Automatic segmentation of clothing for the identification of fashion trends using k-means clustering

*

VISHAL VASOYA *202118013*
VASHISHTH RAVAL *202118026*
BHUMI BOSAMIA *20118040*
DEVANSHI SHAH *202118042*

***Abstract*—Fashion is a fast-growing industry. This project proposes an automatic clothing splitting method. After isolating the images, fashion trends were identified based on color, texture (complex material), and shape. These trends represent true fashion clusters, but further studies with supervised learning models need to be completed.**

I. Introduction

The data set consists of a training set of 60,000 observations and a test set of 10,000 observations. The first column consists of class labels and represents clothing. The remaining columns contain the pixel values of the associated image.

II. Model Pipeline

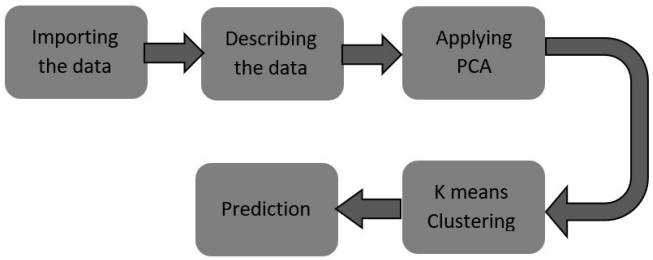

Fig. 1. Model Pipeline

III. Methods

We have started with Exploring and processing the data. We first standardized it to scale all the features in same range and reduced the dimension using PCA for fast computation. Then we have find the clusters for scaled data with PCA and scaled data without PCA.

A. KMeans Clustering

KMeans Clustering is an Unsupervised Learning algorithm, which helps in grouping the unlabeled dataset into different clusters.

"IT IS AN ITERATIVE ALGORITHM THAT DIVIDES THE UNLABELED DATASET INTO K DIFFERENT CLUSTERS IN SUCH A WAY THAT EACH DATASET BELONGS TO ONLY ONE GROUP THAT HAS SIMILAR PROPERTIES."

It allows us to cluster the data into different groups and is a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes an untagged data set as input, divides the data set into a certain number of clusters, and repeats the process until it finds the best cluster. In this algorithm, the value of k must be given in advance.

The K-means clustering algorithm basically does two things.

- Use an iterative process to determine the best value for the K centroid or centroid.
- Assign the closest kcenter to each data point. Data points near a particular kcenter create clusters.

Hence each cluster has data points with some commonalities, and it is away from other clusters. The below diagram explains the working of the K-means Clustering Algorithm:

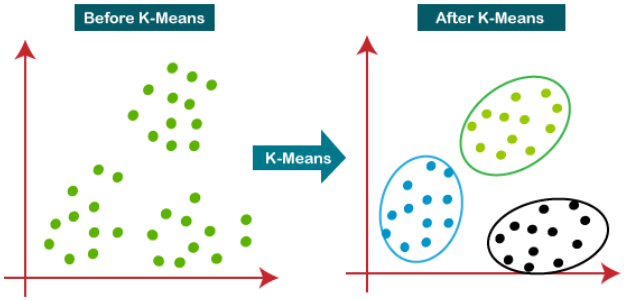

Fig. 2. K Means

B. Elbow Method:

The elbow method is one of the most widely used methods for finding the optimal number of clusters. This method uses the concept of WCSS values. WCSS represents the sum of squares within a cluster and determines the total number of variants within the cluster. The formula for calculating the WCSS value is:

$$WCSS = \sum Pi \text{ in cluster } 1(PiC1) + \sum Pi \text{ in cluster } 2(PiC2) + \sum Pi \text{ in cluster } 3(PiC3) ... + \sum Pi \text{ in cluster } n(PiCn)$$

Pi in Cluster1 Distance(Pi) C1)2: is the sum of the squared distances between each data point and the centroid of cluster1, the same for the other two terms. . To measure the distance between the data point and the centroid, you can use methods such as the Euclidean distance or the Manhattan distance. To find the optimal clustering value, the elbow method performs the following steps: o K-means clustering on a given data set for various K values (ranging from 110). Calculate the WCSS value for each value of K. o Plot the curve between the calculated WCSS values and the number of clusters K. o If a sharp bend point or graph point looks like a shoulder, this point is considered the best K value. Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method. The graph for the elbow method looks like the below image:

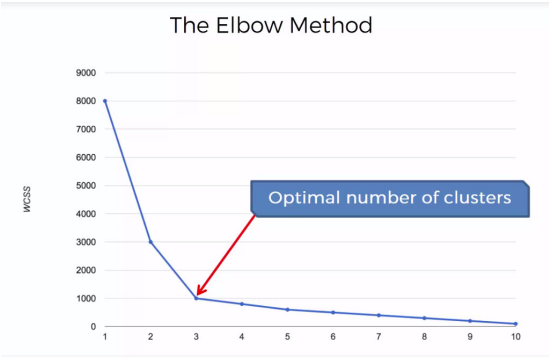

Fig. 3. K Means

C. Logistic Regression:

- Logistic Regression can be applied to binary or more classification problems. It uses a logistic function which is a sigmoid function. Logistic regression is easier to implement, interpret, and very efficient to train, and also logistic regression is better than linear regression because the sigmoid curve fits better to the data than a straight line. But the major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In other words, non-linear problems can't be solved with logistic regression because it has a linear decision surface.
- An alternative approach is to modify the logistic regression model to directly support multi-class label prediction. Specifically, it predicts the probability that an input example belongs to all known class labels. The model accuracy on the dataset is 0.8439.

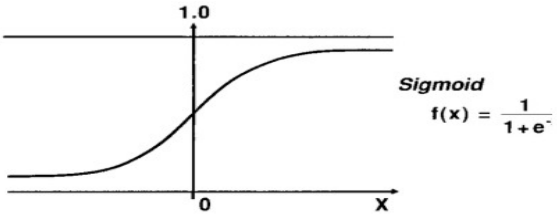

Fig. 4. Logistic Regression

D. SVM:

Models obtained using support vector classification depend only on a subset of the training data.SVM builds hyperplanes in multidimensional space to isolate different classes. The support vector is the closest data point to the hyperplane. This point calculates the margins to better define the dividing line. This has more to do with the construction of the classifier. The model has an accuracy of 0.84 on the data set.

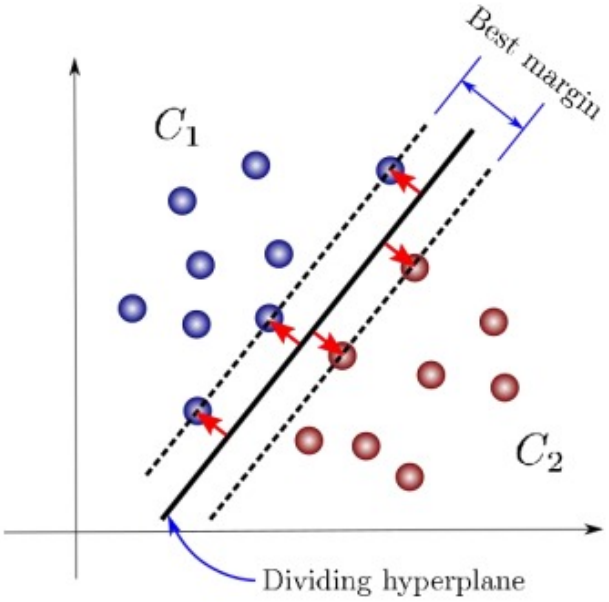

Fig. 5. SVM

E. Decision Tree:

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.

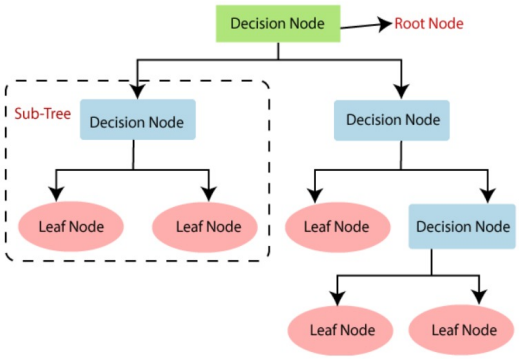

Fig. 6. Decision Tree

F. Random Forest:

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree,the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The accuracy of the model on the dataset is 0.8685

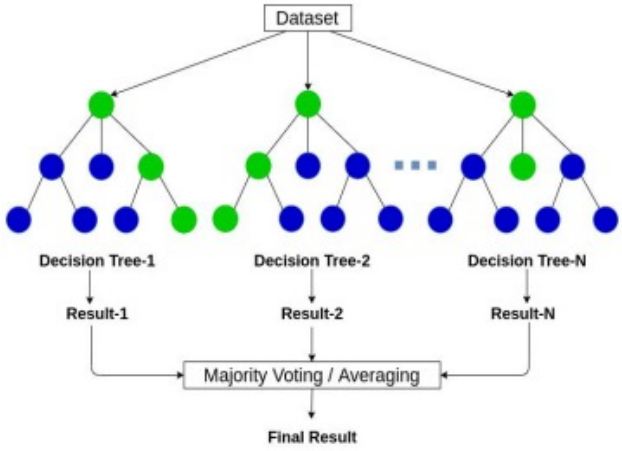

Fig. 7. Random Forest

G. CNN

The last method we employed was CNN. As the images were in grayscale, we applied only one channel. We selected the following architecture: Two convolutional layers with 32 and 64 filters, 3 × 3 kernel size, and relu activation. The polling layers were chosen to operate tiles size 2 × 2 and to select the maximal element in them. Two sets of dense layers, with the first one selecting 128 features, having relu and softmax activation. For loss function, we chose categorical cross-entropy. To avoid overfitting, we have chosen 9400 images from the training set to serve as a validation set for our parameters. We used novel optimizer adam, which improves over
standard gradient descent methods and uses a different learning rate for each parameter and the batch size equal to 64. The model was trained in 50 epochs. We present the accuracy and loss values in the graphs below.
We see that the algorithm converged after 15 epochs and that it is not overtrained, so we tested it. The obtained testing accuracy was equal to 89%, which is the best result obtained out of all methods!
Before proceeding to other methods, let's explain what have the convolutional layers done. An intuitive explanation is that the first layer was capturing straight lines and the second one curves. On both layers we applied max pooling, which selects the maximal value in the kernel, separating clothing parts from blank space. In that way, we capture the representative nature of data. In other, neural networks

perform feature selection by themselves. After the last pooling layer, we get an artificial neural network. Because we are dealing with the classification problem, the final layer uses softmax activation to get class probabilities. As class probabilities follow a certain distribution, cross-entropy indicates the distance from the network's preferred distribution.The model accuracy on the dataset is

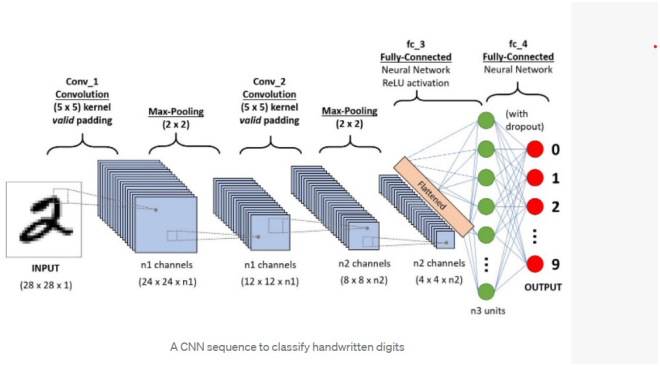

Fig. 8. CNN

IV. RESULT

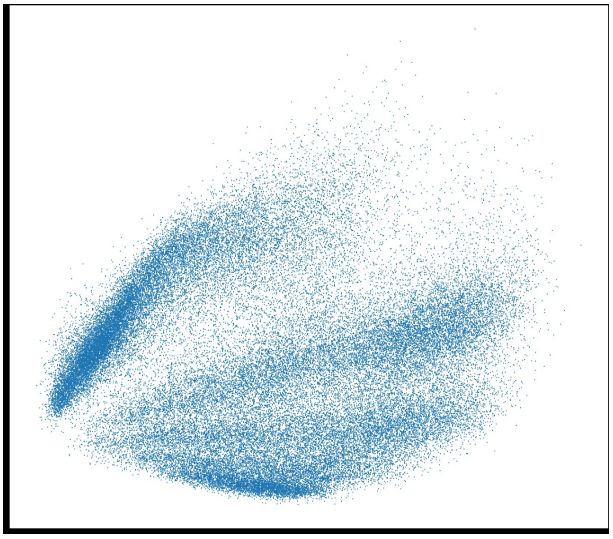

Fig. 9. Data Points

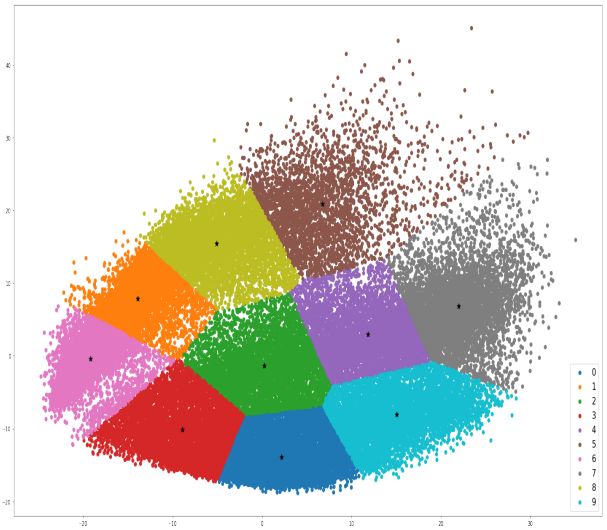

Fig. 10. Cluster With Centroid Using K Means

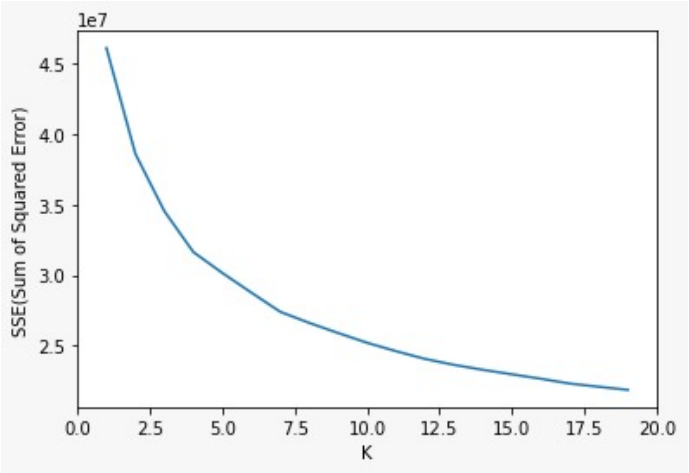

Fig. 11. Caption

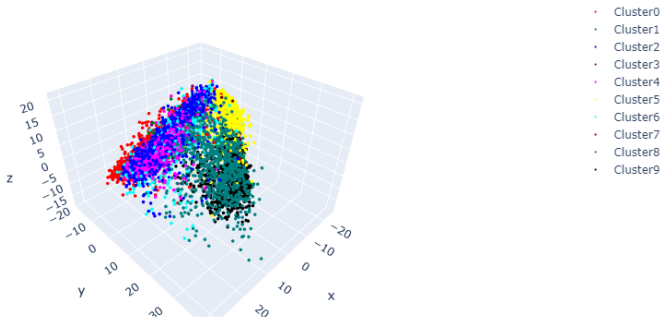

Fig. 12. Logistic Regression

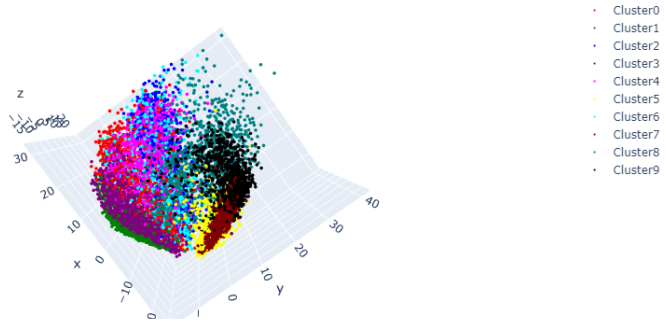

Fig. 13. Logistic Regression

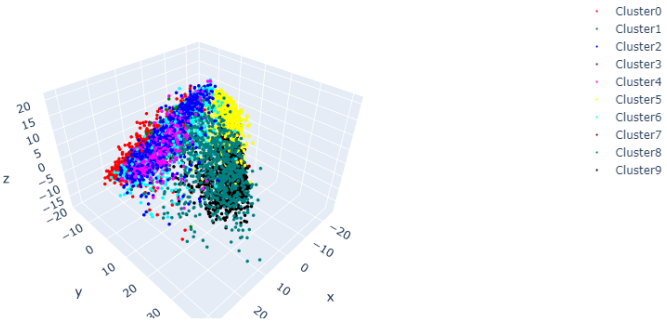

Fig. 14. SVM

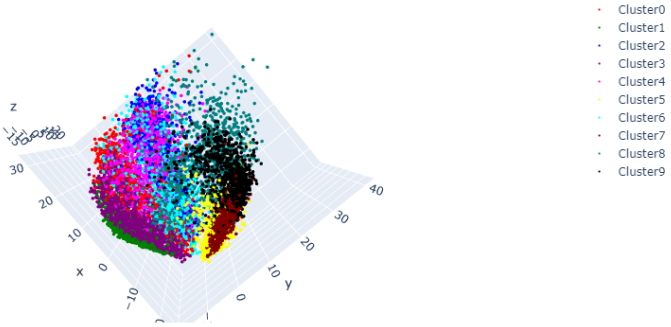

Fig. 15. SVM

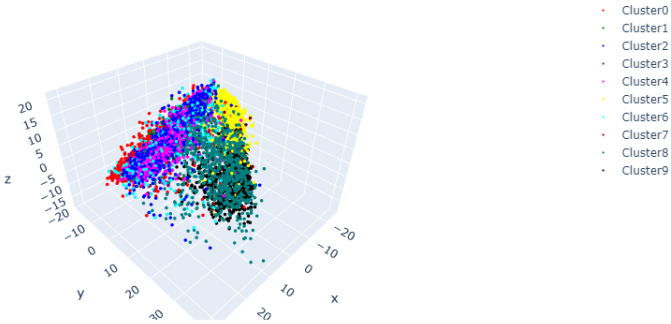

Fig. 16. Decision Tree

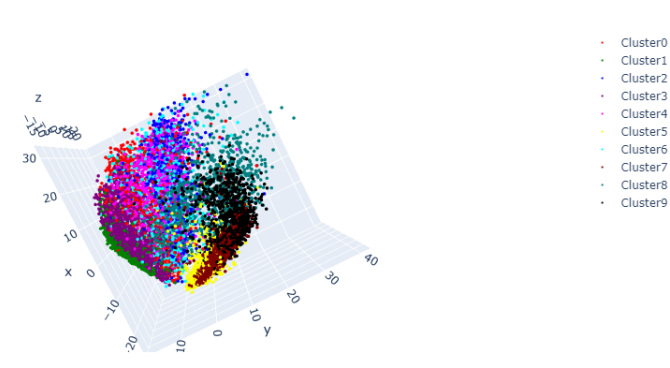

Fig. 17. Decision Tree

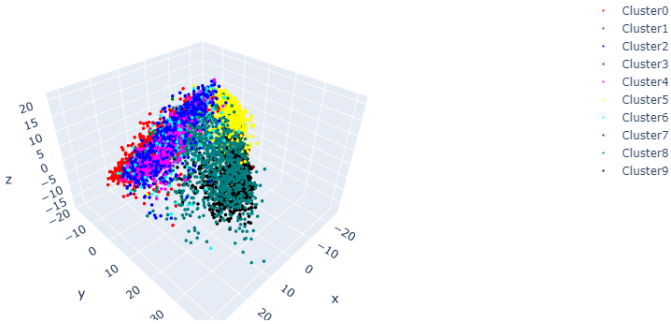

Fig. 18. Random Forest

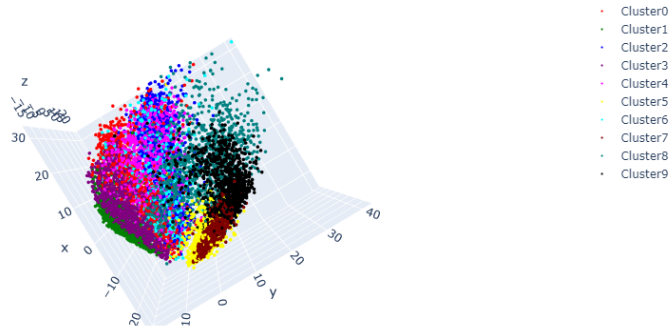

Fig. 19. Decision Tree

Models	With PCA				Without PCA			
	accuracy	precision	recall	f1 score	accuracy	precision	recall	f1 score
K means	0.68	0.55	0.84	0.66	0.68	0.68	0.70	0.69
Logistic	0.84	0.87	0.81	0.81	0.84	0.81	0.81	0.80
SVM	0.87	0.83	0.82	0.84	0.87	0.83	0.82	0.84
Decision Tree	0.75	0.71	0.70	0.71	0.79	0.75	0.74	0.76
Random Forest	0.83	0.78	0.82	0.74	0.86	0.83	0.84	0.82
CNN	-	-	-	-	0.9174	0.97	0.28	0.43

V. CONCLUSION

In this project, we have performed five classification algorithms and then applied voting classifier (both hard and soft) using them. We also performed KNN classifier algorithm. From the table shown above, an ensemble of the first five algorithms, i.e. voting classifier gives the best accuracy, precision, recall and F1 score and SVM gives maximum recall and F1 score.We have obtained fairly accurate predictions on whether a person will click on an ad or not. This can be used by many companies/organizations who want to advertise themselves/ their products globally.

VI. REFERENCES:

[1] DATASET - https://www.kaggle.com/datasets/zalando-research/fashionmnist

[2] Scikit learn for machine learning- https://scikit-learn.org/stable/ Scikit learn for machine learning-

[3] Logistic Regression- https://scikit-learn.org/stable/modules/generated/sklearn.linear$_{m}odel.LogisticRegression.html$

[4] SVM - https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[5] Random Forest- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[6] Decision Tree - https://towardsdatascience.com/understandingdecision-tree-classifier-7366224e033b

[7] K Means- https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[8] CNN - https://towardsdatascience.com/image-classification-with-fashion-mnist-why-convolutional-neural-networks-outperform-traditional-df531e0533c2