where gender information is not explicitly represented.

A screenshot of a computer

Description automatically generated

Discovered 4096 rows with missing gender information within the 'productDisplayName' column, suggesting the need for further data refinement or imputation techniques.

Exploring baseColour column and its relationship with productDisplayName

Displaying unique baseColour

A close up of words

Description automatically generated

Determine the count of rows where the "baseColour" values are absent from the "productDisplayName” column.

A screenshot of a computer screen

Description automatically generated

Found 3401 rows lacking base color information within the 'productDisplayName' column, suggesting potential inconsistencies or incomplete data. Further analysis may be needed to address this issue.

Explored the 'articleType' column to unveil its relationship with other dataset attributes, facilitating a deeper understanding of product categorization and its implications.

Displaying unique articleType

A screen shot of a computer

Description automatically generated

Identified the count of rows where 'articleType' values are absent from the 'productDisplayName' column, indicating instances where product categorization information is not explicitly provided.

A screenshot of a computer

Description automatically generated

Discovered 27,628 rows where the 'productDisplayName' does not correspond with the 'articleType' of the same row, suggesting discrepancies in product categorization or labeling.

**Visualization of the above information**

A graph showing rows of rows

Description automatically generated

Due to the small and unclear nature of the images, extracting brand names is rendered impossible, posing a challenge to comprehensive data analysis.

Due to the small and unclear nature of the images, along with the absence of visible brand names, selecting 'productDisplayName' as the target vector is not feasible. This decision is further supported by the presence of diverse values from 'gender', 'articleType', and 'baseColour', highlighting the challenges in accurately predicting or classifying products based on their attributes.

Utilizing word cloud visualization, we delve into the unique values of the 'subCategory' column to gain insights and uncover patterns within the dataset.  
  
A close up of words

Description automatically generated

Conducting an exploration of the unique values within the 'masterCategory' column

A close up of words

Description automatically generated

Delving into the unique values within the 'articleType' column to gain insights into the diverse types of products present in the dataset.

A word cloud of clothing

Description automatically generated

Visualizing and comparing the unique values of 'articleType', 'subCategory', and 'masterCategory'

A graph of different values

Description automatically generated

By comparing the unique values of 'articleType', 'subCategory', and 'masterCategory' columns, we highlight that 'articleType' exhibits the highest diversity. Thus, we have selected 'articleType' as our primary target vector for its rich variation and representation of product types."

**Feature Extraction**

Manually extracting features involves deriving histograms of gradients and performing edge detection from randomly selected images.

A screenshot of a computer screen

Description automatically generated

Due to the time and computational resources required for manual feature extraction, we have opted to extract features from the VGG16 deep learning model.

We've restructured the VGG16 model solely for feature extraction purposes, abstaining from its use in training or prediction tasks.

Our modified VGG16 model is configured specifically for feature extraction, capturing 4096 features from each image in our dataset.

A screenshot of a computer program

Description automatically generated

In summary of the model we're utilizing, we've removed the prediction layer, repurposing the VGG16 architecture solely for feature extraction.

A screenshot of a computer

Description automatically generated

We've created a new column 'image\_features' where we've mapped the extracted features to their respective image IDs and saved the data to a new CSV file.

A screenshot of a computer

Description automatically generated

Visualizing the count of each unique article type to assess class distribution within the dataset. This will help me determine if there's any class imbalance present.

A graph of a number of people

Description automatically generated with medium confidence

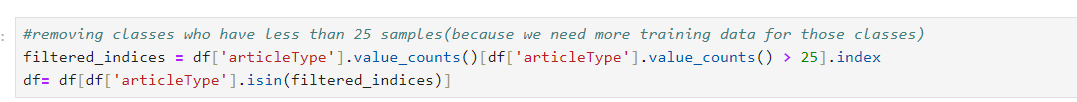
As evident from the above graph, there is a notable class imbalance, particularly with a high count of T-shirts in the 'articleType' category.

Additionally, to address this imbalance, we have removed 2000 rows corresponding to T-shirts to achieve better class balance.

A graph of a number of text

Description automatically generated with medium confidence

Furthermore, from the graph, it's evident that some classes have a very low unique count, which could potentially impact our model's accuracy, recall, and precision. Thus, we have removed 'articleType' counts with values less than 25 to mitigate this issue.



**Label Encoding**

Implemented label encoding to transform categorical data into numerical format, facilitating compatibility with machine learning algorithms.

A screenshot of a computer code

Description automatically generated

**Principal Component Analysis**

Utilized Principal Component Analysis (PCA) to reduce the dimensionality of the data to 100 components, enabling more efficient representation while preserving key information.

A screenshot of a computer program

Description automatically generated

Calculating the cumulative variance explained by principal components to assess what percentage of features can be retained through PCA.

A screenshot of a computer

Description automatically generated

Visualizing the cumulative variance explained by principal components to gain insight into the percentage of features retained through PCA.

A graph with a curve

Description automatically generated

PCA retains 80% of the information from the original 4096 features, demonstrating efficient dimensionality reduction while preserving a significant portion of the dataset's variance.

**Model Selection and Evaluation**

**Logistic Regression Model**

Trained and evaluated a Logistic Regression model with a maximum iteration of 300 to achieve optimal accuracy.

A screen shot of a computer program

Description automatically generated

Attained an accuracy of 0.80 with respectable precision and recall scores, indicating satisfactory performance of the Logistic Regression model.

A screenshot of a number

Description automatically generated

**K-nearest neighbor**

Performed training and evaluation of the K-nearest neighbor model to explore its predictive capabilities.

A screenshot of a computer program

Description automatically generated

Achieved an accuracy of 0.79 with the K-nearest neighbor model, indicating its performance in classification tasks.

A screenshot of a number

Description automatically generated

**Support vector Classifier**

Conducted training and evaluation of the Support Vector Classifier (SVC), identified as the best-performing model for our project due to its superior accuracy.

A screenshot of a computer program

Description automatically generated

Attained an impressive accuracy of 0.83 with the Support Vector Classifier model, showcasing excellent precision and recall metrics, making it the optimal choice for our project.

A table of numbers with numbers on it

Description automatically generated

Generated and visualized the confusion matrix for the Support Vector Classifier model to assess its performance in classification tasks.

A screenshot of a graph

Description automatically generated

**Conclusion**

* In conclusion, our project embarked on the ambitious journey of enhancing e-commerce fashion experiences through advanced image classification models. We judiciously prepared and preprocessed an extensive dataset, derived from a leading online fashion portal, which laid the groundwork for our exploratory analysis and feature extraction. By implementing PCA, we effectively condensed the feature space from a high-dimensional dataset while preserving the essence of the visual information, thereby boosting our model's performance and efficiency.
* The employment of machine learning algorithms, including Logistic Regression, KNN, and SVC, culminated in identifying SVC as the most proficient with an impressive accuracy rate of 82%. This superior performance signifies not only the robustness of SVC in handling image data but also its aptness for real-time classification tasks in a dynamic e-commerce environment.
* Our findings illuminate the potential of such technologies to revolutionize the way consumers interact with fashion products online, offering them a seamless and intuitive shopping experience. By recommending products that are visually similar to users' preferences, our model does not only streamline the shopping experience but also aids in navigation through the vast selection of fashion merchandise.
* As we move forward, we envision integrating our model into a user-friendly application, setting a benchmark for personalized shopping and paving the way for future innovations in the realm of fashion retail. This project serves as a testament to the transformative power of machine learning in e-commerce and opens avenues for continued enhancements in precision, speed, and user engagement.

**References**

<https://towardsdatascience.com/10-papers-you-should-read-to-understand-image-classification-in-the-deep-learning-era-4b9d792f45a7>

<https://www.sciencedirect.com/science/article/pii/S1110866517303225>