Classification of Fruits

Shubham Pal
CSAI
IIIT D
NEW DELHI, INDIA
shubham21564@iiitd.ac.in

Parth Kaushal
CSAI
IIIT D
NEW DELHI, INDIA
parth21548@iiitd.ac.in

***Abstract*—This report focuses on the classification of fruits using machine-learning techniques. The dataset used has 4096 features and 1216 data points, with a target feature consisting of 20 classes based on the ripeness and type of fruit. Outlier detection using the LOF, Dimensionality reduction using PCA and LDA techniques, and K-means clustering was used for grouping similar data points, which also improved the accuracy. The model was trained using Logistic Regression and Random Forest algorithms, ensembled using a voting classifier. The final accuracy obtained was around 81.25**

I. DATASET OVERVIEW

The dataset has 4096 features(columns) representing various parameters upon which the data was made and 1216 data points (rows), i.e., the observed values corresponding to those parameters. The target feature consists of the ripeness and type of fruit; in total 20 classes are possible. We first converted the labels from string to integer so that they could be used directly in the algorithms present in the libraries of Python.

II. OUTLIER DETECTION AND REMOVAL

Outliers are data points whose values differ significantly from other points in the dataset. These data points should be removed from the dataset to train the model accurately. LOF is an outlier detection algorithm that was used. The Local Outlier Factor (LOF) is a statistical method for identifying outliers in a given dataset. LOF measures the local deviation of a data point with respect to its nearest neighbors. The accuracy was tested without outlier removal, and the accuracy obtained was around 79.8 percent, however with the hyperparameters of LOF($nneighbors = 40, contamination = 0.01, metric = "euclidean"$), around 13 outliers were detected, which when removed from the dataset results in around 81.25 percent .

III. DIMENSIONALITY REDUCTION

Dimensionality reduction is a technique used to reduce the number of features or variables in a dataset. It is used to simplify the data while preserving as much information as possible. Working with lower dimensional data is easier and much more efficient. The algorithms which help us in achieving this are PCA and LDA.

A. Principal Component Analysis (PCA)

It is an unsupervised technique used to transform high-dimensional data to a specified number of columns, called principal components, while retaining as much of the original data's variation as possible. The essential features in the data are preserved. By reducing the number of features used in the model, which can reduce overfitting, the accuracy can be improved using PCA. It also works as a preprocessing step and helps build a stable model. For the dataset, the accuracy recorded without using PCA was 39.4 percent, and with PCA (components = 359) around 81 percent. Since PCA was beneficial for overall accuracy, it was used.

B. LDA (Linear Discriminant Analysis)

It is a supervised dimensionality reduction technique and is a dimensionality reduction technique used in machine learning. It aims to find a linear combination of features that maximally separates the data into different classes or categories. It helps reduce unnecessary data while preserving relevant information. The accuracy obtained without using LDA was around 73.5 percent, and using LDA with the parameters (n cluster=19) was around 81 percent.

.

The k-fold cross-validation technique with K=10 was used for both the above techniques to check for the accuracy obtained, and the beneficial algorithm was used.

C. K-means

It is used for partitioning a given set of data points into k clusters. It tries to minimize the sum of the squared distances between data points and their assigned centers of the cluster. It is used to group data points similar to each other into clusters. It can help in identifying the patterns in the data. By grouping similar data points together, K-means can reduce the dimensionality of the dataset and simplify the task of further training the model. The optimal value of the K-fold was found using Silhouette Analysis. Without using K-Means, the accuracy obtained was 79 percent with the values.

IV. MODEL TRAINING

K fold cross validation is used for the evaluation of the overall performance of the model. It involves partitioning the original dataset into K equally sized subsets (or ”folds”) and then using 1 of the folds one by one for training and all the other folds for testing. This process is repeated for each fold one by one. The average of all accuracies obtained is returned at the end. The algorithms used for training are Logistic Regression and Random Forest

A. *Logistic regression*

It is a statistical method for analyzing a dataset in which one or more independent variables determine an outcome. The outcome is measured with a variable (generally used when only two possible outcomes exist but can be further expanded for multiple classes). Here, multi-class classifier logistic regression is used to predict the labels of the testing data based on the features. Logistics gives a good performance on linearly separable data. The hyperparameters used for Logistic Regression were the ones that performed the best in the k-fold cross-validation technique. The parameters, $C = 0.01, dual = False, fit_intercept = True, max_iter = 500, penalty =' l2', solver =' newton - cg'$ were the best. The accuracy obtained was around 80 percent.

B. *Random Forest*

Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset Using it in conjunction with logistic regression improved the accuracy to up to 82 percent.

C. *Ensembling*

Ensembling is an ML technique that combines multiple models to improve the overall prediction accuracy. By combining multiple models, we can reduce the errors a single model can make. The voting classifier combines the prediction of different logistic regression models and predicts which it finds to be the most accurate and recurring in the algorithms. The accuracy obtained without using ensembling and just using Logistic Regression was giving an accuracy of around 80 percent, and on using ensembling, the accuracy increased to 81 percent.

V. Result

The final results are as follows:

- Validation Acc: 84%
- Public Score: 84%
- Private Score: 81.25%

Using all the above steps and knowledge, we can predict the type of fruit and its ripeness.This model can be helpful in various fields of work, like in industries working with fruits or a similar situation as to which batch of products to use in manufacturing and processing according to the requirement of that particular fruit, like raw fruits to be used or ripened fruits.

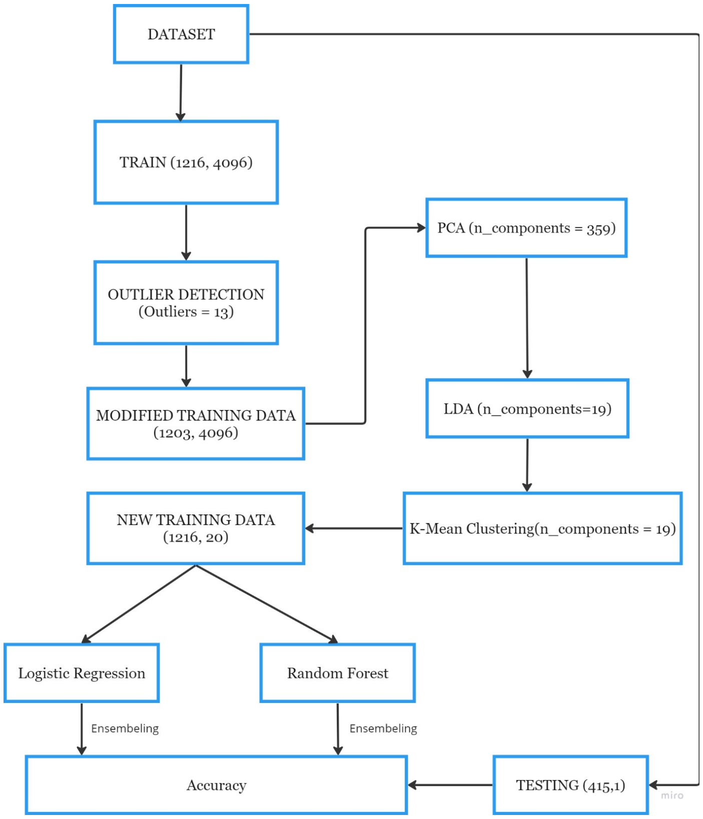

Fig. 1. Workflow

VI. Acknowledgement

We would like to express my sincere gratitude to Prof. Koteswar Rao Jerripothula for his invaluable guidance and support throughout our academic project. His insightful feedback, encouragement, and dedication to teaching have inspired us to reach new heights in our studies. Thank you, Prof Koteswar Rao Jerripothula for your unwavering commitment to your students and your contributions to the academic community.

VII. References

- Algorithms - https://scikit-learn.org/stable/
- Dataset - https://www.kaggle.com/competitions/sml-project/data
- PCA - https://builtin.com/data-science/step-step-explanation-principal-component-analysis