Classification Of Fruits: ML Model

Manav Mittal
CSAI (2021-25)
IIIT-Delhi
India
manav21538@iiitd.ac.in

Utkarsh Venaik
CSAI (2021-25)
IIIT-Delhi
India
utkarsh21570@iiitd.ac.in

***Abstract*—In this report we present our approach to solve a classification problem. We used a combination of techniques like dimensional reduction methods, clustering, outlier detection and a combination of appropriate classification algorithms. The data-set we used in our project was shared to us via Kaggle. The method we used involves the use of Principal Component Analysis and Linear Discriminant Analysis for dimensionality reduction. Then, we used the unsupervised method K-Means for adding clustering labels as additional features to the data-set with the help of silhouette scores for deciding the K value, then we used Local Outlier Detection and a suitable threshold method to complete our preprocessing. We finally used ensemble techniques like Random Forests as well as combination of other classification algorithms like multi classifier logistic regression. We used K-fold cross validation for internal testing of our machine learning model. At the end we were able to achieve a rather high accuracy score in both the public and private testing data.**

***Index Terms*—PCA, LDA, clustering, k-means clustering, otsu thresholding , logistic regression, ensemble techniques, random forests, outlier detection, cross-validation.**

I. Introduction

In this report we discuss how we developed our machine learning model for classification. The model includes preprocessing steps like PCA, LDA, Cluster labelling , Silhouette Analysis and outlier detection while our machine learning model uses the logistic regression model along with the ensemble method of decision trees: Random forests. The predictions were done on a data-set provided to us by the university. Kindly go through the abbreviations and acronyms which will be used throughout the report from hereon instead of using their full forms.

A. Abbreviations and Acronyms

PCA - Principal Component Analysis LDA - Linear Discriminant Analysis ML - Machine Learning RI - Reinforcement LOF - Local Outlier Factor

B. Units

Accuracy is measured in percentage(%)

II. Dataset Description

The dataset is for Fruits Classification. It contains 20 classes and more than 1200 data samples. The feature vector of a data sample contains exactly 4096 features. It contains data to classify a particular fruit sample as Raw or Riped. The following bar graph provides the number of data samples in each class. The exact distribution class-wise is shown in [Table 1].

TABLE I
Classes and number of samples in them

Class	Number of samples
Apple Ripe	81
Apple Raw	78
Banana Ripe	86
Banana Raw	63
Coconut Ripe	58
Coconut Raw	54
Guava Ripe	56
Guava Raw	45
Leeche Ripe	55
Leeche Raw	37
Mango Ripe	46
Mango Raw	72
Orange Ripe	54
Orange Raw	48
Papaya Ripe	77
Papaya Raw	51
Pomengranate Ripe	57
Pomengranate Raw	58
Strawberry Ripe	75
Strawberry Raw	65

III. Data cleaning

We removed all the columns with all zero entries in them. This removed 473 features from the data-set. We also filled all the NA entries with the mean of the feature to which they belonged. After this, we were left with 3621 features for each data sample.

IV. Literature Review

In this section, we provide a brief overview of the concepts used in our model which are taken as secondary literature.

1) ***PCA****:* It is a dimension reduction method often used on data sets with large number of features, it aims at preserving the information of the large data set while reducing the dimensionality which makes it easier to explore and visualise. PCA is widely used in the fields of image processing, bioinformatics and finance.

2) ***LDA****:* A supervised algorithm aimed at finding a linear combination of features that best seperate the classes, it aims at minimizing the intra class variance while maximizing the inter class variance, used for better visualisation and exploration of

Identify applicable funding agency here. If none, delete this.

data. Widely used for preprocessing. Widely used in the fields of computer vision, genetics, speech recognition.

3) ***K-Means clustering****:* A popular clustering algorithm, aims at assigning labels to the data points based upon their distances from the centroids of clusters, has been used in our model for adding labels to the data-set, used for preprocessing in our data for improving the classification. The K value was decided using the Silhouette score.

4) ***Silhouette Score****:* Silhouette score is a metric used to calculate the effectiveness of the clustering technique, the more the metric is that values can be used as the K value for doing the clustering. The silhouette score value ranges from -1 to 1, the closer the silhouette score is to 1, it indicates that the current K value is more favourable. Other applications include image segmentation, market segmentation and customer segmentation.

5) ***LOF****:* LOF is an algorithm that is used to find the outliers present in the dataset based upon density based clustering an outlier considering the density of the neighborhood. It is particularly useful since there are certain outliers found which would not have been removed solemnly based upon distance based clustering. In our model it uses the Eucledian distance metric Other applications of LOF include: fraud detection , intrusion detection and network security.

6) ***Otsu Thresholding****:* A threshold teachnique to formally deecide a LOF score which divides the inliers weith outliers, in the data-set, In our model it uses the Eucledian distance metric It finds the optimal threshold to seperate a foreground and background pixels (inliers and outliers)

7) ***Logistic regression****:* A classification model that predicts a relationship between input and output variables. Since we need to classify between multiple labels we have used the One vs All method which assigns prediction probabilities to being of each label and then selects the one with the maximum probability as final label. Used in applications such as credit scoring and medical diagnosis.

8) ***Random Forest*** *:* An ensemble model of the decision tree algorithm to boost the accuracy of the predictions. In our project, Random Forests have been used as a secondary method which is used when the probability confidence of the logistic regression is low. Also, if the probabilities between the highest and next highest label are not much i.e: there is a high change that the other label may also be s potential solution using logistic regression we have used Random Forest to assist with the decision making. This algorithm has a wide use in the industry in fields such as bioinformatics , finance and image processing . It builds and ensemble of decision trees by selecting bootstrapped subsets to build different decision trees and combines their predictions to give the final predictions.

Our Model Applications of the classification of fruits can have multiple applications like in the farming industry, supermarket industries, etc. Farmers can use the model to determine whether their yield is ripe or raw which can help in increasing the efficiency of the harvesting process as a whole. Health and Nutrition industry can also benefit from the model since the nutrition content such as vitamins and minerals in the fruits vary with their degree of ripeness which can be predicted using our model. This classification model in general, much like other algorithms can be used in image classification using some modification.

V. Model Description

First, we added some of our custom features using unsupervised clustering. For this, we used the Kmeans algorithm. And to decide the number of clusters we should make we did the silhouette analysis. We added three features using three different values of k in Kmeans. We added features corresponding to k=2,8,18. Because we were getting, high silhouette scores for these values. We wanted a varied number of clusters; therefore, we chose 2,8,18. Silhouette scores for every value of k are shown in Fig. 1.

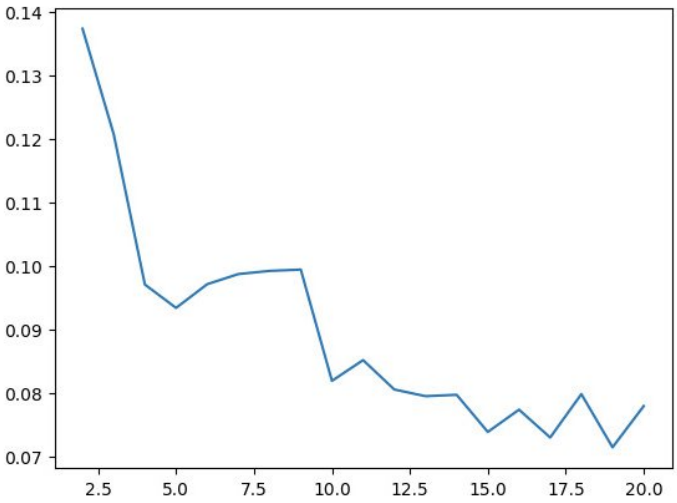

Fig. 1. Silhouette Score vs Number of Clusters before PCA-LDA

Then plotted the graph of the number of PCA components vs the explained variance. From this, we observed that only 661 components out of 3621 components account for 99 percent of the variance in the data. The plot is given in Fig. 2. After performing PCA now, we are left with around 661 features. After PCA to reduce the number of features, further we did LDA so that we could maximize the inter-class variance between the different classes and minimize the intra-class variance within the classes. And since we have already reduced the dimensions very much, we restricted ourselves from doing so more and kept the number of dimensions in LDA at its maximum limit which is 19 since there are 20 classes. Since too much reduction in the dimensionality can cause data loss. The graph of explained variance vs the number of LDA components is given below in Fig. 3. Though it says 18 components are enough just to be on the safe side we took 19 components. We then detected outliers using LOF and removed them so that data could be more refined and better predictions could be made. After this, we detected and removed outliers using Otsu thresholding on LOF scores. This removed about 32 outliers. After this, we

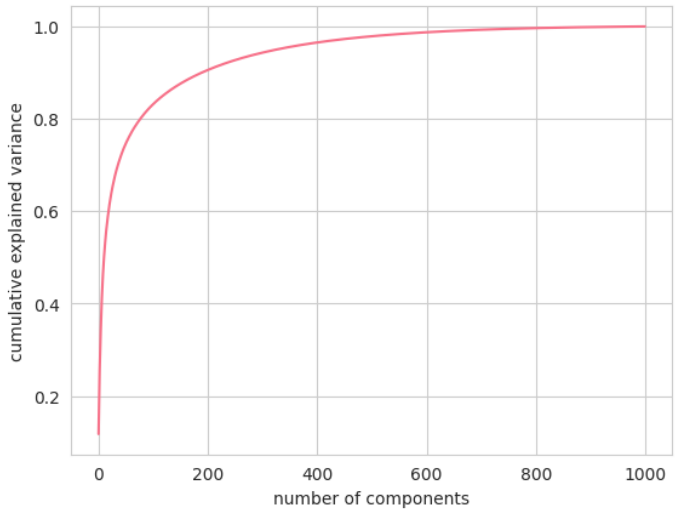

Fig. 2. Number of PCA components vs Explained variance

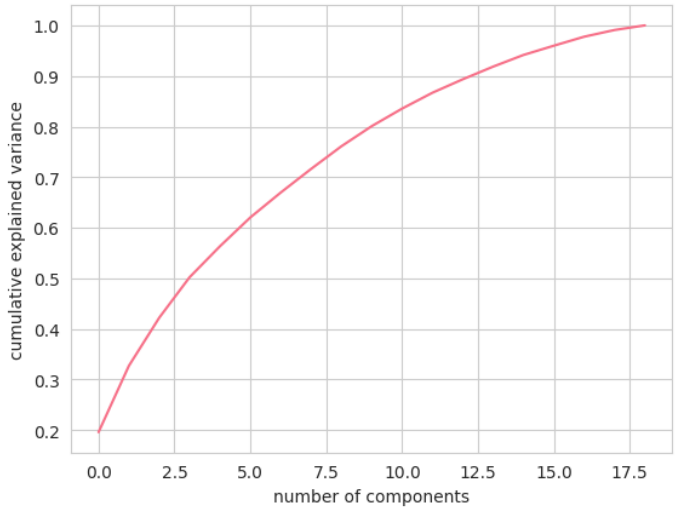

Fig. 3. Number of LDA components vs Explained variance

used Logistic Regression on the dataset to predict outputs. But after a detailed analysis of the probabilities predicted by logistic regression, we learned that there are some data samples for which Logistic Regression predicts classes with low probabilities. There we many cases in which we found out that the probabilities of belonging to different classes were very close. In fact, the difference was even less than 0.1. So to improvise upon it we use the Ensemble method Random Forest Classifier to replace such predictions with its prediction. We replaced all those labels predicted by Logistic Regressions, which were predicted by greater than just 0.2 probability of its successor. We chose 0.2 because we got the highest cross-validation accuracy of 82.3 percent on the training dataset.

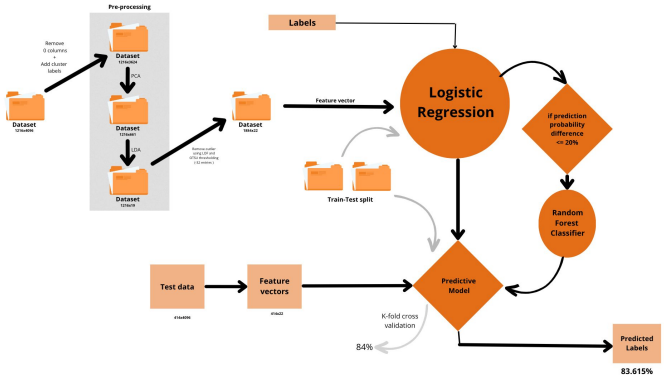

Fig. 4. Model Summary

VI. Model Figure

VII. Result

We got a record high of 84.057 percent on the first 50 percent of the test data and 83.173 percent on the second 50 percent. Making the average percentage 83.615 percent for the complete test dataset..

References

[1] https://towardsdatascience.com/

[2] https://scikit-learn.org/stable/supervised-learning.html#supervised-learning

[3] https://www.kaggle.com/competitions/sml-project/data

[4] Statistical Machine Learning Slides (Google Classroom) by Koteswar Rao Jerripothula

[5] https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

[6] https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[7] An Introduction to Principal Component Analysis (PCA) by J. Austin Pruszko (https://builtin.com/data-science/step-step-explanation-principal-component-analysis) (https://chrisalbon.com/machine_learning/fundamentals/linear_discriminant_analysis_fo

[8] Outlier Detection with Local Outlier Factor (LOF) by Jason Brownlee (https://machinelearningmastery.com/local-outlier-factor-for-outlier-detection/)

[9] Clustering Metrics: Silhouette Coefficient by S. Raschka (https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)

[10] Clustering with K-Means in Python by Kavita Ganesan (https://towardsdatascience.com/clustering-with-k-means-in-python-1e07a8bfb7ca)

[11] Random Forest: A Gentle Introduction to Random Forests by Jason Brownlee (https://machinelearningmastery.com/random-forest-ensemble-in-python/)

[12] Logistic Regression: Logistic Regression for Machine Learning by Jason Brownlee (https://machinelearningmastery.com/logistic-regression-for-machine-learning/)