SML Project

Team Name : Vortex

1st Maanas Gaur
Indraprastha Institute of Information
Technology Delhi (IIIT-Delhi)
Delhi, India
maanas21537@iiitd.ac.in

2nd Lakshya Kumar
Indraprastha Institute of Information
Technology Delhi (IIIT-Delhi)
Delhi, India
lakshya21536@iiitd.ac.in

***Abstract*—This paper presents a machine learning approach for fruit classification using the Kaggle competition dataset.The goal of this project is to predict the category of fruits in the test set based on the given features. Our approach includes preprocessing steps along with appropriate classification algorithms and ensemble methods. The literature review also provides a brief description of different applications of these methods in similar classification tasks. The proposed approach achieved a high accuracy of 98.77 % on the validation set and 84.06 % and 80.77 % on the first and second halves of the test set, respectively.**

***Index Terms*—Ensemble, Multi-layer Perceptron classifier, Principal Component Analysis, Linear discriminant analysis**

I. Introduction

Fruit classification is an important problem in the food industry and agriculture.In this project, we aim to classify fruits based on the given features using machine learning algorithms.

II. Dataset Description

A. Train and Test Dataset

The dataset provided for the fruit classification problem consists of two files: train.csv and test.csv. The train.csv file contains 4098 columns, with the first and last columns providing the 'ID's and target values ('category'), respectively. The remaining columns (n0 to n4095) contain the features used for classification. The target variable 'category' has 20 distinct classes, namely
'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', 'Strawberry_Ripe'.

The test.csv file contains unlabeled data with the same structure as the train.csv file, with the exception of the missing 'category' column.

III. Methodology

Preprocessing is an essential step in any machine learning task, as it helps in improving the quality of the data and enhances the performance of the model. However, the preprocessing steps required for each dataset may vary based on the nature of the data and its quality. Therefore, preprocessing requires numerous iterations to identify the optimal preprocessing steps that can improve the accuracy of the model. In this project, we performed multiple iterations of preprocessing to identify the best techniques to be used for the given fruit classification problem

A. Preproccessing Steps

Removal of ID: The ID column in the given dataset does not contribute to the preprocessing or the classification task. Therefore, we removed it from the dataset to avoid any unnecessary computation and to simplify the data.

Several preprocessing techniques were explored in this project, including Local Outlier Factor (LOF), k-means clustering, Principal Component Analysis (PCA), and Linear Discriminant Analysis (LDA).

LOF was initially considered to detect and remove any outliers in the dataset. However, it was found to be less useful in increasing the accuracy of the model due to its poor performance in validation accuracy.

Similarly, k-means clustering was also considered to group similar data points together, but it did not perform well on this dataset, leading to its exclusion from the final approach.

On the other hand, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) were found to be useful in reducing the dimensionality of the dataset and increasing the accuracy of the model. PCA is a technique used to reduce the number of dimensions in a dataset while retaining the maximum amount of information.It creates new features that are linear combinations of the original features, and these new features capture the most significant variation in the data.

LDA is also a technique used for feature extraction that is commonly used in classification problems. It transforms the dataset into a new coordinate system that maximizes the separation between classes while minimizing the variance within each class. In other words, it creates new features that

'https://scikit-learn.org/stable/'

can discriminate between different classes and help improve the classification accuracy.

In this project, both PCA and LDA were applied to reduce the dimensionality of the dataset and to identify the most significant features that contribute to the classification task. The performance of the models improved significantly after applying PCA and LDA, which suggests that they were successful in reducing the noise in the data and identifying the most informative features.

B. *Classification Models*

The MLPClassifier is a multi-layer perceptron algorithm that is commonly used for classification tasks. It uses a neural network with one or more hidden layers to predict the target variable. The hyperparameters that can be adjusted include the number of hidden layers, the number of nodes in each layer, the activation function, the optimizer, and the learning rate. In this project, we used two MLPClassifiers with different activation functions, namely logistic and tanh.

Logistic regression is a linear model used for classification tasks. It estimates the probability of a sample belonging to each class and then predicts the class with the highest probability. The hyperparameters that can be adjusted include the regularization parameter, penalty, solver, and maximum number of iterations. In this project, we used two logistic regression models with the same regularization parameter and penalty but with different solvers, namely newton-cg and lbfgs.

C. *Ensemble Method*

After exploring various classification algorithms and preprocessing techniques, we used an ensemble method to improve the performance of our model. Specifically, we used the Voting Classifier from scikit-learn library to combine the predictions from four classifiers - nn1, nn2, lr1, and lr2 - with a hard voting strategy.

The voting classifier takes a list of estimators and combines their predictions by taking the majority vote. In our case, the four classifiers nn1, nn2, lr1, and lr2 were included in the estimator list. The hard voting strategy means that the final prediction is based on the mode of the predicted classes by each individual classifier.

Using the voting classifier helped to improve the accuracy of our model by reducing the variance and bias errors. It also provided a more robust prediction, as it combined the strengths of each individual classifier. Moreover, it allowed us to use different types of classifiers, each with its own strengths, to create a more accurate and balanced final prediction.

Overall, the ensemble method of the Voting Classifier was an effective way to combine the predictions of multiple classifiers, and helped us achieve a high accuracy on the test set.

D. *Figures*

Figure illustrates the algorithmic flow for PCA, LDA, classification, ensemble, and anticipated output of test data following training on train data.

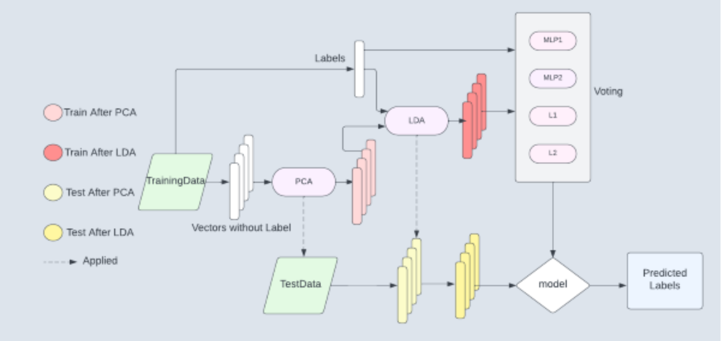

Fig. 1. Pipeline of Algorithm.

IV. Applications

A. *PCA (Principal Component Analysis)*

Some of the uses of PCA include Image Compression, which compresses digital images by lowering their dimensionality while maintaining the most significant information, and used in Facial Recognition too for the same reason. It is also utilised in Neuroscience for spike-triggered covariance analysis. It is also used in Quantitative Finance to assist discover the most essential elements that drive asset returns, minimising risk and maximising profit.

B. *LDA (Linear Discriminant Analysis)*

It has many applications like Identification, i.e., to identify the type of customers that is likely to buy a certain product in a store. Used to classify gene expression patterns based on their relevance to a particular phenotype or condition. It is also used in Pattern recognition and in finance to identify the most important factors that drive investment returns.

C. *Neural Network (MLP Classifier)*

It has numerous applications in various fields. Speech Recognition, to classify different words and sounds. It is also used to detect fraudulent transactions or activities in finance and banking. They are used in medical diagnosis to classify patients into different categories based on their symptoms and medical history. They are used in natural language processing to classify text into different categories.

D. *Logistic Regression*

Logistic Regression has various applications, such as Credit Scoring, which shows great results in it as it helps lenders to assess the creditworthiness of borrowers and make informed lending decisions. It is a popular choice in Natural Language Processing tasks. Logistic regression can be used to diagnose various diseases such as cancer, diabetes, and heart disease.

References

[1] https://scikit-learn.org/stable/
[2] https://activewizards.com/blog/5-real-world-examples-of-logistic-regression-application
[3] https://people.revoledu.com/kardi/tutorial/LDA/Applications.html
[4] https://iq.opengenus.org/applications-of-pca/
[5] https://www.analyticssteps.com/blogs/8-applications-neural-networks