SML Project (Fruit Classification)

1st Akash Kushwaha
CSAI (2021-2025)
IIIT Delhi
New Delhi, India
akash21514@iiitd.ac.in

2nd Vivek jain
CSAI (2021-2025)
IIIT DELHI
New Delhi, India
vivek21218@iiitd.ac.in

***Abstract*—This report represents an approach to solve a fruit classification problem by applying several machine learning algorithms, given data set contains 1216 data points and each data point contains 4096 features, First of all we applied Kmeans clustering algorithm to add the generated labels as an additional feature to the data and then reduced the number of features using dimensionality reduction algorithms like PCA and LDA, we also used Local Outlier factors to remove the outliers from the data set, Lastly we used logistic regression as a classification algorithm to predict the final labels.**

I. Introduction

Given data set contains 20 classes of fruits needed to be classified. The objective of this project is to develop a machine learning model that can accurately classify different types of fruits based on their features. The dataset used in this project contains 1216 datapoints, each with 4096 features. The goal is to achieve high accuracy in fruit classification, which can be used in various real-world applications. In this report, we will provide a detailed description of our methodology, including the machine learning algorithms and techniques used to build our fruit classification model. We will also present our results and discuss the insights gained from our analysis.

II. Dataset Description

Given data set contains 1216 data points and each data point contains 4096 features and there are total 20 classes of fruits which we want to classify. In order to train and evaluate our fruit classification model, the dataset was split into a training set and a testing set. The training set contains 80 percent of the datapoints, while the testing set contains the remaining 20 percent. This split was chosen to ensure that the model is trained on a sufficient amount of data while still having a separate dataset to evaluate its performance. Given able provides a more detailed description of given data set.

III. Methodology

We used the following algorithms to prepare our model well:-

Clustering Algorithm:- Kmeans

It is a popular unsupervised clustering algorithm used in machine learning and data analysis. The goal of the algorithm is to partition a dataset into k clusters, where each data point belongs to the cluster with the nearest mean. The algorithm starts by randomly selecting k initial cluster centers from the data points. It then assigns each data point to the nearest cluster center based on the Euclidean distance between the data point and the cluster center. After assigning all data points to clusters, the algorithm recalculates the mean of each cluster and moves the cluster center to the new mean. This process is repeated until the cluster assignments no longer change or until a maximum number of iterations is reached.

TABLE I
Fruit ripeness percentages

Fruit	Ripeness percentage
Banana_Ripe	86
Apple_Ripe	81
Apple_Raw	78
Papaya_Ripe	77
Strawberry_Ripe	75
Mango_Raw	72
Strawberry_Raw	65
Banana_Raw	63
Coconut_Ripe	58
Pomengranate_Raw	58
Pomengranate_Ripe	57
Guava_Ripe	56
Leeche_Ripe	55
Coconut_Raw	54
Orange_Ripe	54
Papaya_Raw	51
Orange_Raw	48
Mango_Ripe	46
Guava_Raw	45
Leeche_Raw	37

Outlier removal:- Local Outlier Factor

It is an unsupervised anomaly detection algorithm that identifies outliers by measuring the density of data points in their local neighborhood. The algorithm calculates the local reachability density (LRD) for each data point based on its distance to its k-nearest neighbors, where k is a user-defined parameter. Data points with a significantly lower density than their neighbors are identified as outliers based on their LOF score, which is the ratio of the LRD of the data point to the LRD of its k-nearest neighbors. Data points with high LOF scores are considered to be outliers, as they have significantly lower density than their neighbors and are thus less representative of the underlying data distribution.

Dimensionality reduction methods:-

Principal Component Analysis (PCA):
It is a popular dimensionality reduction technique that identifies the most important features in a dataset and transforms the data into a new coordinate system. The algorithm calculates the principal components of the data, which are linear combinations of the original features that capture the maximum variance in the data.

Linear Discriminant Analysis (LDA):
It is is a supervised dimensionality reduction technique that aims to find a lower-dimensional space that maximizes the separation between classes in the data. The algorithm identifies the most discriminative features by calculating the between-class and within-class scatter matrices of the data. LDA projects the data onto a new space that preserves the class separation.

Classification Algorithms:- Logistic regression
It is a popular classification algorithm that models the probability of a binary or categorical outcome based on one or more predictor variables. The algorithm estimates the coefficients of the predictor variables using a maximum likelihood estimation method and uses these coefficients to predict the probability of the outcome for new data.

IV. Literature review

The methods used in this study, including K-means clustering, LOF outlier detection, PCA, LDA, and logistic regression, have been widely applied in various fields.

- K-means clustering has been used in image segmentation, customer segmentation, and network traffic analysis.
- LOF outlier detection has been applied in credit card fraud detection, intrusion detection, and medical diagnosis.
- PCA has been used in face recognition, gene expression analysis, and speech processing.
- LDA has been applied in face recognition, text classification, and bioinformatics.
- Logistic regression has been applied in churn prediction, spam filtering, and sentiment analysis.

These methods have demonstrated their effectiveness in solving various real-world problems and are commonly used in data science and machine learning applications.

V. Results and Observations

In this study, we developed a sophisticated machine learning model to classify different category of fruits based on the given features, After loading the data, we split it into training and testing sets, with a 80-20 split ratio. We then trained several classification models, including k-th Nearest Neighbour (kNN), Random Forrest classifier and Logistic Regression and LOF to remove Outliers.
Logistic Regression was found to be the best classification model, outperforming kNN and Random Forest classifiers, irrespective of whether we implemented dimensionality reduction or not. This finding suggests that Logistic Regression can be a reliable and robust tool for fruit classification, especially in situations where the number of features is high.
kNN scored 0.657, Random Forrest scored a maximum of 0.80 and Logistic regression alone itself scored a maximum of 0.82.
Moreover, our application of various clustering algorithms, such as K-means, DBSCAN, Agglomerative, and Spectral clustering, highlights the potential of clustering techniques in augmenting generated labels as new features for enhanced fruit classification. Interestingly, K-means outperformed other clustering algorithms, indicating that this algorithm can be an optimal choice for clustering-based fruit classification. Then, we removed other clustering algorithms and applied K-means 3 times because k-means is a randomized algorithm, meaning that the initial cluster centroids are randomly selected, and different initializations can result in different final cluster assignments. By running the algorithm multiple times with different initializations and selecting the solution with the lowest overall error, one can increase the likelihood of finding a more stable and accurate clustering solution, thus we obtained more better results.
Furthermore, our use of Dimensionality Reduction techniques, such as PCA and LDA, elucidates their potential for enhancing fruit classification performance. After applying K-means and Logistic Regression with reduced dimensions, we achieved a commendable public score of 0.85024. This result indicates that Dimensionality Reduction techniques can be instrumental in effectively reducing the number of features while retaining important information for fruit classification, thus improving performance.
We performed an extensive hyperparameter tuning exercise to identify the optimal values for various hyperparameters, including the number of clusters in K-means, epsilon value in DBSCAN, number of neighbors in kNN, maximum iterations in Logistic Regression, and the number of features in PCA and LDA.
Our results indicate that hyperparameter tuning can significantly improve the performance of the classification models. For instance, the maximum public score of 0.85024 and maximum private score of 0.85096 we achieved was a result of extensive hyperparameter tuning.

In summary, our study successfully developed a machine learning model for fruit classification, with Logistic Regression emerging as the most effective classification model. The application of clustering algorithms and dimensionality reduction techniques further improved performance. Our observations highlight the importance of hyperparameter tuning in machine learning and the need for more efficient techniques for optimization.