Fruits classification

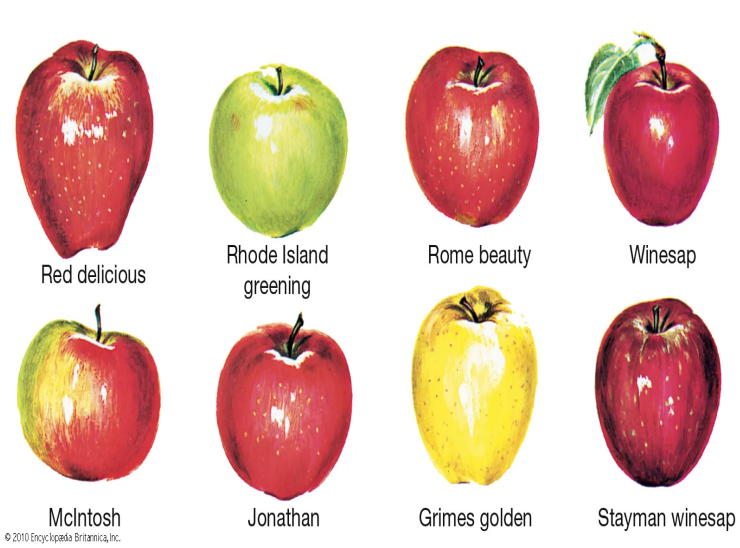
Group-16:Pradeep Kumar, Pranay, Shubham Kumar
B20AI029, B20AI030, B20AI039

Abstract

This paper reports our experience with building a Fruit Classifier. We have a dataset generated using the filmed images using a Logitech C920 camera. Fruits and vegetables were planted in the shaft of a low-speed motor (3 rpm) and a short movie of 20 seconds was recorded. Behind the fruits, a white sheet of paper was placed as a background. We use various classification algorithms and compare their results in this report.

INTRODUCTION

The following fruits and vegetables are included in the dataset: **Apples (different varieties: Crimson Snow, Golden, Golden-Red, Granny Smith, Pink Lady, Red, Red Delicious), Apricot, Avocado, Avocado ripe, Banana (Yellow, Red, Lady Finger), Beetroot Red, Blueberry, Cactus fruit, Cantaloupe (2 varieties), Carambula, Cauliflower, Cherry (different varieties, Rainier), Cherry Wax (Yellow, Red, Black), Chestnut, Clementine, Cocos, Corn (with husk), Cucumber (ripened), Dates, Eggplant, Fig, Ginger Root, Granadilla, Grape (Blue, Pink, White (different varieties)), Grapefruit (Pink, White), Guava, Hazelnut, Huckleberry, Kiwi, Kaki, Kohlrabi, Kumsquats, Lemon (normal, Meyer), Lime, Lychee, Mandarine, Mango (Green, Red), Mangostan, Maracuja, Melon Piel de Sapo, Mulberry, Nectarine (Regular, Flat), Nut (Forest, Pecan), Onion (Red, White), Orange, Papaya, Passion fruit, Peach (different varieties), Pepino, Pear (different varieties, Abate, Forelle, Kaiser, Monster, Red, Stone, Williams), Pepper (Red, Green, Orange, Yellow), Physalis (normal, with Husk), Pineapple (normal, Mini), Pitahaya Red, Plum (different varieties), Pomegranate, Pomelo Sweetie, Potato (Red, Sweet, White), Quince, Rambutan, Raspberry, Redcurrant, Salak, Strawberry (normal, Wedge), Tamarillo, Tangelo, Tomato (different varieties, Maroon, Cherry Red, Yellow, not ripened, Heart), Walnut, Watermelon.**



The total number of images: 90483.

Training set size: 67692 images (one fruit or vegetable per image).

Test set size: 22688 images (one fruit or vegetable per image).

The number of classes: 131 (fruits and vegetables).

Image size: 100x100 pixels.

Datasets

The Training dataset has been split into train and test dataset with test size of 0.3.

Image dataset :

- The Training image dataset has 47384 data samples,
- The Testing image dataset has 20308 data samples,
- The Validation image dataset has 47384 data samples.

I. METHODOLOGY

OVERVIEW

There are various classification algorithms present out of which we shall implement the following

- *Random Forest Classification*
- *KNN*
- *Logistic Regression*
- *MLP*
- *Gaussian Naive Bayes*
- *SVM with linear Kernel*

We have used PCA for dimensionality reduction with number of components=120

Exploring the dataset and pre-processing

First of all we have imported all the images using the "glob" module to extract all the images into an array and then we extracted all the label names using the sub-folder names.

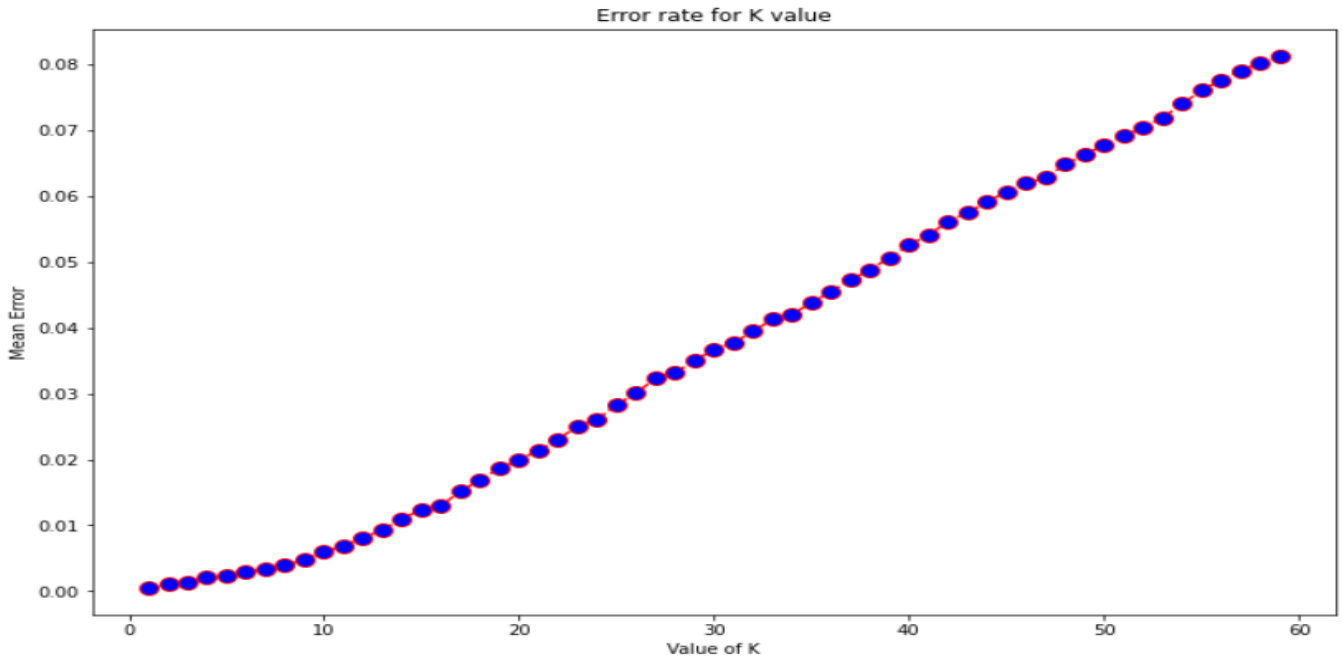
We extracted all the images in color format using the OpenCV module. In OpenCV, when reading a color image file, OpenCV imread() reads a image as a NumPy ndarray of row (height) x column (width) x color (3) and we have converted the order of color to BGR (blue, green, red).

Then finally resized all the images to size=45X45 and applied standard scaling to all the image dataset to make the computation Faster.

Implementation of classification algorithms

- ***Random Forest Classifier* :** Random Forest Classifiers use boosting ensemble methods to train upon various decision trees and produce aggregated results.It is one of the most used machine learning algorithms.
  We have used a Random Forest Deep tree classifier with a max depth of 48.

- **KNeighborsClassifier (KNN)***:* KNeighborsClassifier are supervised algorithms which classify on the basis of distance from similar points.Here k is the number of nearest neighbors to be considered in the majority voting process.
  We have used a Simple KNeighbor with number of neighbors=2 on the PCA reduced dataset.
  Here we have used the number of components=2 as it gives us the lowest error.



Looking at the graph we get the lowest error for n_neighbors=2

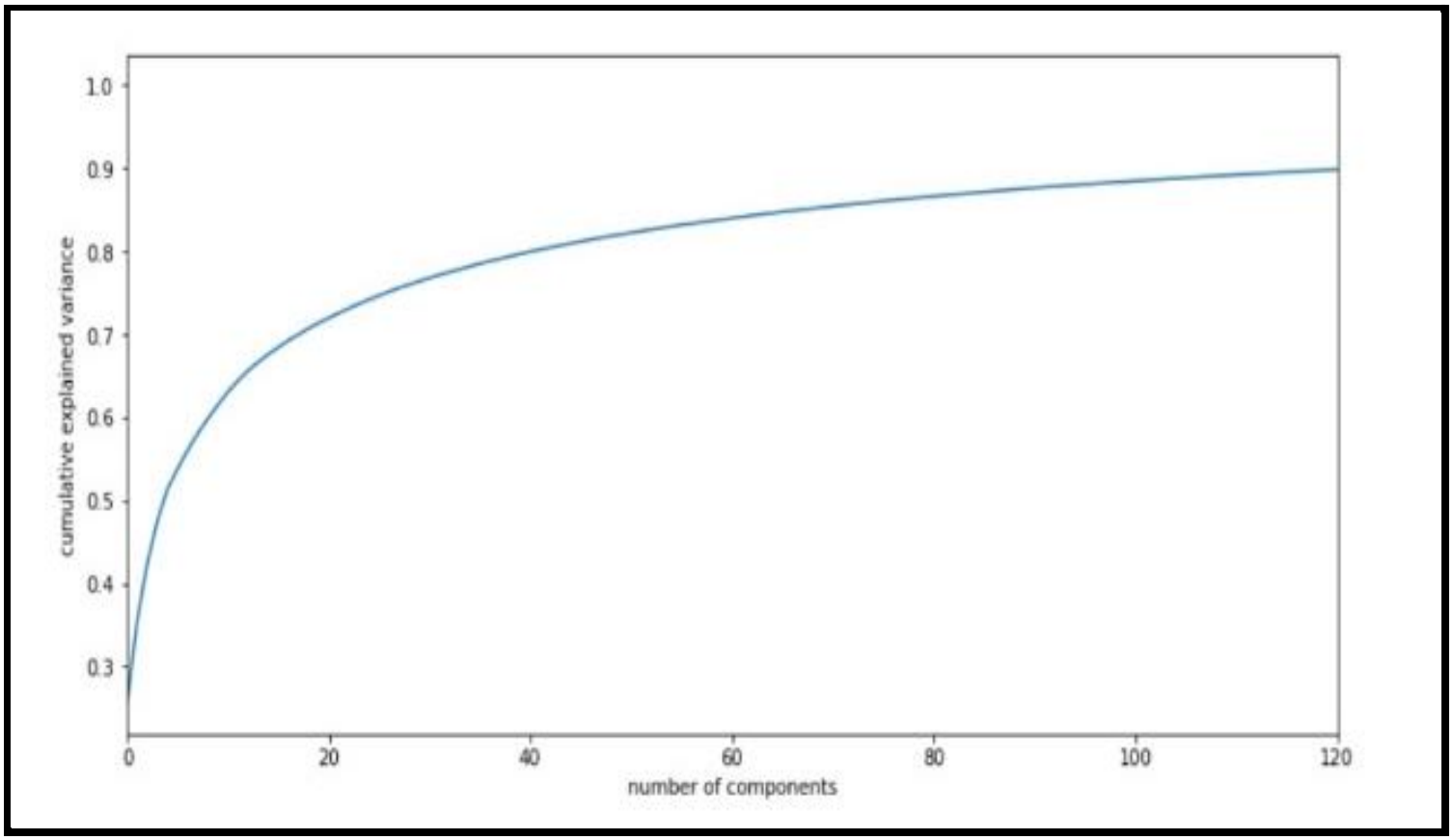
- ***Gaussian Naive Bayes*** *:*GNB is a type of Naive Bayes classifier that assumes that the distribution of data is gaussian and classifies data based on this assumption.
  We have used a Simple Gaussian Naive Bayes Classifier on PCA reduced dataset

- ***Logistic Regression*** :Logistic Regression model is widely used for binary as well as multiclass classification and hence is good suited for multiclass classification into various Fruits.
  We have used a Simple logistic regression on PCA reduced dataset

- ***Support Vector Machine*** :In SVM , data points are plotted into n-dimensional graphs which are then classified by drawing hyperplanes.
  We have used a Simple SVM with linear kernel on PCA reduced dataset

- ***Multilayer Perceptron*** :MLP is a feedforward Neural Network which uses backpropagation to update weights and improve the results.
  We have used a Simple MLP on PCA reduced dataset

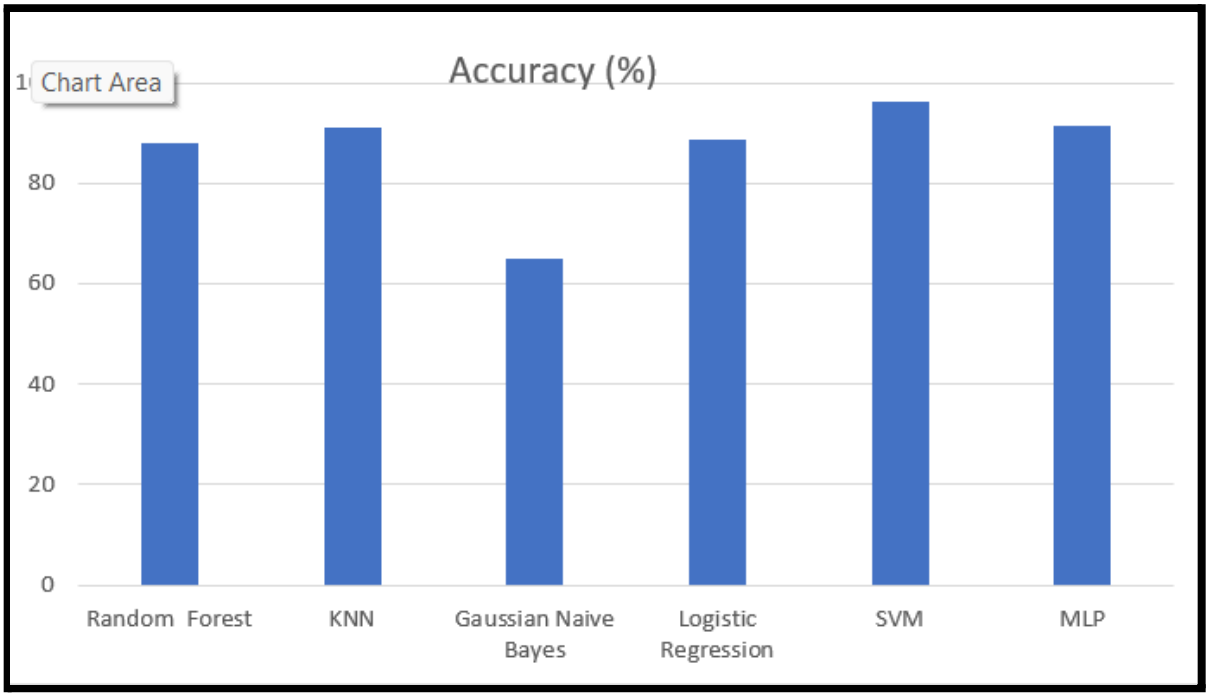
Dimensionality reduction techniques used **: PCA** with number of components=120 as 120 components is able to capture approx 90% variance.



II. EVALUATION OF MODELS

The models implemented were evaluated using techniques like - Classification report: precision, recall, f1 score and support, Confusion matrix, accuracy score.

Model	Accuracy (%)
Random Forest	88.0024
KNN	91.136283
Gaussian Naive Bayes	65.122532
Logistic Regression	88.51
SVM	96.055
MLP	91.3875



III. Results and Analysis

The table shows that the Gaussian Naive Bayes classifier had the worst efficient performance in all of the models. Logistic Regression, Multilayer Perceptron and Logistic Regression nearly have the same performance. SVM with linear kernel is the best classifier among all the models. Dimensionality reduction technique 'PCA' is used for speeding up classification.

CONTRIBUTIONS

The learning and planning was done as a team.The individual contributions are as given:

- Pradeep Kumar (B20AI029):Random Forest , Logistic Regression , Report.
- Pranay (B20AI030):KNN, Multi-Layer Perceptron , SVM ,Report.
- Shubham Kumar(B20AI039) : Data pre-processing and exploratory analysis ,Gaussian Naive Bayes, Report.

REFERENCES

[1] Support Vector Machine |www.javapoint.com|

[2] Gaussian Naive Bayes|machinelearningmastery.com

[3] Logistic Regression|Towards Data Science

[4] SVM|javapoint.com

[5] MLP|geeksforgeeks.com

[6] Random Forest | analyticsvidhya.com|Towards Data Science

Thank you !