DATA 270: Data Analytics Processes

# E-Commerce Recommendation System: Using Conventional Machine Learning Approach Based on Visual Similarity

## GROUP 8

November 28, 2022

Under Guidance of Dr. Eduardo Chan


Team

**Deepak Halliyavar**
**Lohitha Vanteru**
**Mahe Jabeen Abdul**
**Pranavi Sandrugu**

Department of Applied Data Science, San Jose State University

# Agenda

- Background
- Motivation
- Technology and Literature Survey
- Problem Statement
- Scope
- Project Requirements
- Deliverables
- Project Resource Requirements & Plan
- Data Preparation
- Machine Learning Models
- Conclusion & Future scope

“Recommendation systems have always been in demand in every industry and domain as matching a user’s preference is always the most important thing in modern-day business”

# Background

- What is a Recommendation System?
- Type of Recommendation System.
- What is Visual Similarity?

**Recommender Systems**

**Popularity-based RS** | **Content-based RS** | **Collaborative Filtering**

https://pub.towardsai.net/product-recommendation-based-on-visual-similarity-on-the-web-machine-learning-project-end-to-end-6d38d68d414f

# Motivation

- ## Age of Pandemics and Digital gadgets:

  ***User Preference –*** *Online Shopping/E-commerce application.*

- ## Existing Recommendation Systems :

  *Rely on* ***keyword matching*** *techniques or* ***Past purchasing history***

- ## Drawbacks

  *-Not effective due to varying* ***semantic usage*** *across multiple websites.*

  *-Quickly generate* ***unnecessary*** *suggestions.*

- ## Motive:

  *To Provide consumers a similar enriched experience as in person retail shopping.*

https://www.lynda.com/Data-Science-tutorials/Content-based-recommendations-Recommending-based-product-attributes/563030/600814-4.html

# Technology and Literature Survey

| References | Dataset | Summary | Results |
|---|---|---|---|
| Pandey et al. (2016) [1] | <u>**Several Datasets**</u>:<br>ZuBuD -1005 images<br>WANG – 1000 images<br>Caltech101 – 8677 images<br>Web images | • Uses **agglomerative hierarchical clustering algorithm**.<br>• The proposed clustering compares only the representative images of clusters at any time.<br>• Novel idea of tracking the loss of information to get clusters automatically using proportional reduction in error method. | Best precision for all datasets.<br>ZuBuD- 0.7<br>WANG – 0.57<br>Caltech101 – 0.35<br>Web Images – 0.80 |
| Putri et al. (2020) [2] | Movielens – available online (Ratings, tags, films, users, genres etc.) | • Proposes a **movie recommendation** system using various unsupervised ML algorithms such as s **K-Means, K-Means Mini Batch, Birch Algorithms** & compares various performance measures. | Best computation time 13.75ms- Minibatch Kmean is good<br>(Calinski-Harabasz) -59.42<br>Birch of 1.24<br>(Davies-Bouldin) |

| References | Dataset | Summary | Results |
|---|---|---|---|
| Pitolli et al. (2020) [3] | • Publicly available Windows malware.<br>• Samples between 2006 and 2016 – 4,100 samples. | • MalFamAware uses **BIRCH algorithm.**<br>• Comparison how MalFamAware performs in terms of family identification (against other clustering algorithms) and malware classification | • **Good accuracy** in family identification and high accuracy in malware classification<br>• Very **low execution time** |
| Ullah et al. (2019) [4] | • Amazon product image with 3.5million images of 20 classes;<br>• Randomly selected 100 images from each class (**2000** images) | • Deep learning<br>• Deep learning with RF | • RF= **75%** accuracy;<br>• RF with DL = **84%;**<br>• **98% correct** recommendations |
| Asiroglu et al. (2019) [5] | • Stanford University's Clothing Attributes<br>• Dataset(**1856** images) - Feature training; e -survey with use google forms | • Proposes a smart clothing recommendation system.<br>• The model used is inception based **Deep learning.** | Accuracies<br>**86%** : Gender, clothes color prediction<br>**98%** : Clothes pattern prediction |

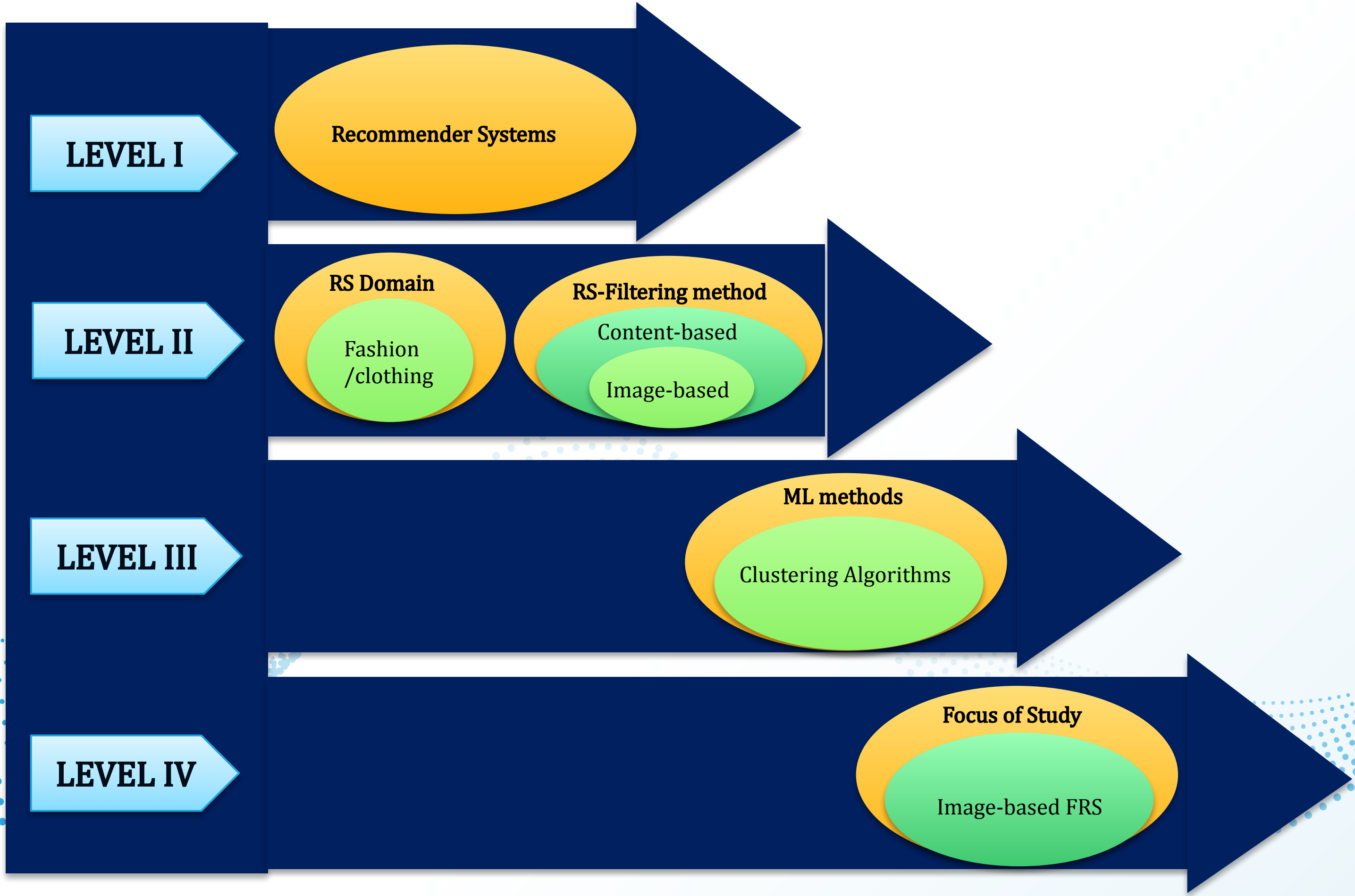
# Problem Statement

- Detecting all products in an image and retrieve similar images from the database.

- Our goal is to increase the satisfaction of customers by displaying similar products and providing them with potentially superior outcomes by pulling out more relevant items from the provided images based on their features.


https://www.robertoreif.com/blog/2018/05/14/product-recommendations-using-image-similarity-yy76x

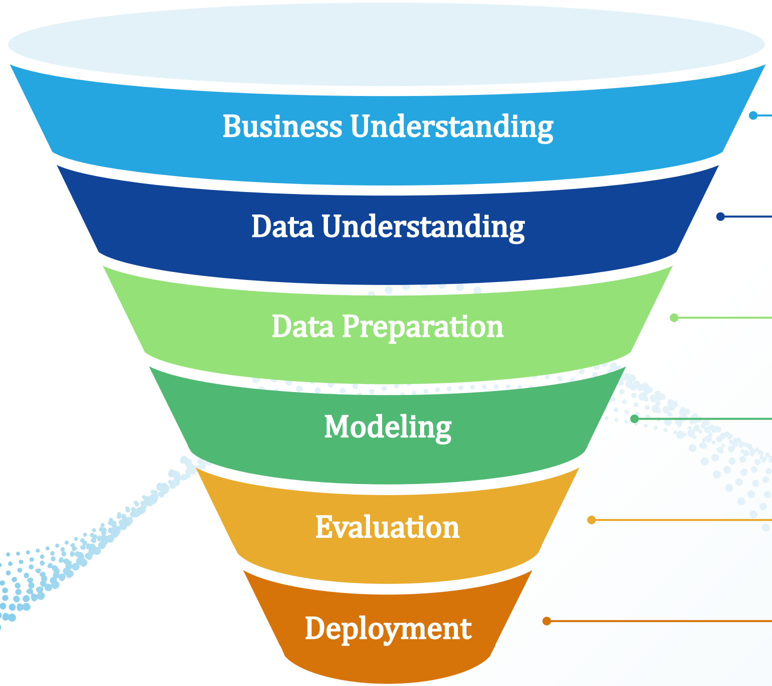
# Scope and focus of this project



# Project Requirements

| Functional Requirements | Data Requirements | AI Requirements |
|---|---|---|
| • Cloud storage to store images<br>• Virtual Machines with enough computation power to process image data<br>• Project Management Tool | • Collect High Definition Clothing Images with diverse categories.<br>• Important Parameters: Image id, Format, Resolution<br>• Target Labels : Shirt, Skirt, Top, Dress e.t.c., | • Dimensionality Reduction Techniques<br>• Clustering Machine Learning Algorithms<br>• Similarity Measures |

# CRISP DM Approach




Identifying Problem and proposing a plan

Exploratory Data Analysis

Preparing  data ready for ML models

Selection & application of ML models

Performance metrics & optimal model selection

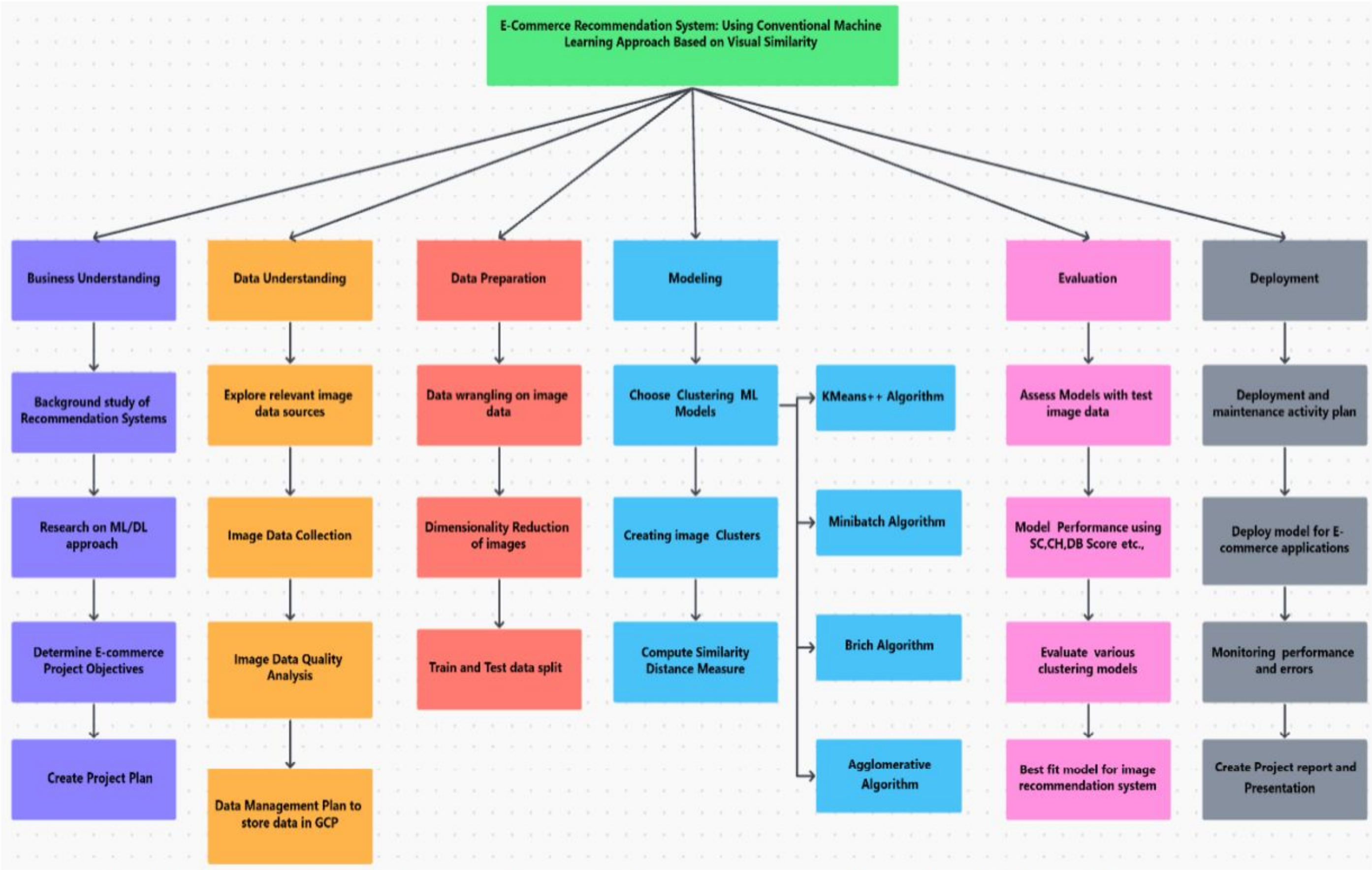
Make the project accessible for customers

# Deliverables

2022

- **I** — SEP 19 — Project Proposal
- **II** — OCT 11 — Project Organization Plans (WBS, Gantt & PERT Charts)
- **III** — OCT 25 — Data Collection & Data Exploration Plans
- **IV** — NOV 4 — Data Engineering Report
- **V** — DEC 5 — Model Development (Individual report)
- **VI** — DEC 11 — Model comparison & Performance Evaluation (Final group report)

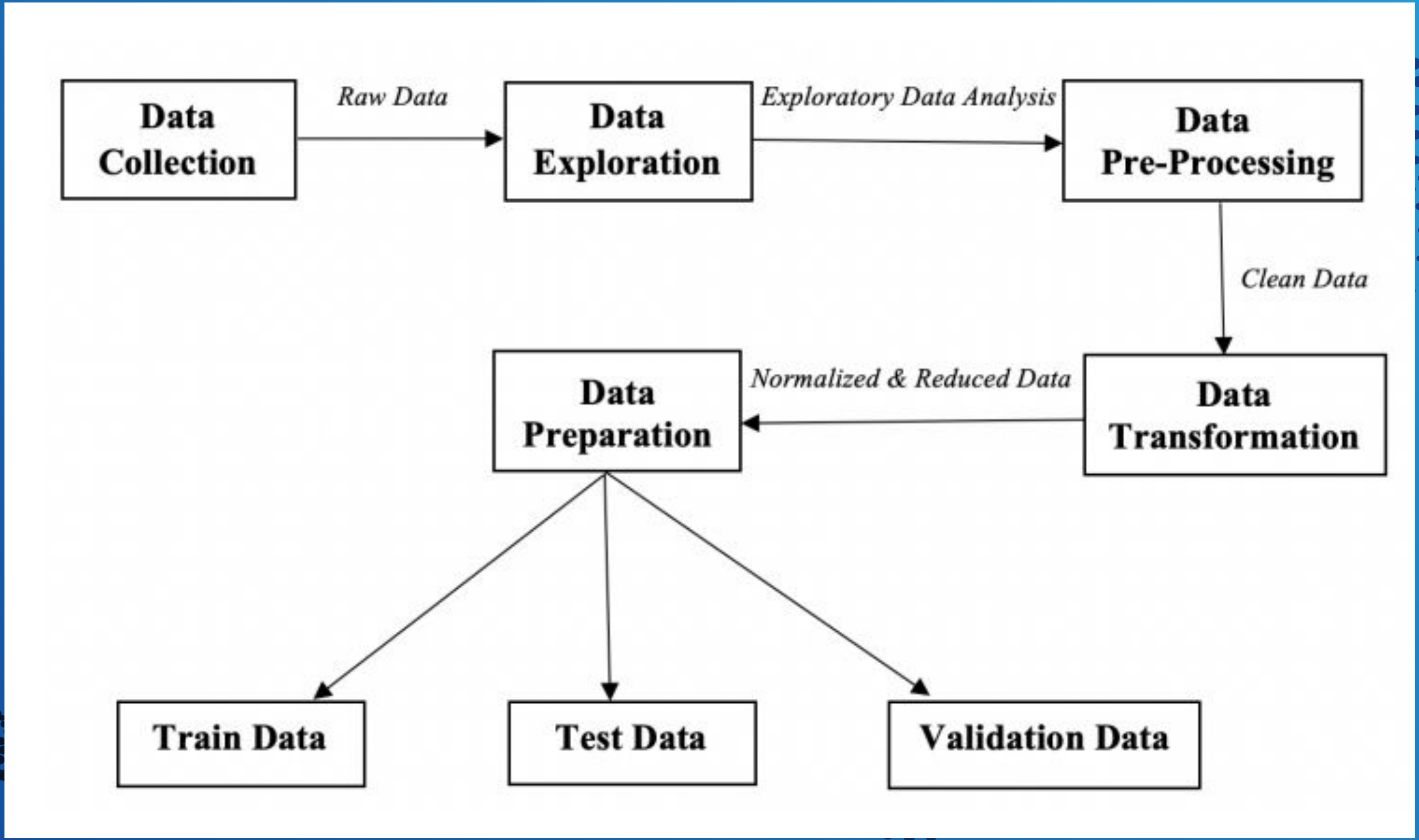
# Work Breakdown Structure (WBS)



# Project Resource Requirements and Plan

| Resource | Type | Purpose | Time Duration | Cost in USD |
|---|---|---|---|---|
| Google Cloud Storage Bucket | Hardware | Data Storage | ~4 months | $11.20 |
| GC VM n1-standard-16 (vCPUs: 16, RAM: 60GB), NVIDIA TESLA T4 GPU:2 | Hardware | Virtual Machine | ~2.5 months | $733.75 |
| Google Colab Python 3.7 Version | Software | Data Preprocessing | ~2.5 months | Free |
| ClickUp version 2.19 | Tool | Project Management | ~3 months | $40 |
| GitHub version 3.6.2 | Tool | Data Redistribution | ~2.5 months | Free |
| Zoom version 5.12 | Tool | Project Work Collaboration | ~3 months | Free |
| MS Office 365 Suite version 2209 | Tool | Project Documentation | ~3 months | Free (Student License) |
| | | | **Total Cost Estimation of the Project** | $784.95 |

# Data Preparation



# Dataset collection process

Created a call for action to be shared on social media

Defined set of rules

Created Air Table Forms to upload images

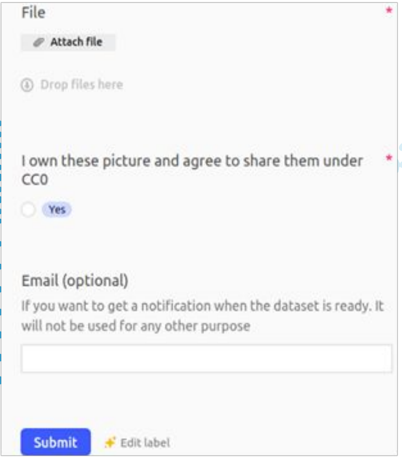
Toloka Crowdsourcing Platform

Tagias- Data Collection Company

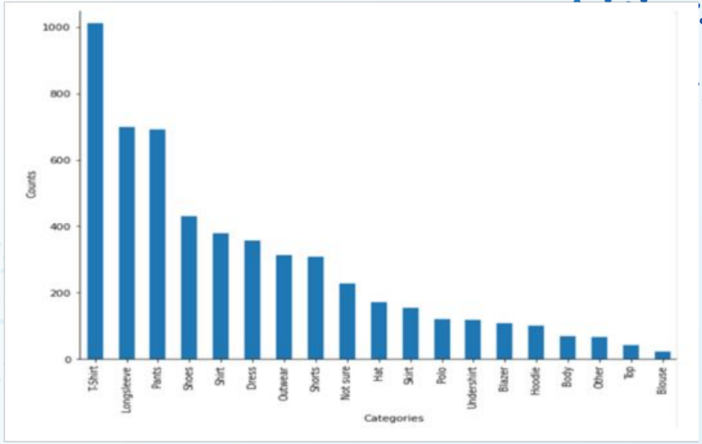
7.07 GB

5000 + images

20 Categories







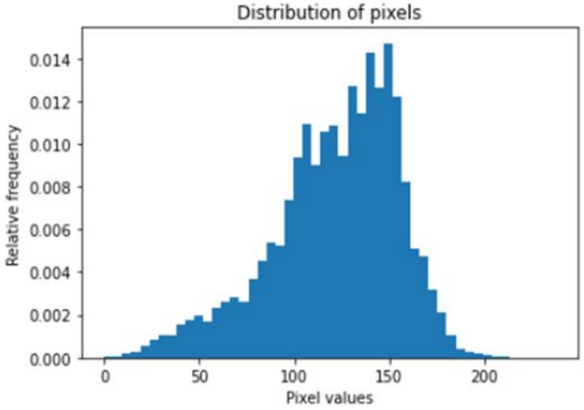

Source: https://medium.com/data-science-insider/clothing-dataset-call-for-action-3cad023246c1

# Raw Data Set samples


| | image | sender_id | label | kids |
|---|---|---|---|---|
| 4376 | 90db4c3d-2d36-4086-9a2d-c1548b922787 | 50 | T-Shirt | False |
| 5230 | 5176fdc6-3c9b-4291-854e-487861248b4b | 204 | Shorts | False |
| 2694 | 3c9c2f96-3e55-4c75-9a21-bbbc6c7d96e7 | 181 | Longsleeve | False |
| 3764 | 0637d9a1-60fa-4b26-86db-f2511a17db62 | 50 | T-Shirt | False |
| 4333 | 98424108-fcd3-46a4-8a27-fcf4db664f4c | 181 | Longsleeve | False |

```
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   image      5384 non-null   object
 1   sender_id  5403 non-null   int64
 2   label      5403 non-null   object
 3   kids       5403 non-null   bool
dtypes: bool(1), int64(1), object(2)
memory usage: 132.0+ KB
```

| Image_Name | Resolution |
|---|---|
| d8445328-a318-434b-9c4e-fa29cf04be41.jpg | 1500x2000 |
| ff8ca4c9-b3dc-4d84-8dba-095581f94dbe.jpg | 3024x4032 |
| 39d92542-ffcb-4b3b-896f-d158eccf8172.jpg | 3024x4032 |
| e2e8ba7f-f36b-4ccf-b116-b08265f2ee84.jpg | 4032x3016 |
| a307671b-f4cd-4f75-bf32-5d4a72320b93.jpg | 1456x2592 |
| 911a104e-2ca1-4983-a732-7cff59fcbec1.jpg | 3024x4032 |
| f341e492-7c08-49dd-ba45-b1095d0239f1.jpg | 3024x4032 |
| e2dfcb33-19ba-4c82-998b-66f125086e63.jpg | 3024x4032 |
| f3ca5280-6e74-44c6-ad14-f04ca1aee182.jpg | 3024x4032 |
| 769b4c86-0e6b-4576-8993-0fa927ace181.jpg | 3000x4000 |



# Data Pre-Processing

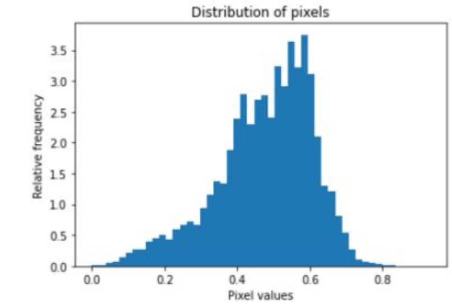
| Issues Found | Suggested Fixes |
|---|---|
| Invalid Images | Duplicate images , other random object images, images directly downloaded from internet are discarded. |
| Raw and unlabeled data | Manually annotated the labels using domain knowledge and Python widget tool. |
| Labeling mistakes | Applied high learning rate neural network on data and corrected the labels that differs from model predictions |
| Irrelevant Columns | Dropping all the columns except “image” and “label”. |
| Non uniform Image properties | Resize the images such that all are of same size, format, and resolution. |
| Noisy and Inconsistent data | Denoised and enhanced all the images to be consistent. |

# Pre-Processed Data Samples

```
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   image   5384 non-null   object
 1   label   5384 non-null   object
dtypes: object(2)
```

| Image_Name | Resolution |
|---|---|
| 85dd32b5-b370-4547-9677-4e3c1e2ac477.jpg | 600x600 |
| 74d4f5f3-d0fc-46df-80bd-c00f7144feff.jpg | 600x600 |
| 0b0457d9-32f9-4b34-915d-017e5525d9f6.jpg | 600x600 |
| 2b41ceef-e71c-4229-86fc-e7b50a6ef8c5.jpg | 600x600 |
| fe94de4f-958b-4a4f-a8d9-37748a70717d.jpg | 600x600 |
| ddc3da65-2ba7-4012-adf4-0b6e636fad6b.jpg | 600x600 |
| 74222128-e39b-4787-afb2-f88a92b8e537.jpg | 600x600 |
| 7fa97b45-380a-4c9c-8de4-522976cb4972.jpg | 600x600 |
| fec3f552-e894-4737-8867-a9447f68f6be.jpg | 600x600 |
| d9e84490-185d-48f9-ac16-4ef3360616d5.jpg | 600x600 |

*Distribution of pixels in the transformed image*




Noisy Image

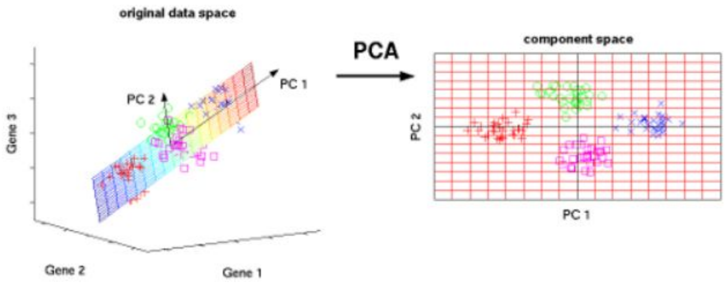
Denoised Image

Original Image

Enhanced Image

# Transformed Data Samples

## Data Reduction :Principal Component Analysis





**Reduced to 50 Dimensions.**

## Data Augmentation : Image Rotation and Flipping


*Note.* Comparison between image before rotation and after rotation.


Source: http://www.nlpca.org/pca_principal_component_analysis.html

# Data Preparation

- Considering only Top 10 labels of data to avoid class imbalance.

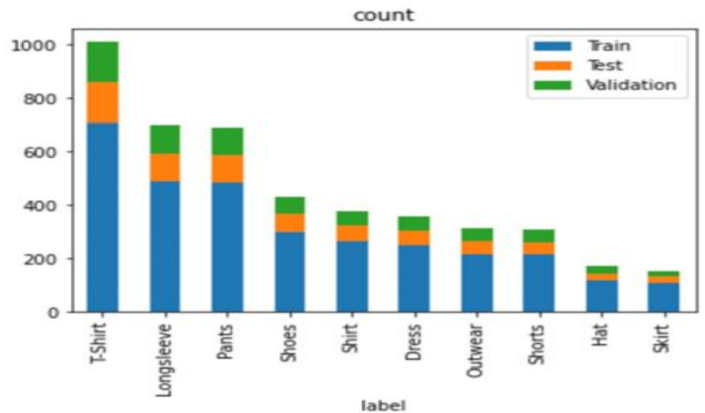
| Data in each stage | Count |
|---|---|
| Raw data | 5403 |
| Pre Processed Data | 5384 |
| Transformed Data | 17,286 |
| Data for Modeling | 11,343 |
| Training Data | 7940 |
| Test Data | 1702 |
| Validation Data | 1701 |

**Train** 70 %

**Test** 15 %

**Validation** 15 %

*Label Data Distribution*

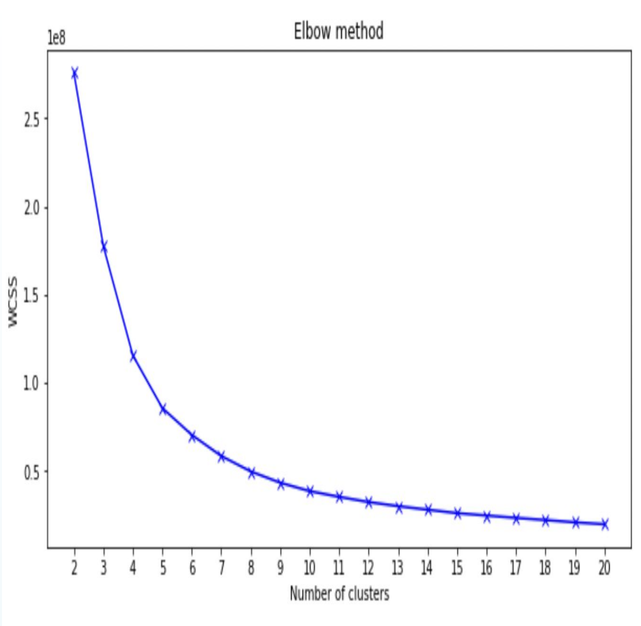



*Note.* Distribution of labels with each being split into Train, Test, and Validation sets

# Machine Learning Models

We have used unsupervised clustering algorithms such as

| Clustering models | Use cases | Limitations |
|---|---|---|
| **K-Means++** | K-means++ is a smart centroid initialization method for the K-means algorithm. It is both faster and provides better performance. | Not efficient to perform large datasets with limited resources like memory and slower CPU. |
| **MiniBatch** | It reduces the temporal and spatial cost of the algorithm. | Results are not better than standard algorithms. |
| **Birch** | For large datasets, Outlier removals | It can only process metric attributes. i.e., no categorical attributes should be present. |
| **Agglomerative** | Easy to use and implement, no need of information about no. of clusters. | Outliers can cause model less optimal. Time and Space complexity |

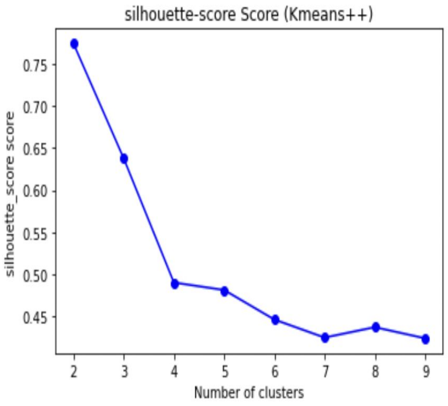


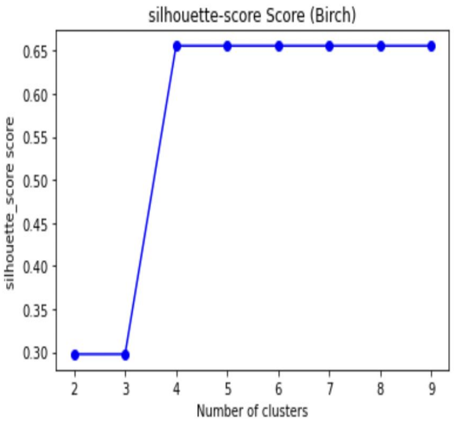
# Model Evaluation Methods

- Silhouette Coefficient (SC)
- Calinski-Harabasz score (CH)
- Davies-Bouldin score (DB)

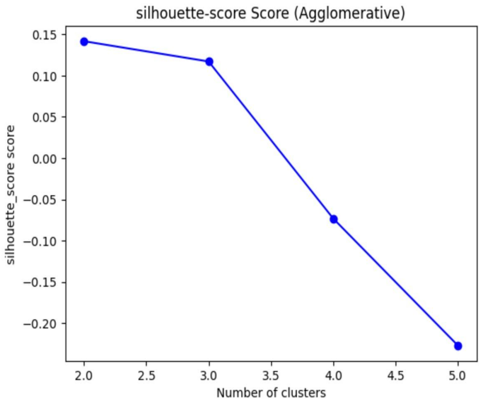
| Clustering Algorithm | SC Score | CH Score | DB Score |
|---|---|---|---|
| K-Means++ | 0.48034 | 2.26007e+06 | 0.70369 |
| MiniBatch | 0.48152 | 2.25916e+06 | 0.70162 |
| Birch | 0.67207 | 1.07556e+06 | 0.32108 |
| Agglomerative | -0.22688 | 29842.5 | 12.279 |

# Silhouette Coefficient

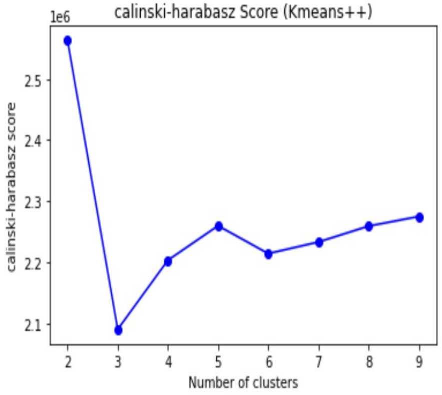


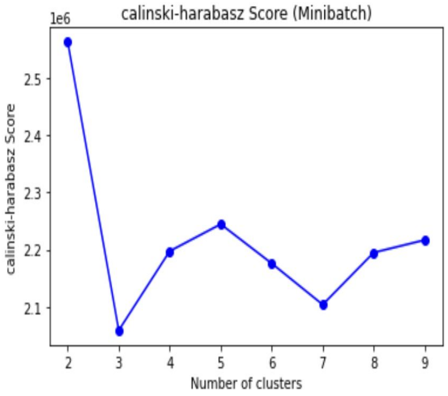



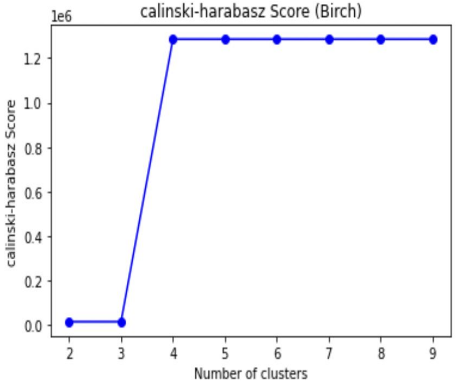


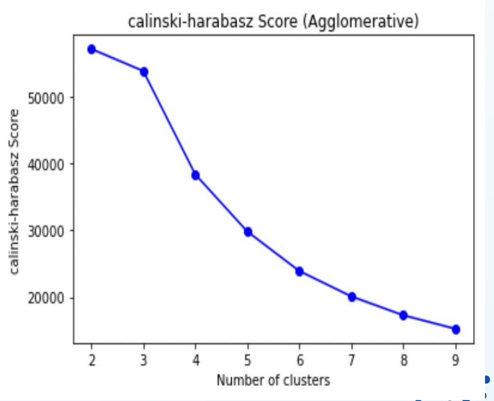


# Calinski-Harabasz coefficient

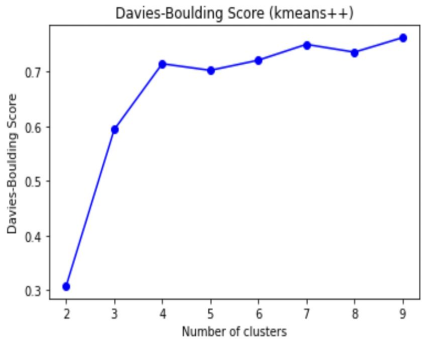


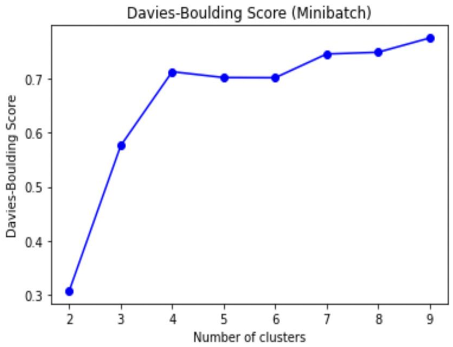



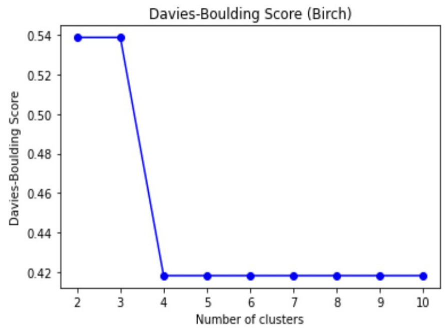



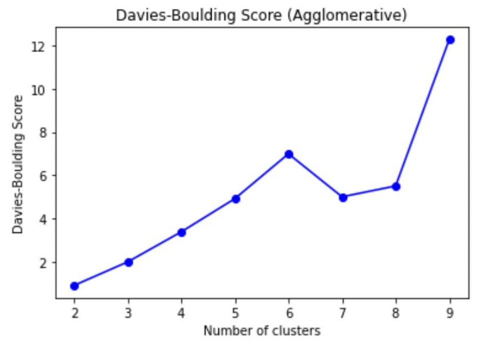




# Davies-Bouldin index



# Conclusions & Future Scope

- The purpose of our model is to fetch the similar images based on the user input image.
- With dimensionality reduction technique, by using only 50 components, we can keep around 95% of the variance in the data.
- We have performed four unsupervised ML models such as K-means++, MiniBatch, Birch, Agglomerative clustering models.
- Among the three metrics evaluated with these models, SC (0.67) & DB (0.3) scores are good for Birch where as CH score is less in comparison to K-means++ and MiniBatch.
- The research can extend by including more images in various categories  including human detection for images including persons.
- Also, the use case can be extended to  Deep Learning Algorithms and can verify their accuracy over conventional machine learning algorithms.

# Thank You!

## Any questions?

# References


[1] Pandey, S., & Khanna, P. (2016). Content-based image retrieval embedded with agglomerative clustering built on information loss. *Computers &Amp; Electrical Engineering*, *54*, 506–521. https://doi.org/10.1016/j.compeleceng.2016.04.003.

[2] Putri, D. C. G., & Leu, J. S. (2020b). Towards the Implementation of Movie Recommender System by Using Unsupervised Machine Learning Schemes. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 99–113. https://doi.org/10.1007/978-3-030-52988-8_9

[3] Pitolli, G., Laurenza, G., Aniello, L., Querzoni, L., & Baldoni, R. (2020). MalFamAware: automatic family identification and malware classification through online clustering. *International Journal of Information Security*, *20*(3), 371–386. https://doi.org/10.1007/s10207-020-00509-4

[4] Pitolli, G., Aniello, L., Laurenza, G., Querzoni, L., & Baldoni, R. (2017). Malware family identification with BIRCH clustering. 2017 International Carnahan Conference on Security Technology (ICCST), 1-6.

[5] ASIROGLU, B., ATALAY, M. I., BALKAYA, A., TUZUNKAN, E., Dagtekin, M., & ENSARI, T. (2019). Smart Clothing Recommendation System with Deep Learning. 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). https://doi.org/10.1109/ismsit.2019.8932738