Classification of Fruits

Pranav Aggarwal
2021551
pranav21551@iiitd.ac.in

Prerak Gupta
2021552
prerak21552@iiitd.ac.in

***Abstract*---We have developed and trained our model to classify a Kaggle dataset into categories of fruits. After using appropriate clustering algorithms, dimensionality reduction algorithms were applied on the dataset. We then removed the outliers in the dataset and used classifiers with ensemble methods. Finally, k-fold cross validation technique was used to validate the model. The objective is to predict the category of fruits from a test dataset.**

I. INTRODUCTION

We will discuss key concepts and methods used in classification. We will also look at data preprocessing steps that were followed. This report aims to provide a comprehensive understanding of machine learning classification techniques and their practical implications.

A. Dataset Description

Datasets Name: train.csv and test.csv

Data source : Kaggle

Data size : 4096 features, 1216 data points

Data format : CSV Format

Features : 4096 features (labelled n0 to n4095)

Label : Category of the fruits

Following are the categories present in the label column :- Apple_Raw, Apple_Ripe, Banana_Raw, Banana_Ripe, Coconut_Raw, Coconut_Ripe, Guava_Raw, Guava_Ripe, Leeche_Raw, Leeche_Ripe, Mango_Raw, Mango_Ripe, Orange_Raw, Orange_Ripe, Papaya_Raw, Papaya_Ripe, Pomengranate_Raw, Pomengranate_Ripe, Strawberry_Raw and Strawberry_Ripe.

II. DATA PREPROCESSING

The following preprocessing steps were applied on the data:

- After loading the dataset, we **removed the ID column** and stored the data and labels separately using slicing with the help of "iloc" function.
- **Feature Scaling** is applied on the dataset to **normalize** it for fair comparison.
- The labels were **encoded** from String values to Numerical values. There were 20 distinct labels so after encoding they ranged from **0 to 19**.
- For dimensionality reduction, **Principal Component Analysis** (PCA) followed by **Linear Discriminant Analysis** (LDA) was used. This is important because of the curse of dimensionality. It becomes increasingly difficult to analyze the data as the number of dimensions increase.
    1. PCA is a statistical method that transforms high dimensional data into lower dimensional space while retaining as much of the original variation in data as possible by performing a linear transformation on the data.
    2. LDA is a statistical method that transforms the dataset into a new coordinate system that maximizes the ratio of the between-class variance and the within-class variance resulting in well separated classes.
- **k-Means Clustering** is then applied to have cluster labels as additional features. This will group similar data points into clusters using distance of each data point to its nearest centroid and then updating each cluster till convergence. Silhouette Score was used to find the best value for k and it expectedly came out to be 20.
- Finally, the outliers were removed using the **Local Outlier Factor** (LOF). LOF measures the local density deviation of a data from its k-nearest neighbors to assign it an outlier score. We found that LOF gave 0 outliers.

III. CLASSIFICATION MODELS

The following classification models were used :

- **MLP Classifier** has been used as one of the voting classifiers for our ensemble method. This is a type of feedforward neural network classifier which used one or more hidden layers to predict the target variable. The hyperparameters (size of hidden layers, alpha, learning rate, max_iter, batch_size) were adjusted to minimize the error in the predicted variable and target variable using **GridSearchCV**. GridSearchCV gives the best combination of all possible sets of hyperparameters. We

have used two MLP Classifiers with activation functions logistic and relu respectively.

- **Logistic Regression** is a statistical method for classification tasks. It estimates the probability of a certain class based on the input features. Here the hyperparameters adjusted are max_iter, solver, penalty, and regularization parameter C. We have chosen penalty l2 because it is good for sparse dataset. We have used two logistic regression models with solvers newton-cg and liblinear respectively.

IV. ENSEMBLE METHOD

We used an ensemble method to combine our multiple individual models of MLP Classifier and Logistic Regression to improve the overall accuracy and stability of the final model.

For this we have used the Voting Classifier to combine the prediction of the four classifiers mentions in the previous section. By doing this, we are able to combine the strength of all our models. Hard voting was used because it is said to be better for sparse dataset and also gave better prediction for the dataset. Hard voting means final prediction is based on majority vote of the individual models.

Voting classifiers are particulary effective in reducing bias and improving generalization performance.

V. APPLICATIONS

- **Principal Component Analysis**

PCA is used for data compression. It is used to recognize patterns and also in image processing. It is also used to analyze genetic data to identify genetic markers associated with certain diseases. It is used in finance to analyze factors that contribute to performance of stocks and other financial instruments.

- **Linear Discriminant Analysis**

LDA is used in computer vision applications to classify images based on their features. It is also used in speech recognition. It is used in marketing to classify customers based on their demographic and behavioral data and similarly to classify patients based on their symptoms and medical history. Just like PCA, it is also used in finance.

- **k-Means Clustering**

k-Means is used in image segmentation into distinct regions. It is also used for customer segmentation and recommendation system to group similar products together. K-Means can also be used to identify outliers or anomalies in a dataset.

- **Local Outlier Factor**

LOF is used for anomaly detection. It can be used to detect fraudulent behavior in financial transactions. It is also used to identify anomalies in medical data and to detect abnormal traffic patterns.

- **Neural Network**

Neural networks are used in image recognition and natural language processing. It is also used in recommender systems and financial forecasting. One of the main applications of neural networks are autonomous systems such as drones and robots.

- **Logistic Regression**

Logistic Regression is used in medical diagnosis to predict the likelihood of a patient having a particular disease. It is also used in credit scoring and fraud detection. It is useful for sports analytics and sentiment analysis.

References:

[1] https://scikit-learn.org/stable/user_guide.html

[2] https://medium.com/machine-learning-researcher/dimensionality-reduction-pca-and-lda-6be91734f567

[3] https://vitalflux.com/k-fold-cross-validation-python-example/

[4] https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a#:~:text=Grid%20search%20is%20a%20tuning,us%20time%2C%20effort%20and%20resources.

[5] https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e

[6] https://towardsdatascience.com/anomaly-detection-with-local-outlier-factor-lof-d91e41df10f2

[7] https://medium.com/edureka/what-is-a-neural-network-56ae7338b92d

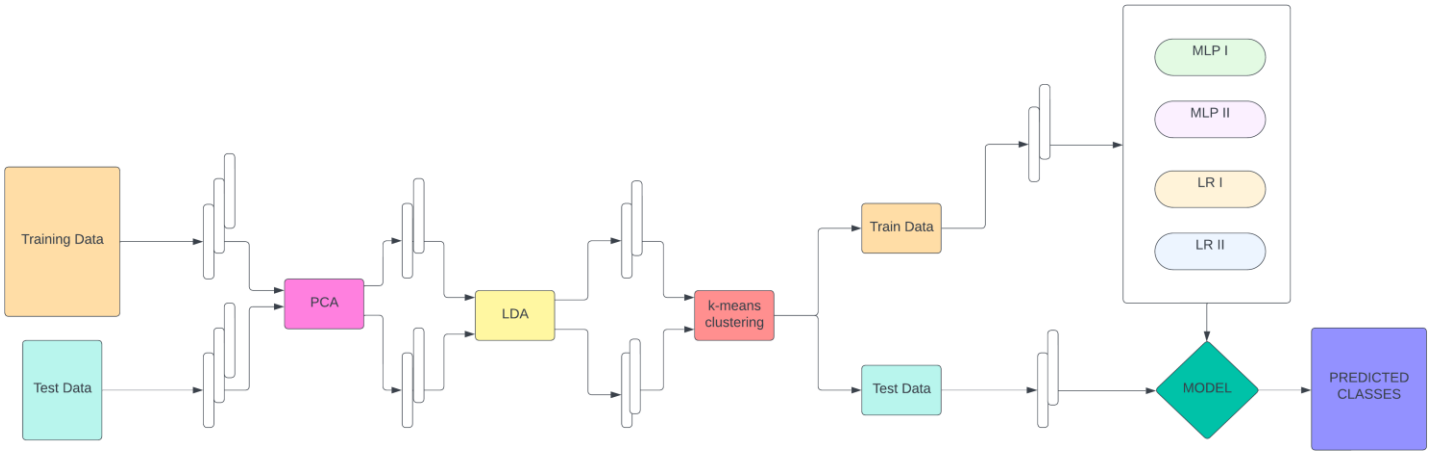

FIG1 Summary of the Project