DS-203 Project- Group-07

Shravani Kode(210070082)
Shristi Shrivastava(21d070069)
Dhrumil Lotiya(21d070026)

PROBLEM STATEMENT: Optimizing Architectural Design with Machine Learning

Design Family Identification:

Goal: Group designs by shape for better standardization.

Methods: Use clustering and deep learning to categorize similar designs.

Layout Complexity Classification:

Goal: Define complexity levels for better understanding.

Plan: Analyze designs and use machine learning to classify complexity.

Layout Retrieval for Faster Design:

Goal: Quickly find past designs based on parameters.

Solution: Develop a system to predict suitable design families.

Q1

Grouping in families

- Group images into families based on their features

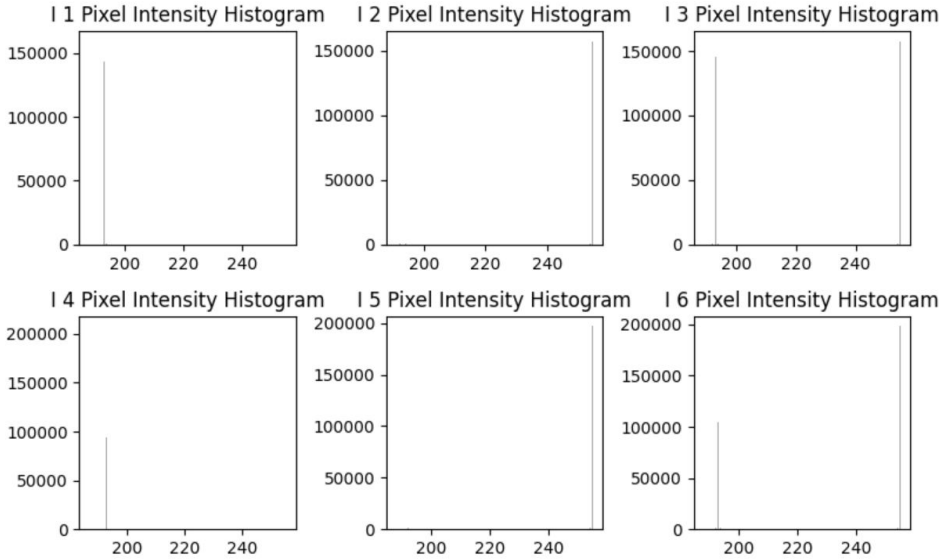
DATASET SUMMARY

- There are 1183 images in the dataset, each of dimension 640x480 pixels.
- Sample set of some images in the dataset are shown below:

DATASET ANALYSIS

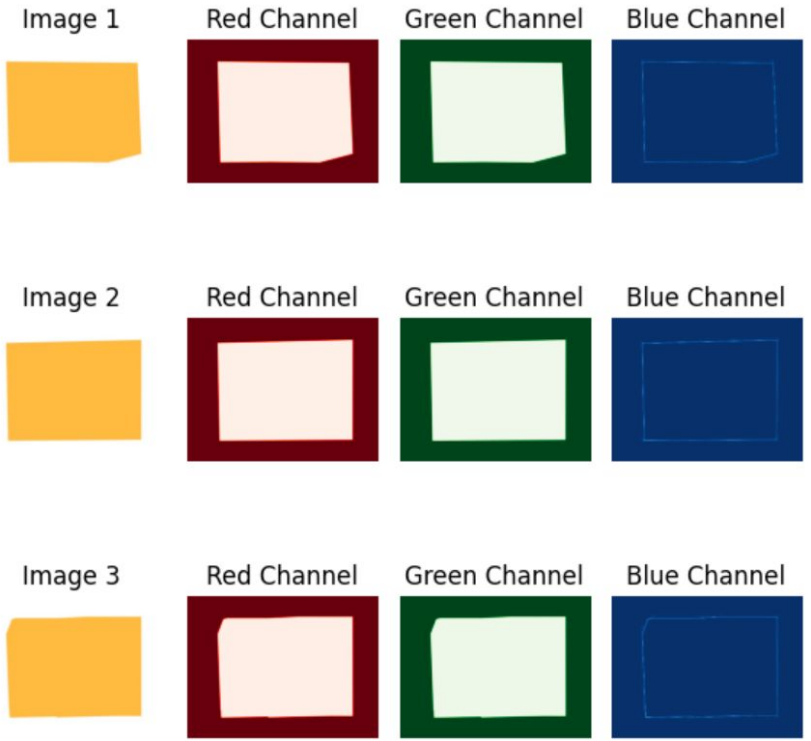
- By plotting histograms for a sample of images from the dataset, we observe the distribution of pixel intensity values across these images

- We observe that there are only two pixel intensity levels in the images



VISUALIZING COLOR CHANNELS

- Analyzing color channels provides insights into the composition and distribution of colors within images
- By visualizing RGB channels, we can identify prominent colors and their intensity levels
- Detected the composition of yellow color within the images
- Utilized the identified target color (yellow) in the code to effectively detect relevant features

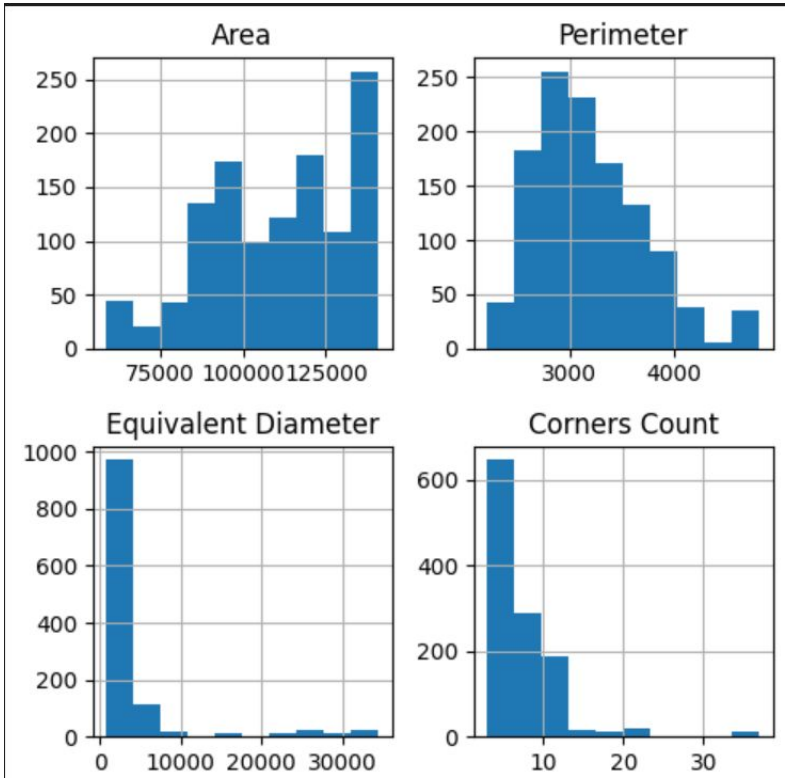


FEATURES USED FOR DECIDING FAMILY

- Area: Mean area of regions with the specified color, providing quantitative measures of the size and boundary of regions. This analysis reveals insights into the distribution and variation of object sizes within the dataset
- Perimeter: Total perimeter of regions with the specified color, offering further insights into the shapes and boundaries of objects
- Equivalent Diameter: Average diameter of regions with the specified color, serving as another size metric for the objects in the dataset
- Shape Descriptors: Extracted Hu moments to characterize shape features, providing detailed information about the shape properties of objects, useful for classification or further analysis
- Compactness: Ratio of perimeter to the square root of the area, providing a measure of how compact or spread out the shape of an object is

FEATURES USED FOR DECIDING FAMILY

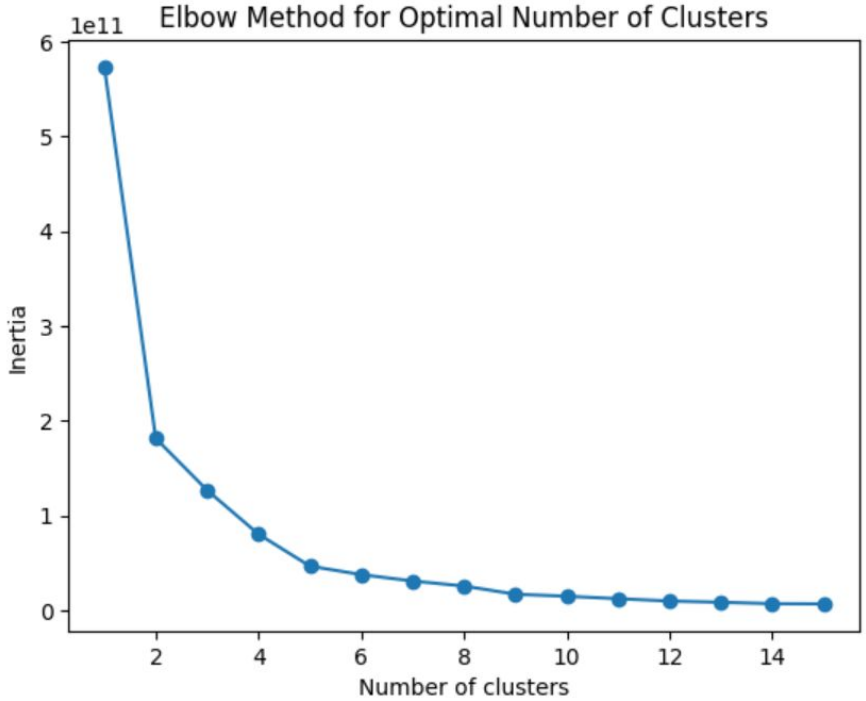
Distribution of some features across the dataset



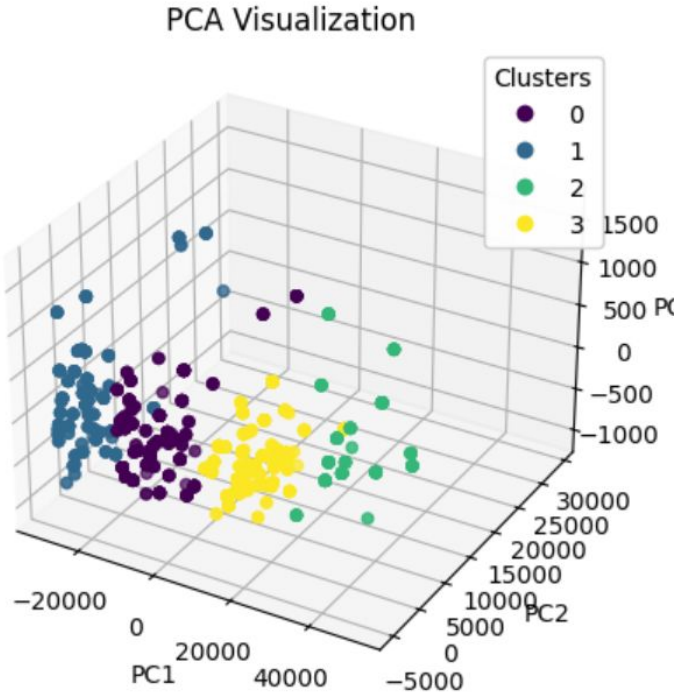
APPROACH 1- KMEANS CLUSTERING

After obtaining the required features, we performed k-means clustering to divide the dataset into families

Used Elbow-method to decide the optimum number of clusters

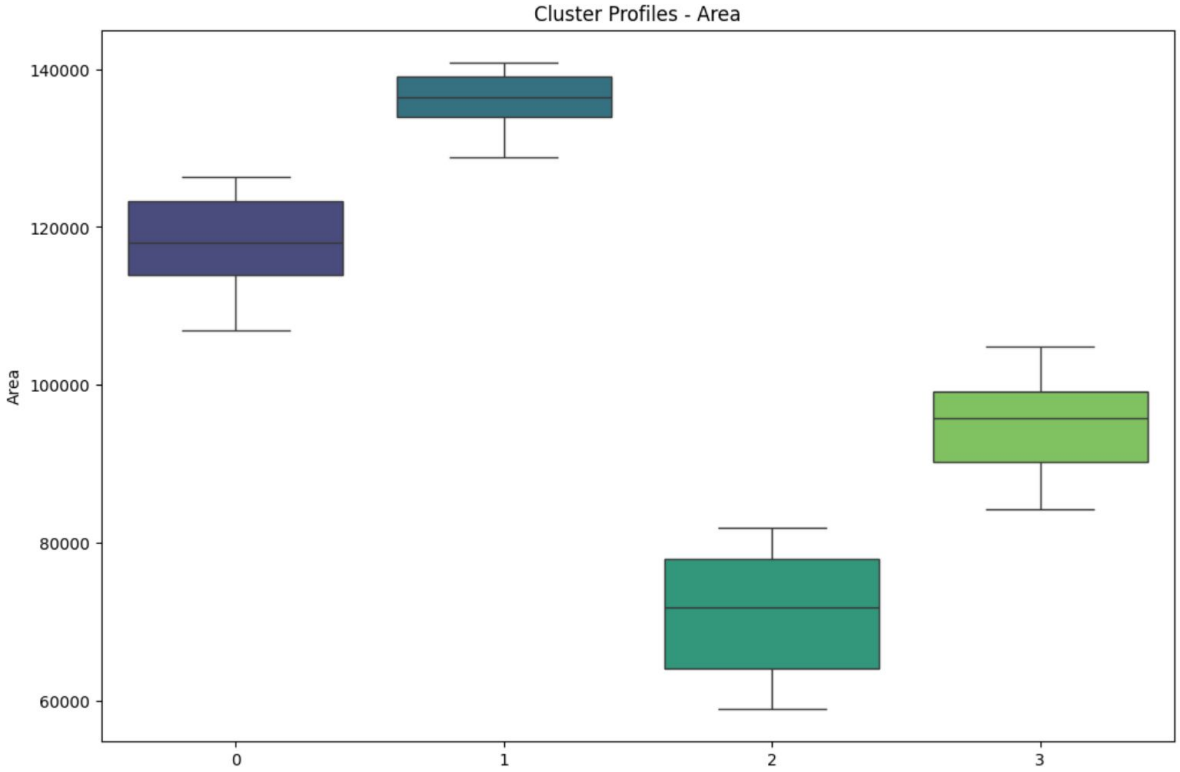


APPROACH 1- KMEANS CLUSTERING



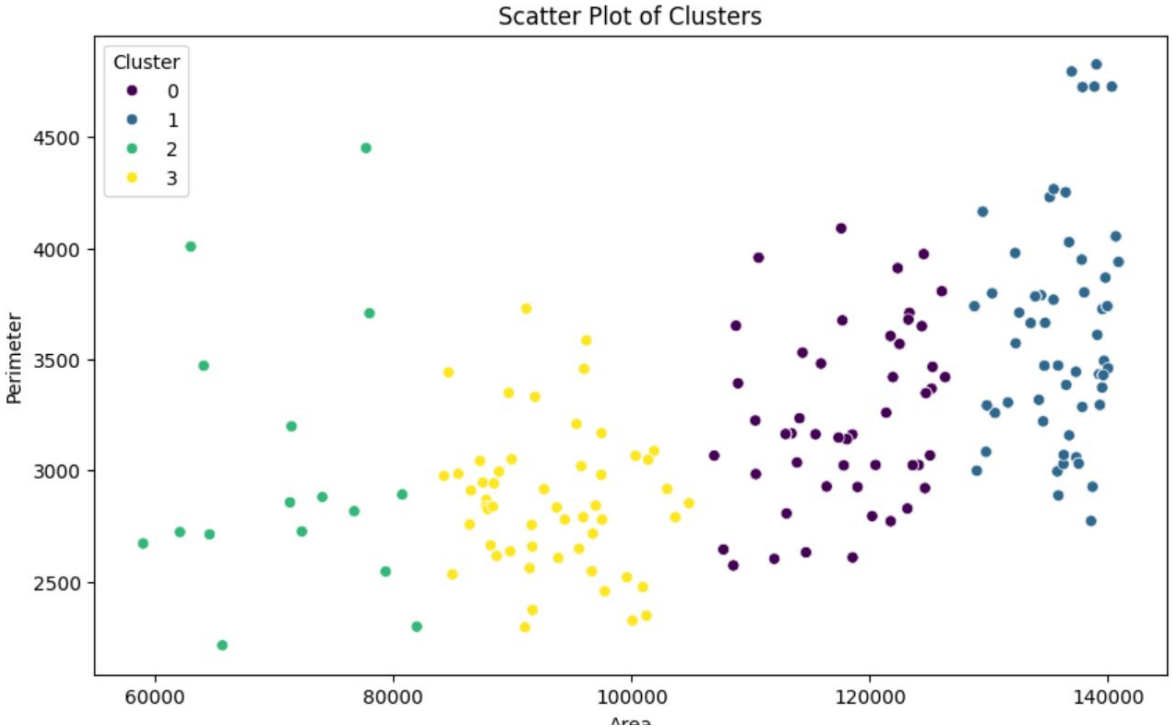
The data is divided into 4 families

RESULT VISUALIZATION



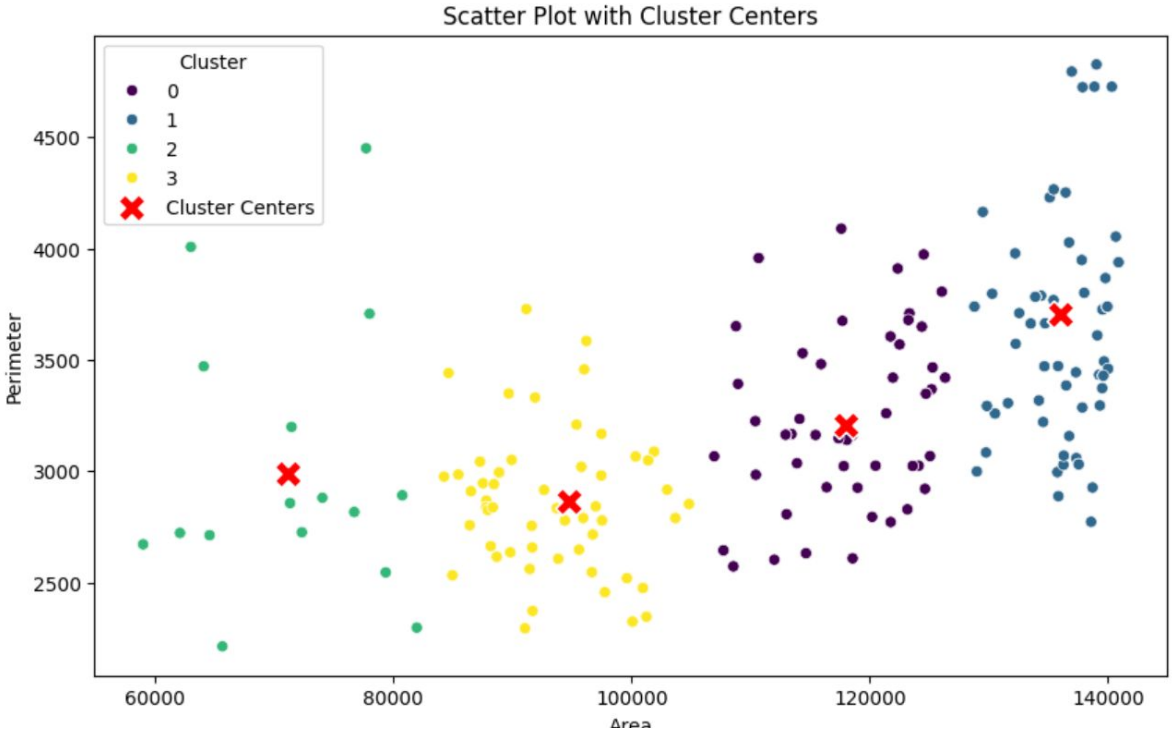
Visualizing the cluster profile: we observe that the clusters do not overlap much with each other

RESULT VISUALIZATION



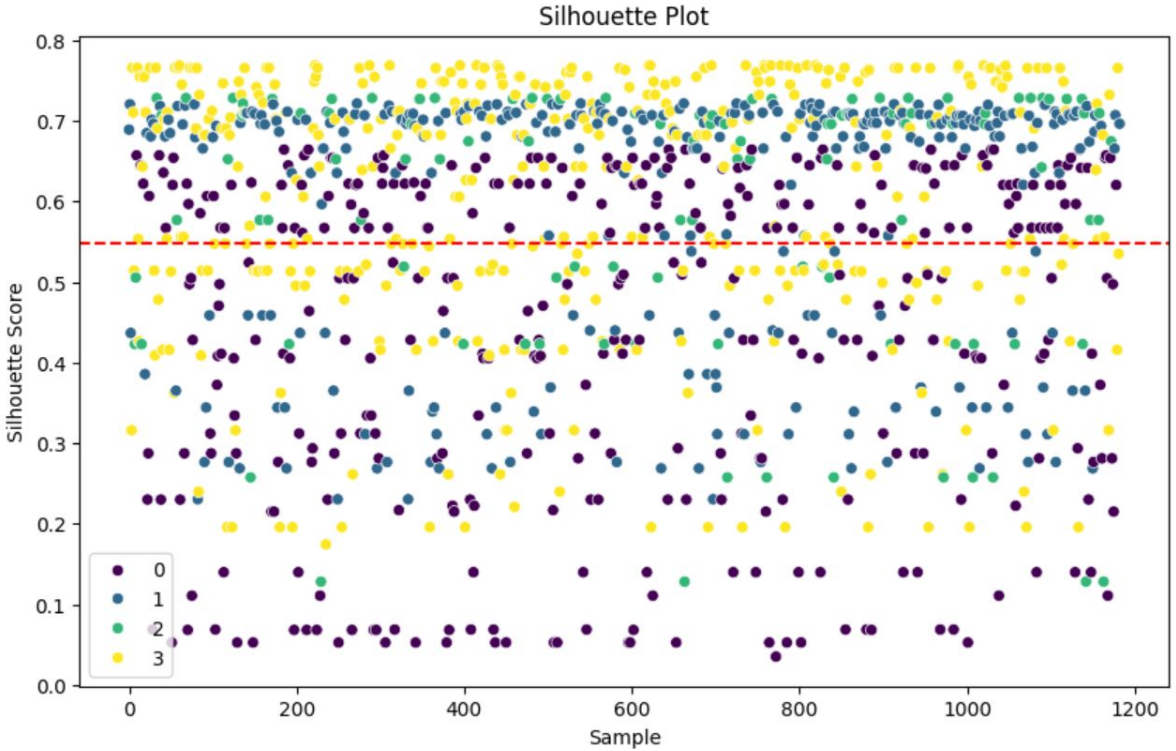
- This shows the scatter plot of the two features: Area and perimeter
- Here well-separated clusters exhibit distinct groupings in the scatter plot

RESULT VISUALIZATION



- This shows the scatter plot of the two features: Area and perimeter
- Here well-separated cluster exhibit distinct groupings in the scatter plot

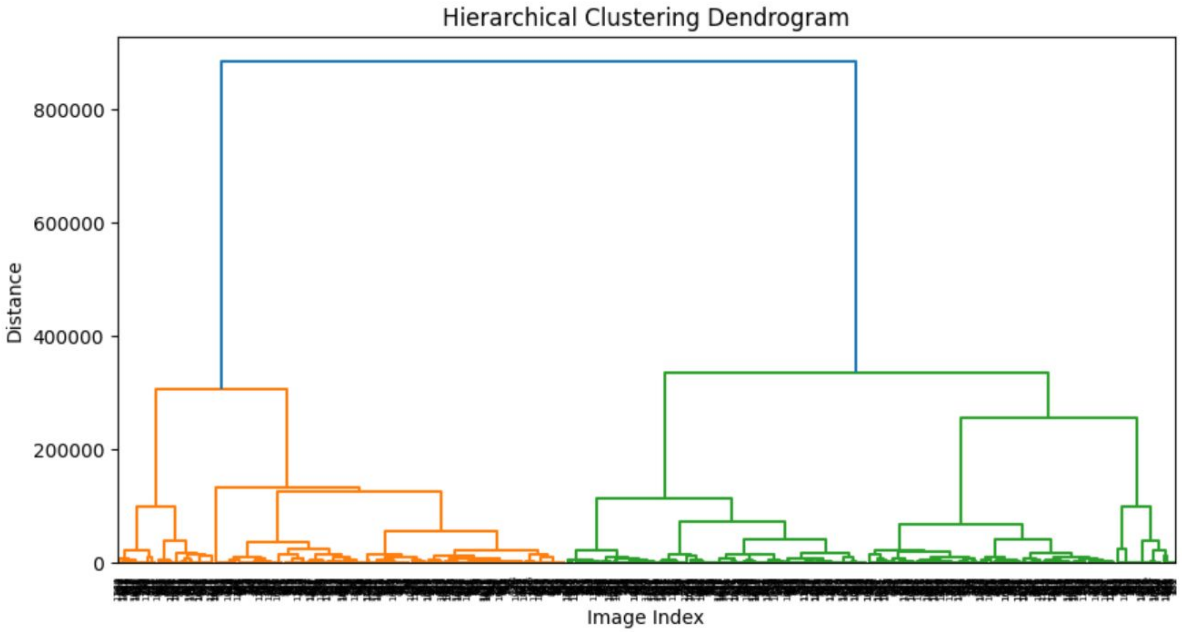
RESULT VISUALIZATION



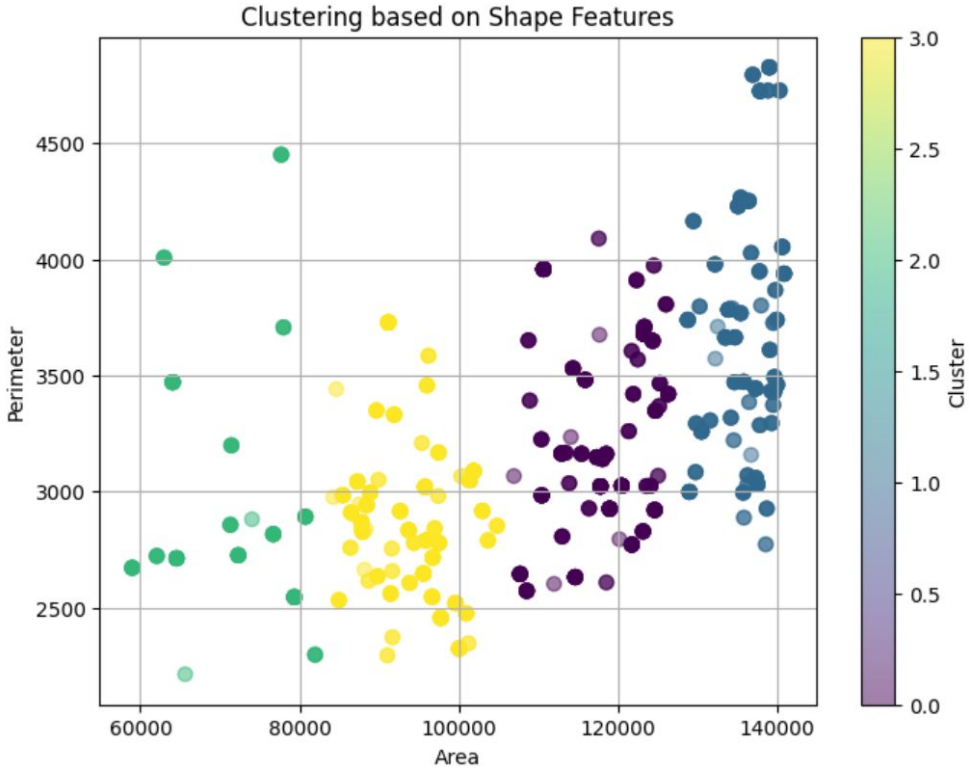
- Ideally, in a well-clustered dataset, most silhouette coefficients will be close to +1, indicating that the data points are well-clustered and are far from neighboring clusters
- Here the average Silhouette Score obtained was 0.5479767049661552

APPROACH 2- HIERARCHICAL CLUSTERING

After obtaining the required features, we performed hierarchical clustering to divide the dataset into families

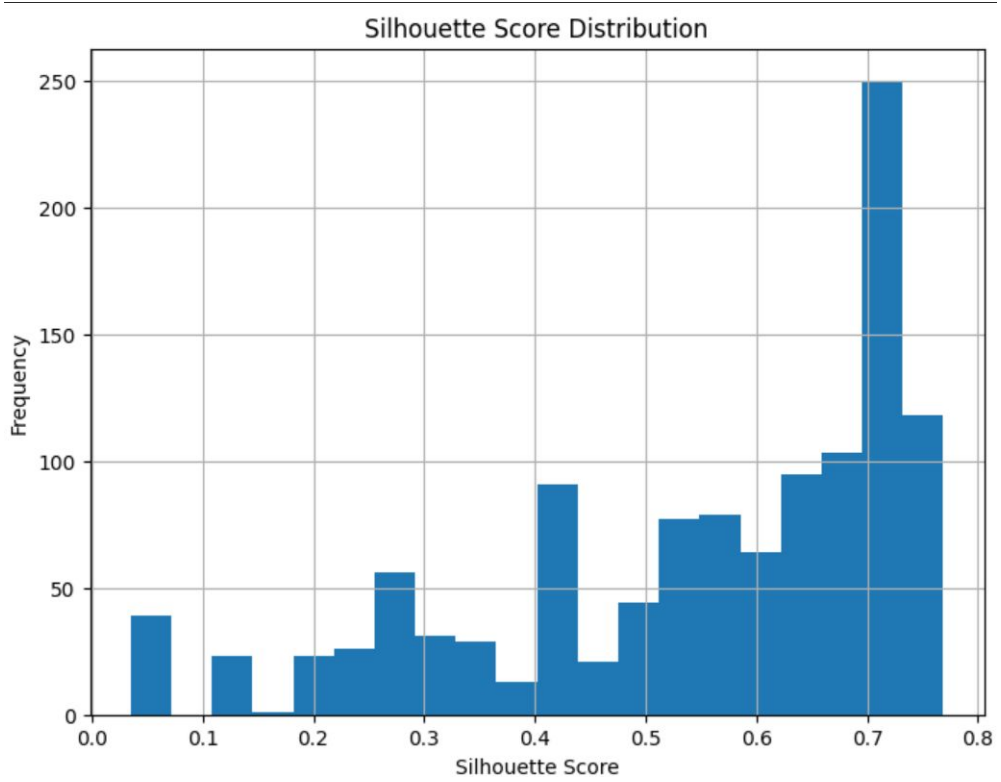


RESULT VISUALIZATION



- This shows the scatter plot of the two features: Area and perimeter
- Here well-separated clusters exhibit distinct groupings in the scatter plot

RESULT VISUALIZATION



- Ideally, in a well-clustered dataset, most silhouette coefficients will be close to +1, indicating that the data points are well-clustered and are far from neighboring clusters
- Thus we see that more data points have high silhouette score here

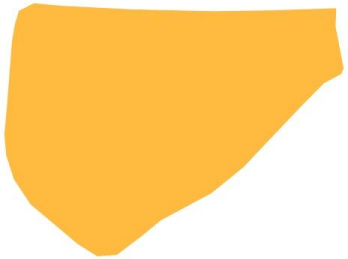
Q2

Complexity Classification

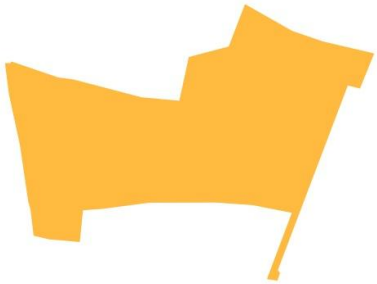
- High Complexity
- Medium Complexity
- Low Complexity

OUR UNDERSTANDING OF COMPLEXITY

(1) Low complexity image: It has a few corners and it almost completely fills the space



(2) Medium complexity image: It has a few twists and curves which makes it more complex than (1)



(3) High complexity image: It has a many twists and does not fill the space completely

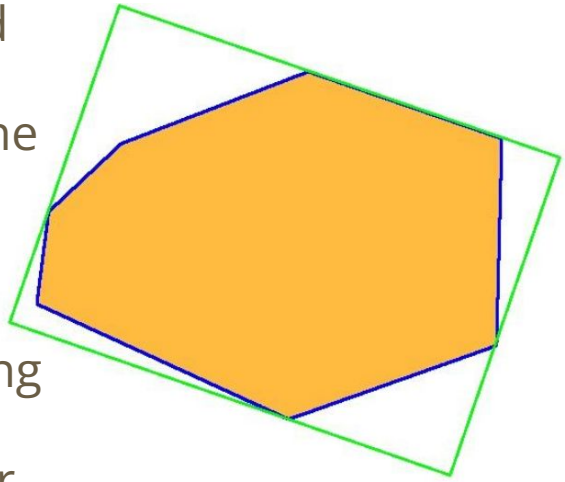
FEATURES USED FOR DECIDING COMPLEXITY

1. Elongation

   - Elongation refers to a measure of how stretched the shape is.
   - It's calculated by comparing the longer side of the bounding rectangle to the shorter side.

2. Extent

   - Extent is a measure of how much of the bounding box is filled by the contour.
   - It's calculated by dividing the area of the contour by the area of the bounding box.



Bounding rectangle - Tight fitting box

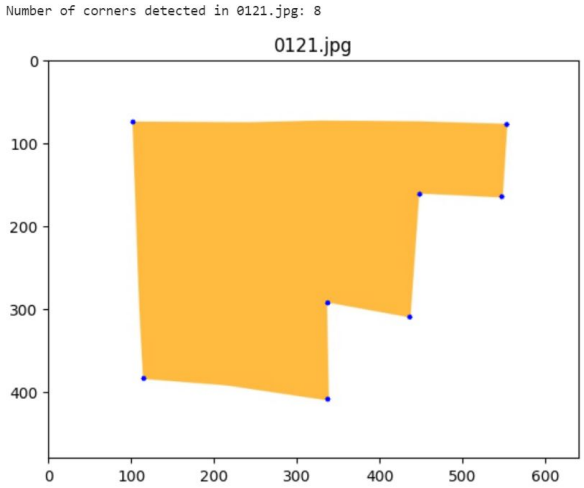
FEATURES USED FOR DECIDING COMPLEXITY

3. Compactness

- Compactness is a measure of how "compact" or "dense" a shape is.
- It's calculated by dividing the square of the perimeter by the area of the contour.

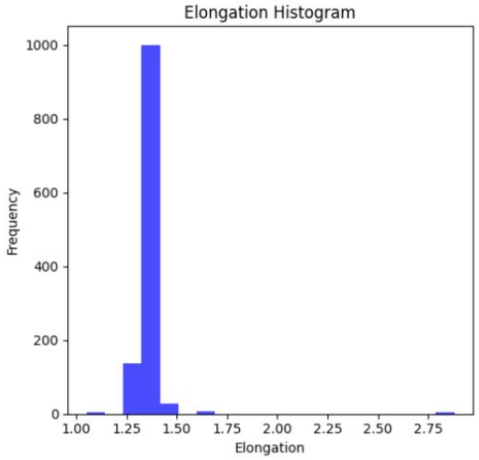
4. Number of corners

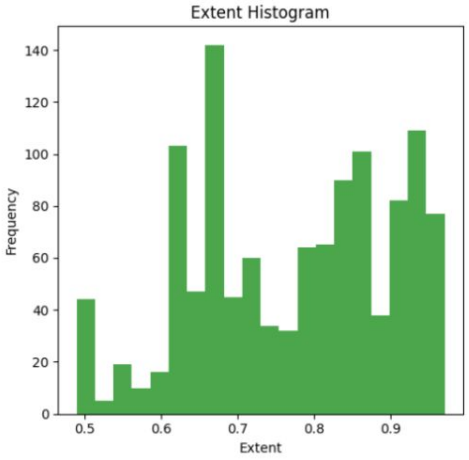
- Corners were detected using Shi- Tomasi corner detection method.

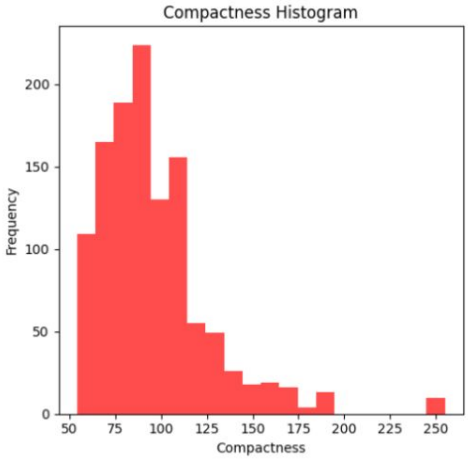


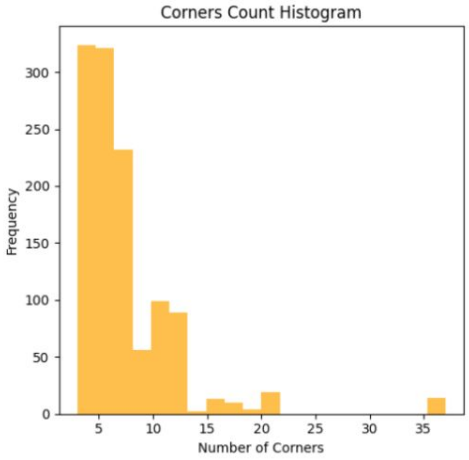
FEATURES ANALYSIS

Histogram of the features calculated for the images:







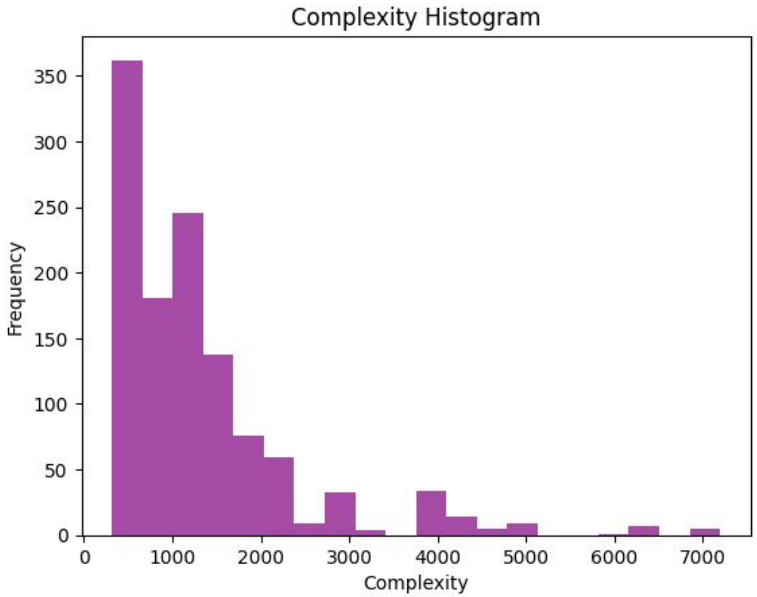


FEATURES ANALYSIS

As per the definition of each feature, complexity increases when :

- Elongation increases
- Extent decreases
- Compactness increases
- Number of corners increases

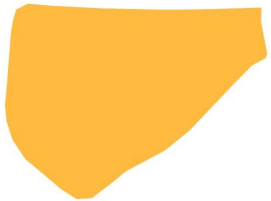
So, for each image we define complexity as:
elongation*corners_count*compactness/extent



COMPLEXITY ANALYSIS

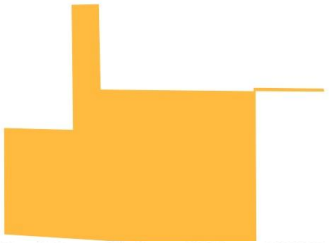
Complexity values obtained for some of the images:

Complexity of img 2 : 549.04



Complexity of img 112 : 2287.94

Complexity of img 63 : 3848.6



Complexity of img 991 : 4869.33

LOW VALUE

MEDIUM VALUE

HIGH VALUES

COMPLEXITY CRITERIA

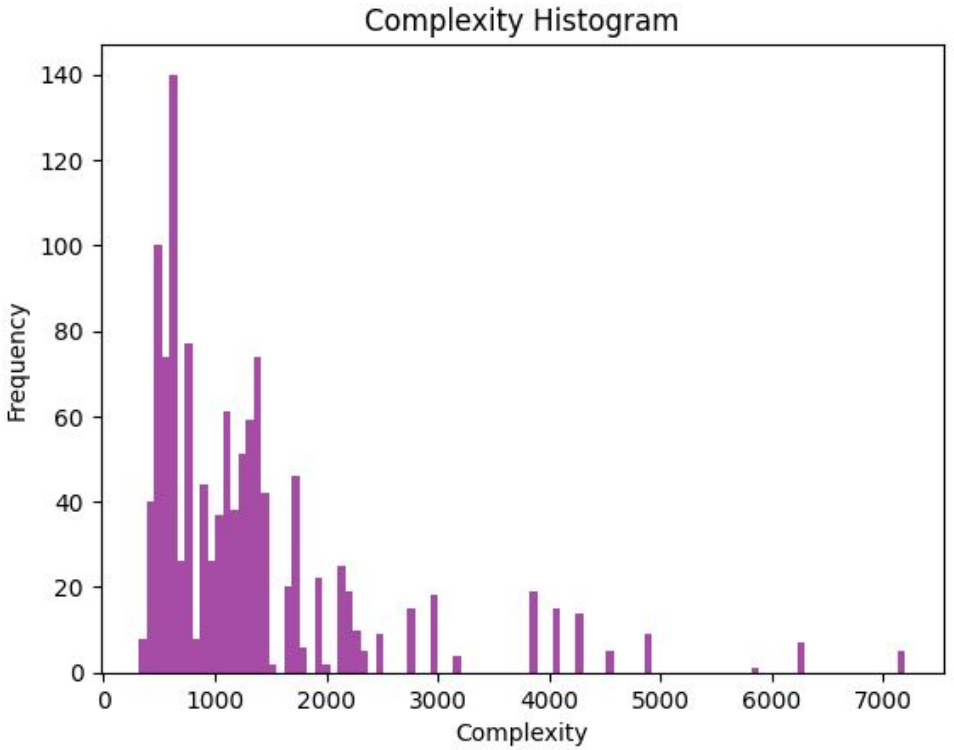
To find the threshold, we check the histogram plot of complexity:

Complexity > 3374 : High complexity

Complexity < 1350 : Low complexity

Else : Medium complexity

(Thresholds calculated by carefully checking images at the boundary value in histogram analysis and deciding their classification)



COMPLEXITY CLASSIFICATION

Based on the complexity value of each image, it is divided into low, medium, high complexity. Results:

```
Number of images in low complexity: 790
Number of images in medium complexity: 318
Number of images in high complexity: 75
```

(Results of which image has which complexity for all the 1183 images is in the python notebook, some of them were shown in the 18th slide)

Q3

Developing a ML model to predict suitable design families from layout parameters

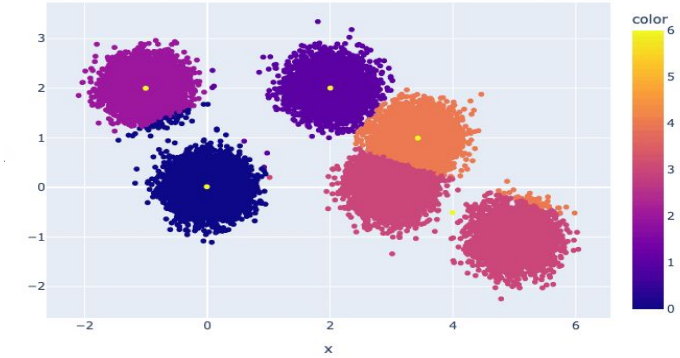
- Parameters which are input:
  - Layout area, length, width of tight fitting box.
  - Layout complexity
- Model will take these parameters and then predict the closest design family.

MODEL SELECTION

- Predict the design family of layout implies classification problem.
- Started with logistic regression. Observed that it was not giving satisfactory output .
- Problem is multi-class because and since we have formed 4 clusters, so this is 4-class classification.
- Switched to RandomForestClassifier model after doing some research on which model would be better for multi-class classification.
- Also used Support Vector Classification(SVC).

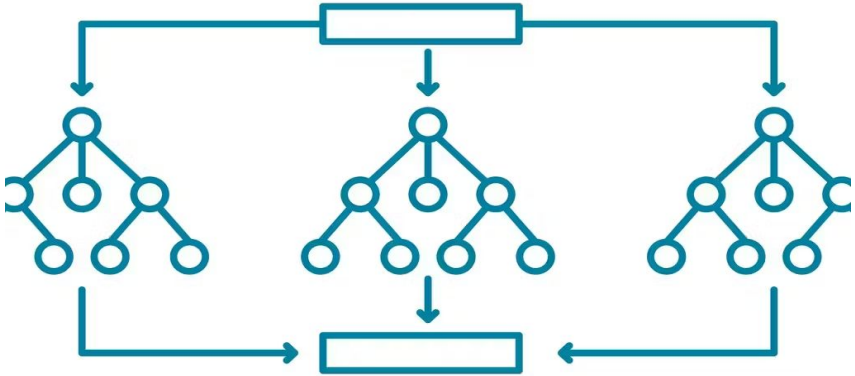
ALGORITHM

- Split the image dataset into train-test set.(80:20 ratio)
- Performed K-mean clustering on train dataset to obtain 4 cluster labels which are mapped to design-families directly.
- Used image processing function of cv2 to obtain tight box layout of the images and then correspondingly obtained its dimensions and area.
- Corresponding complexity was obtained from methodology used in Q2.

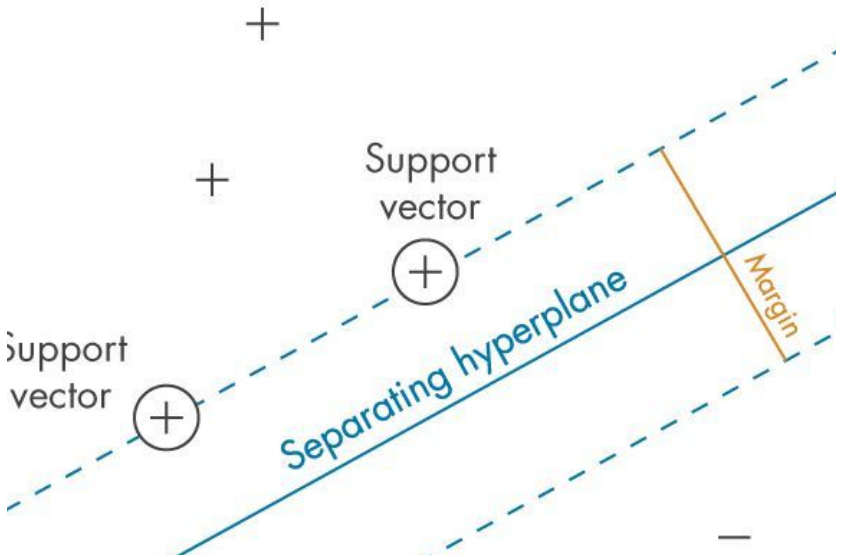


MODELS USED

- Used RandomForestClassifier to train the model using above features of training image dataset.
- Input to the model is tight box layout dimensions and area, image complexity.
- Output is the cluster which is directly mapped to layout family.



- Used SVC with RBF kernel to train the model since it also is one of the reliable algorithm for classification problems.
- Methodology same as Random Forest Classifier, i.e, used contour features to first create clusters and image complexity.
- Then trained model using tight box layout dimensions and area along with complexity as input.
- Output is cluster label mapped to layout family.



METRICS DEFINED

Accuracy Score:

What: Measures how many labels the model got right out of the total number of predictions.

Why: Helps gauge the model's explanatory power. Higher accuracy indicates better model prediction.

Confusion matrix:

What: How many of a classifier’s predictions were correct, and when incorrect, where the classifier got confused.

Why: Directly answers if it perform equally well for each class and were there any pairs of classes it found especially hard to distinguish.

Selection Reasoning:

- Combined, they ensure our model captures data variation (confusion matrix) and predicts with minimal deviation (accuracy), vital for multi-class classification.
- Accuracy score indicate the proportion of correctly predicted design families out of all predictions made. It gives an overall understanding of how well the model is performing in classifying layouts into the correct design families based on the given parameters.
- However since our dataset has class imbalance, accuracy may not provide a complete picture of model performance. Hence, comes confusion matrix.
- The confusion matrix tells us exactly where mistakes were made.

Combining both mitigates the risk of over-relying on one metric, ensuring robust model validation.

***Note:-** Metrics are calculated on the test dataset.

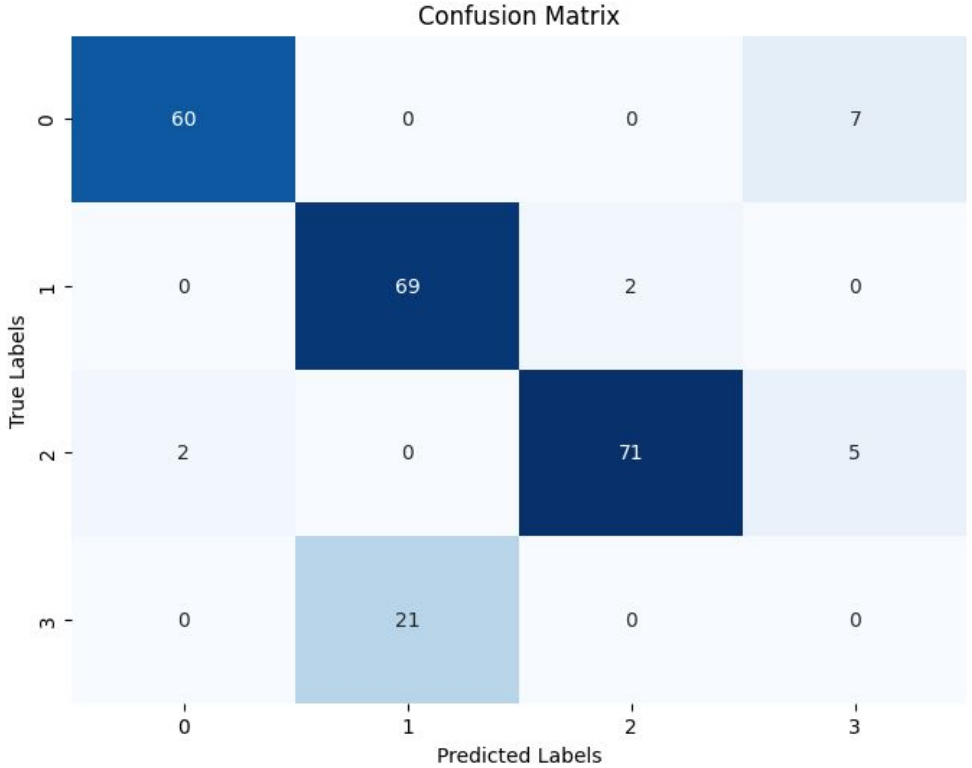
METRICS FOR MODELS

- **Random Forest Classifier:**

  Accuracy: 0.8438818565400844

Accuracy is out of 1, so equivalent to 84 % accuracy.

From confusion matrix, we can see that model can predict correctly for all families except for just family 4(label 3)



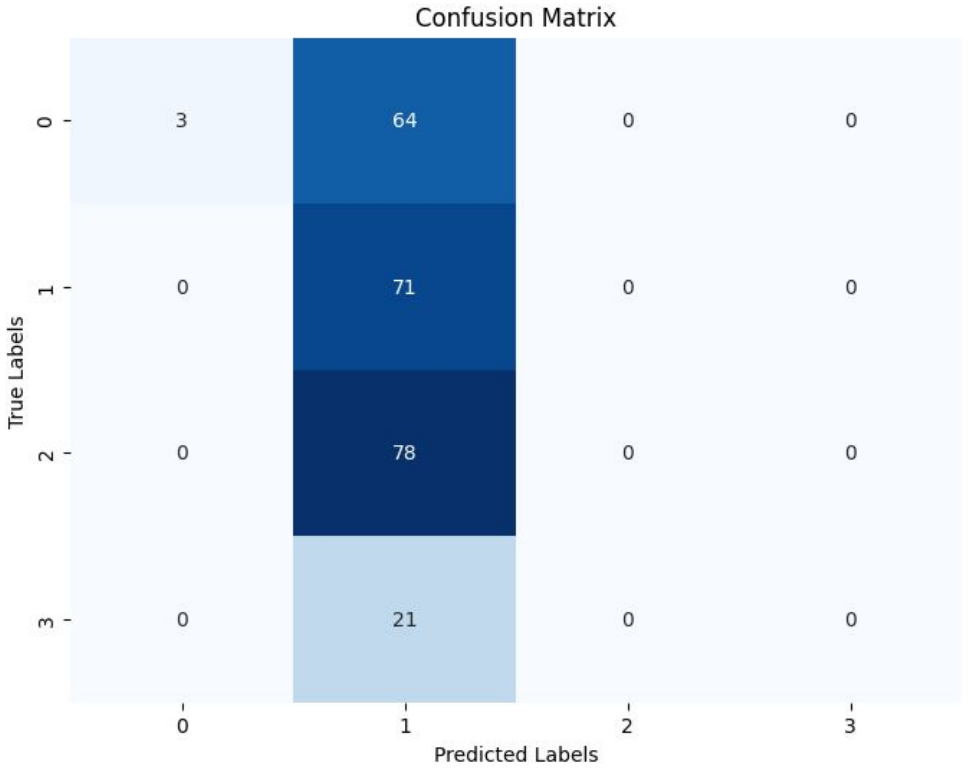
- **Support Vector Classification:**

  Accuracy: 0.3122362869198312 4

This means accuracy is 31%.

Confusion matrix is very poor, as model is always predicting label 1 irrespective of true label. Hence, not reliable.

Hence, we can see that Random Forest Classifier is more reliable model for the problem at hand



SUMMARY OF MODEL

- Final model used is Random Forest Classifier.
- Image dataset split into train-test subset(80-20 ratio).
- Performed K-Means clustering on train dataset and obtained 4 families.
- Trained model with input as gross parameters and output as family
- Similarly, performed K-Means on test set and then used model with input as gross parameters of test set.
- Prediction reliability checked by different metrics like accuracy score and confusion matrix.

CONCLUSIONS

- LEARNINGS
- CHALLENGES
- ANSWERS TO QUESTIONS IN EVALUATION CRITERIA

KEY LEARNINGS

- Analysis of various properties of images
- Distinguishing features of shapes
- Behaviour of different ML models for multi-class classification
- Creating model with just 4 gross parameters as input and predicting family, created from clustering, hence giving closest design layout

CHALLENGES FACED

- Deciding the most relevant features of the image for grouping the images into families
- Deciding threshold value for complexity analysis was challenging, since there was no set labels given for the image which ones are of low, medium or high complexity. We had to look at the images and their complexity histograms to decide that.
- Selecting optimum model and corresponding hyperparameters for best result was a challenge

ANSWERS TO QUESTIONS IN EVALUATION CRITERIA

• Are all the questions posed by the problem statements effectively addressed with solutions?

Proper and complete approach and results to all the questions are described in details in the slides.

• Are the solutions relevant, and correctly applied, and backed up with proper logic / reasons?

The features used for deciding family/complexity were properly analysed and none of the features were used blindly. Hence they are relevant. Results indicate they are correctly applied. Logic of using them are explained in the slides. Validity of model used is backed by its metrics (confusion matrix and accuracy score)

ANSWERS TO QUESTIONS IN EVALUATION CRITERIA

• Has there been any creative thinking and innovation while solving the problems?

What features to include and what not to while grouping into families and complexities were creatively thought out. The thresholds for complexity were decided based on our own analysis.

• Have any possibilities, other than those asked for, been implemented, or listed in the presentation?

We tried to implement CNN to extract features and then k-means to divide them into clusters

ANSWERS TO QUESTIONS IN EVALUATION CRITERIA

• Are the major steps of data analysis diligently followed and correctly applied and documented (wherever required ...)?

We have analysed the data to get the relevant features for further classification and prediction and is shown in slides for Q1.

• Quality of results: are they backed with appropriate metrics, comparisons, analysis, explanations, and justifications?

Quality of results is shown by the metrics like confusion matrix and accuracy score. Comparison between different model is shown on the basis of above metric and corresponding analysis of metric given and model selection is justified.