Fruits Classification

Statistical Machine Learning Project

Ankit Gautam
Computer Science with Artificial Intelligence
Indraprastha Institute of Information Technology Delhi
Delhi, India
ankit21518@iiitd.ac.in

Kartikey Dhaka
Computer Science with Artificial Intelligence
Indraprastha Institute of Information Technology Delhi
Delhi, India
kartikey21534@iiitd.ac.in

***Abstract*—Machine learning is becoming increasingly popular and can be applied in many different areas of our lives. In this project, we have applied and experimented with different statistical machine-learning techniques to classify fruits. The results of this study can be used to classify fruits based on their features and predict whether the fruit is ripe or raw. Industry experts can use this project for various purposes. This project can be used in computer vision, and then The evaluation results for our machine learning model show the best accuracy with a score of 0.85024 in the public leader board and 0.81250 in the private leader board.**

I. Introduction

In the agro-industrial field, machine learning can be used in fruit sorting machines. Besides that, it can be used to scan fruits in supermarkets. Fruit identification and classification is one of the most used tasks on farms and supermarkets. On farms, fruit identification and classification can increase product packaging efficiency. Meanwhile, these tasks can increase the efficiency of arranging fruit on supermarket shelves. Thus, the identification and classification of fruit are necessary. This data collection will help develop future fruit classification solutions in these localities.

II. Dataset pre-processing

We were provided with two dataset files, i.e., "train.csv" and "test.csv." train.csv contained 4097 columns, of which 4096 were the features, and the last column consisted of class labels. On the other hand, test.csv contained only 4096 columns. We were required to train our model using the features and class labels present in train.csv and then generate a new result.csv containing the output values for our test.csv. While developing a machine learning model, the first step is pre-processing the given data. To pre-process the given data we used the following preprocessing techniques.

A. Cluster labels as external features

We used Knn-classification on our train dataset to generate cluster labels and then used those cluster labels as external features in our dataset. As we already know, clustering the data is a good practice that helps enhance the accuracy of our machine-learning model and helps reduce noise from the dataset.

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

B. Reducing dimension of the data

We used a combination of principal component analysis (PCA) and linear discriminant analysis (LDA) on the data set to reduce the number of features and increase class separation. We kept the value of n components as 450. PCA is a statistical algorithm that reduces the number of features in a data set. LDA is also a statistical algorithm. LDA finds a linear combination of features that maximally separates the classes in a data set. Applying both PCA and LDA increased the accuracy of our model.

C. Removing outliers

We used the Local Outlier Factor (LOF) to remove outliers from our dataset. Outliers are the data points that add noise to the resultant model and reduce the overall accuracy. Outlier detection also has many real-life applications, such as fraud detection. Services like Instagram use outlier detection to detect suspicious login attempts.

III. Classification

We have used Logistic Regression (LR) and Random Forest Classifier (RFC) as the classifier for our dataset. We have chosen 10,000 as the number of iterations for the logistic regression and a max depth of 100 for the random forest classifier. We were asked to use ensemble methods in our model. Therefore, we used Voting Classifier for LR and RFC models to combine their predictions and generate an ensemble prediction.

IV. Validation

Finally, we used the K-fold cross-validation technique to validate our model. For the application of validation of our model, we chose five as the hyperparameter value (k). Cross-validation is a procedure used to evaluate machine learning models on a dataset. The algorithm calculates accuracy in a group of "k" data points on the entire dataset and then calculates the mean of all the accuracies. After this model was complete, we got a mean accuracy of 0.99. This is close to 1. Therefore, we can say that our model is working correctly and providing outstanding results.

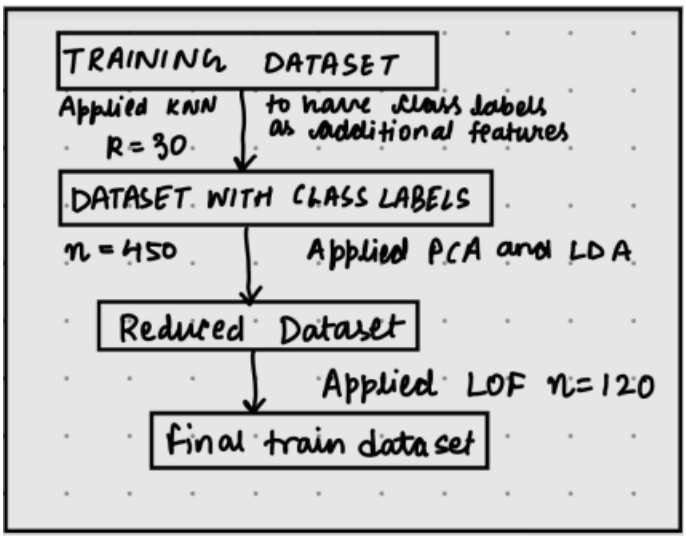

Fig. 1. Pre-processing of training data-set.

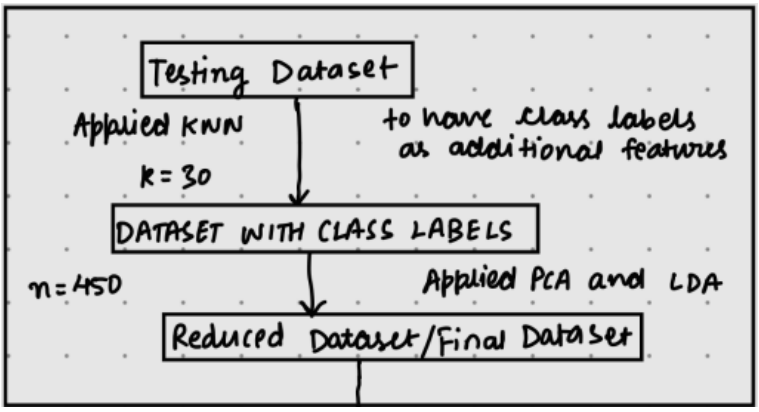

Fig. 2. Pre-processing of testing data-set.

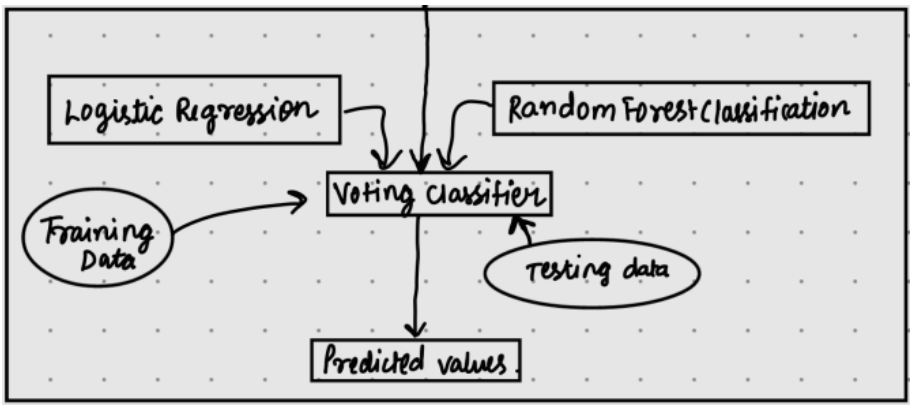

Fig. 3. Combining training data, testing data, and the classification model.

ACKNOWLEDGMENT

We want to thank Dr. Koteswar Rao Jerripothula, who provided the dataset for training and testing our model and was our instructor for the Statistical Machine Learning course at the Indraprastha Institute of Information Technology Delhi during the Winter semester of 2023.

REFERENCES

[1] Lecture Slides.
[2] https://stackabuse.com/implementing-lda-in-python-with-scikit-learn/
[3] https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/
[4] https://www.geeksforgeeks.org/ensemble-methods-in-python/