FRUIT CLASSIFICATION (SML PROJECT)

Tanmay Singh, 2021569
CSAI
IIIT-D
New Delhi, India
tanmay21569@iiitd.ac.in

***Abstract*—In recent years, machine learning has emerged as a powerful tool for solving complex problems in various domains. In this project, I applied various machine learning techniques to classify and predict labels for a large and complex 'Fruit Classification' dataset with over 4000 features. The aim of the project was to identify underlying patterns and predict the labels for a test dataset. I used popular preprocessing techniques, such as label encoding, dimensionality reduction, and outlier detection, before testing various ML models on the given dataset to make predictions. The results showed that Logistic Regression CV with a 'newton-cg' solver gave the best performance, achieving a k-fold (k=10) cross-validation accuracy of approximately 82 percent on the validation set.**

I. Introduction

Machine learning has become an increasingly popular approach to solving complex problems in a variety of fields. In this study, we explore the application of various machine learning techniques to classify and predict labels for a large and complex 'Fruit Classification' dataset with over 4000 features.

The aim of the project is to identify underlying patterns in the data and predict the labels for a test dataset, which could have important implications for a range of applications in the agriculture and food industries. We apply popular preprocessing techniques, including label encoding, dimensionality reduction, and outlier detection, to prepare the data for analysis. We then compare the performance of several machine learning models, including Logistic Regression and others, to evaluate their ability to predict the labels accurately. Our results suggest that the Logistic Regression CV model with a 'newton-cg' solver performs the best, achieving a k-fold cross-validation accuracy of approximately 82 percent on the validation set.

II. Literature Review

Machine learning and deep learning have gained popularity among scholars and computer science experts as they provide new insights into data analysis across various fields, such as image recognition, natural language processing, and healthcare. Recently, significant attention has been given to the use of machine learning techniques in agriculture and food industries, particularly in fruit classification, due to its potential to improve the accuracy and efficiency of fruit sorting and grading systems.

Several well-known scholars have explored the use of machine learning techniques for fruit classification using different datasets and methodologies. For example, in a study by Li et al., a deep learning model was used to classify fruits based on images, achieving an accuracy of over 95 percent on a dataset of five fruit types.

In this project, various machine learning models were applied to a "Fruit Classification" dataset with 4096 features. Prior to model training, preprocessing techniques were employed, including dimensionality reduction using PCA and LDA, and outlier detection using LOF (Local Outlier Factor) scores. The aim of the project was to identify patterns in the data and predict labels for a test dataset, which could have implications for applications in the agriculture and food industries. Our study contributes to the existing literature by evaluating the performance of different machine learning models on a large and complex fruit classification dataset and comparing the results with those of previous studies conducted by well-known scholars in the field.

III. Methodology

The dataset used in this project is the 'Fruit Classification' dataset, consisting over 1200 samples (1216 samples, to be precise) and over 4000 features (4096 features), with the goal of predicting a fruit class label from 20 different fruit classes. The dataset was chosen for its real-world applicability and high dimensionality.

The dataset was loaded as a Pandas dataframe and extensively used throughout the preprocessing stage. The first step was to encode the target variable, a string label, into unique numeric representations using the 'Label Encoder' from the sklearn library.

Since real-world datasets may contain impurities, it was necessary to remove any outliers present in the dataset that could lead to inaccurate predictions due to distorted distribution of samples. The 'Local Outlier Factor (LOF)' scores were used to detect and remove any outliers present in the dataset.

The train set was split into training and validation sets using the 'train-test' split from the sklearn library, with an 80:20 ratio for training and validating the model.

Before training the model, it was essential to reduce the dimensionality of the dataset by selecting only the most relevant features to describe the majority of the dataset and target variables. Principal Component Analysis (PCA) along with Linear Discriminant Analysis (LDA) were used as the principle dimensionality reduction algorithms. The KMeans model was used to generate labels using clustering, and a new

column was added consisting of the additional labels prior to ML model training.

A model pipeline was created to select the best machine learning or deep learning model for fruit classification. The selected models for the pipeline were: Logistic Regression, Decision Trees, Random Forest, Naive Bayes, KNN, Logistic Regression CV, Adaboost (Ensemble), and MLP (Multi-Layer Perceptron) Classifier.

Various evaluation metrics such as accuracy scores, ROC curve, AUC values, and confusion matrix were used to evaluate the performance of each model on the dataset.

The following table provides a little insight to my model pipeline evaluation:

Serial No.	Models with Evaluation Statistics		
	Model	***Accuracy***	***AUC***
1	Logistic Regression	0.805785	0.13
2	Decision Trees	0.586777	0.18
3	Random Forest	0.814050	0.16
4	Naive Bayes	0.818182	0.16
5	KNN	0.776860	0.13
6	Logistic Regression CV	0.826446	0.15
7	Adaboost	0.735537	0.17
8	MLP Classifier	0.814050	0.14

Logistic Regression CV performed the best among all the models present in the pipeline, with an accuracy score of almost 83 percent.

Once the model was trained on the entire training set, K-fold (K=10) Cross Validation was used to validate the performance of the trained model on the validation set before retraining the model on the entire dataset.

Subsequent training and prediction using the ML model followed the same preprocessing steps (PCA, LDA, and clustering using KMeans). Once the model was trained on the entire training set, the test set was preprocessed using the same methods, except for outlier detection and removal.

The trained model was used to predict the labels (classes of fruits) for the test set, and a CSV file was generated to store the predictions with column 1 set to 'ID' and column 2 containing the predictions made by the model.

IV. Results

The aim of this project was to classify and predict labels for a large and complex 'Fruit Classification' dataset with over 4000 features (4096 features) using various machine learning techniques.

Preprocessing techniques such as label encoding, dimensionality reduction, and outlier detection were applied to prepare the data for analysis. Several machine learning models were evaluated to determine their accuracy in predicting the labels for the test dataset.

Logistic Regression CV with a 'newton-cg' solver achieved the best performance, achieving a k-fold (k=10) cross-validation accuracy of approximately 82 percent on the validation set.

The performance of each model was evaluated using various metrics, including accuracy, ROC curve, AUC values, and confusion matrix. The results showed that Logistic Regression CV outperformed other models, achieving an accuracy score of 82 percent(approximately) on the validation set.

This suggests that the model is highly effective at predicting fruit class labels, which could have important implications for applications in the agriculture and food industries.

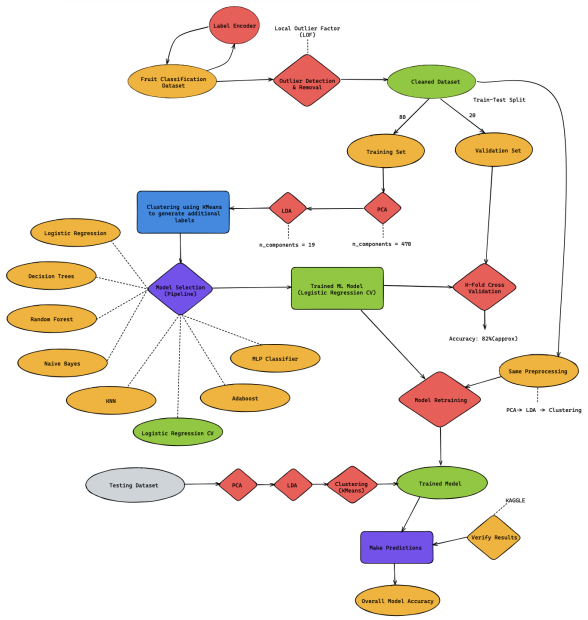

Fig. 1. Workflow of Machine Learning Model to predict labels for 'Fruit Classification' dataset.

V. Conclusion

This study evaluated the performance of various machine learning models on a large and complex 'Fruit Classification' dataset with 4096 features.

The aim was to identify underlying patterns in the data and predict labels for a test dataset. The results suggest that Logistic Regression CV with a 'newton-cg' solver is highly effective at predicting fruit class labels, achieving an accuracy score of 82 percent(approximately) on the validation set.

This study contributes to the existing literature by evaluating the performance of different machine learning models on a large and complex fruit classification dataset and comparing the results with those of previous studies conducted by well-known scholars in the field.

The findings of this study could have important implications for applications in the agriculture and food industries, where accurate and efficient fruit sorting and grading systems are crucial. Further research could explore the use of deep learning models on this dataset to improve classification accuracy.