CLASSIFICATION OF FRUITS

Statistical Machine Learning Project

Samridh Girdhar
samridh21282@iiitd.ac.in
Computer Science with Artificial intelligence
Indraprastha Institute of information and Technology, Delhi, India

Abhijeet Anand
abhijeet21526@iiitd.ac.in
Computer Science with Artificial intelligence
Indraprastha Institute of information and Technology, Delhi, India

Abstract- This project aims to develop a machine learning model for the classification of different fruits using a machine learning model. The report also discusses the various pre-processing steps applied to the dataset, the architecture of the ML model, and the evaluation metrics used to measure the performance of the model. The model was trained on the dataset using a various machine learning algorithms and achieved an accuracy of 82.211% on the complete test data. The findings of this project could have practical implications for automated fruit sorting and grading systems in the agriculture industry.

I. Introduction

Previous research has shown that our brain has specialized nerve cells responding to specific local features of a scene, such as lines, edges, angles or movement. Our visual cortex combines these scattered pieces of information into useful patterns. Our model aims to extract these meaningful pieces of information and put them together into a useful representation in order to perform a classification/ identification task on them.

In this project, the following various methods were used for the model:

i. **LOF:** *for removal of the outliers.*
ii. **K-means Clustering**: *form 20 clusters of fruits.*
iii. **PCA:** *reducing dimensions/features.*
iv. **LDA:** *final preprocessing.*
v. **Random Forest:** *ensemble classification method.*
vi. **K-Fold cross validation:** *for validating the trained model.*

We look at how these methods perform on our data, and how a machine learning model for fruit classification has the potential to improve efficiency, reduce costs, and increase productivity in various industries. By automating the sorting and grading of fruits, producers can save time and resources, while ensuring that the fruits meet the required quality standards.

II. Dataset & Pre-processing

The data used in this project is provided by the NeatAI servo-lab IIIT Delhi hosted on kaggle.com as a Contest. The dataset obtained contained 2 files: *train.csv* and *test.csv*, each containing *4096* features *(i.e., columns)*. The features were carefully selected to capture visual characteristics in the form of size, shape, color etc. in the form of numerical data. train.csv was used to train the model, and test.csv was used to evaluate the performance of the model in the contest.
The dataset was preprocessed to ensure that the data was clean and ready to use by the algorithm. The following schemes were used for our model:

A. Local Outlier Factor: removal of outliers

As the first stage of preprocessing the data, LOF (works by measuring the local density of a point in a dataset and comparing it to the densities of its neighbors to identify outliers) was used to reduce the influence of noisy data points and improve accuracy and reliability of the ML model, data points with high LOF score were removed considering them as an outlier. Schematically,

$$\text{Normalization} \longrightarrow \text{Detection}_{\text{10-neighbours}} \longrightarrow \text{DataFrames}$$

B. K-means Clustering: forming clusters as external features

To make the data more manageable for the model, grouping of similar data points was done into 20 clusters (since the model is a classifier for 20 fruits-kinds) using K-means clustering (works by partitioning a dataset into k clusters, where each cluster is represented by its centroid and each data point is assigned to the nearest centroid based on its distance) based on its simplicity, computational efficiency, and effectiveness in capturing underlying patterns in the data, and then used those cluster labels as external features in the dataset.

C. *Principal Component Analysis: reducing dimensionality*

For transforming the dataset into a lower-dimensionality space while retaining the important information, PCA was used. It is an unsupervised method of reducing dimensions. The dataset initially had 4096 features which were greatly reduced to 600. Where each new dimension was formed as a linear combination of the original features and ordered by their ability to explain variation in the dataset, making it further easier for the model to create accurate predictions.

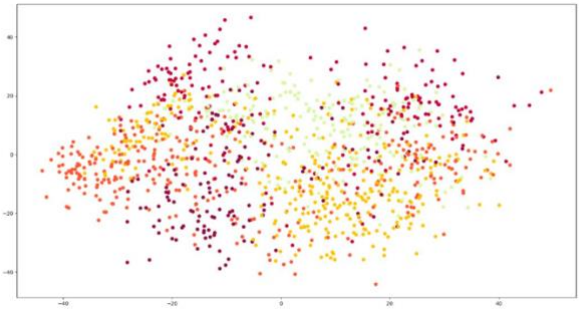

Figure 1: Dataset after operating PCA, with dimensions reduced to 600.

D. *Linear Discriminant Analysis: supplemented dimensionality reduction*

With dimensions now reduced to 600, we use a supervised learning algorithm LDA which finds a linear combination of features that maximises the separation between classes, to further reduce the dimensions into 19 features that best separates the different classes in the dataset.

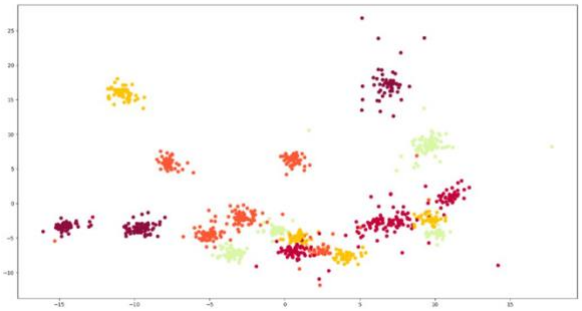

Figure 2: Dataset after PCA+LDA, with dimensions reduced to 19.

III. Classification

With the data gone through pre-processing, we then used *Random Forest Classifier (RFC)* for classification of the data, which is based on building multiple decision trees and combines their predictions to make a final prediction.

Because of its ability to handle high-dimensional datasets, being robust to noisy and missing data, and less prone to overfitting than other algorithms made it our best-bet for classification for generalizing a new, unseen data.

IV. Validation

To assess our model's performance, we use K-fold cross-validation technique. Which by splitting the dataset into k-subsets, trains the model on k-1 subsets and tests it on remaining subset, repeating it k times.

For our model, after experimenting with the hyperparameters, we chose k=5 and found that the model achieved an accuracy of 0.99 (nearly 100%) which suggests that the model is performing well on unseen data.

Acknowledgement

We would like to express our sincere gratitude to our proffessor Dr. Koteswar Rao Jerripothula of our course Statistical Machine Learning (CSE-342) at IIIT Delhi for organizing the kaggle contest and providing us with the opportunity to work on this challenge, I am also grateful to my seniors and colleagues for their guidance throughout the project.

References

[1] Lecture Slides
[2] https://scikit-learn.org/
[3] http://cs229.stanford.edu/
[4] https://www.geeksforgeeks.org/