# SML Project

Aryan Dhull
*2021520*
*IIIT-Delhi*
Delhi, India
aryan21520@iiitd.ac.in

Deepanshu
*2021524*
*IIIT-Delhi*
Delhi, India
deepanshu21524@iiitd.ac.in

**Abstract**

*In this project, we used clustering, dimensionality reduction, and outlier detection algorithms to preprocess a dataset and improve the performance of classification algorithms. We applied ensemble methods to the classification algorithms and used k-fold cross-validation to validate our models. Our results show that our approach improved the accuracy of our classification models and contributed to the field by demonstrating the effectiveness of these techniques.*

***Keywords*— k-fold cross-validation**

## I. INTRODUCTION

Machine learning algorithms have become increasingly important crucial for various applications, including image recognition, speech recognition, and natural language processing. However, these algorithms often require large amounts of data and significant preprocessing to achieve high accuracy. In this project, we aimed to improve the performance of classification algorithms by using clustering, dimensionality reduction, and outlier detection algorithms to preprocess our data. We then applied ensemble methods to the classification algorithms to improve accuracy. We used the k-fold cross-validation technique to validate and demonstrate our models' effectiveness. By demonstrating the effectiveness of these techniques, we contribute to the field of machine learning and help improve the performance of these algorithms for real-world application.

## II. DATASET DETAILS

The dataset used in this project contains 1216 rows of data for fruit classification, with each row corresponding to a single fruit sample. The dataset has 4096 different features numbered from n0 to n4095. The dataset consists of 10 different fruits, including strawberry, leeche, orange, apple, pomegranate, banana, coconut, guava, papaya, and mango. Each fruit was further divided into raw and ripe, thus making a total of 20 categories. Each category contains a different number of samples. This dataset presents a challenging classification problem due to the high number of features and the similarity between specific fruit categories.

## III. PREPROCESSING THE DATA

Before applying any algorithm to the given dataset, we divided the data into X and Y, where X contains all the features, and Y contains the category.

## IV. METHODOLOGY AND MODEL DETAILS

### *A. Principal Component Analysis*

PCA is a statistical technique that reduces the dimensionality of large data sets while retaining important information. The technique aims to identify the underlying structure of the data by finding new variables known as principal components. The features are selected based on the variance they cause in the output. The feature that causes the highest variance is the first principal component. The principal components are uncorrelated to each other.

PCA has various applications in data preprocessing, feature extraction, data visualization, and exploratory data analysis, and it is used in finance, marketing, biology, and computer vision.

- The dataset after preprocessing had 4096 features. Since the number of features was very high, it was required to reduce its dimensionality.
- We used PCA to reduce the dimensionality of the dataset to 400 features, uniquely identified by numbering them as 0 to 399.

### *B. Linear Discriminant Analysis*

LDA is a statistical machine learning technique used to identify the underlying relationship between a set of independent features (predictors) and a categorical dependent variable (response or target variable). It is a supervised machine learning algorithm that is commonly used for classification problems.

The algorithm calculates the means and covariance matric for each class, and uses these values to compute a set of coefficients that can be used to predict the class of a new observation based on its predictor values.

LDA has several applications in fields such as biology, medicine, finance, and engineering. It is often used for tasks such as credit risk assessment, disease diagnosis, and image recognition.

- The dataset after applying principal component analysis had 400 features. Since the number of features was 400 and the number of categories were 20. It was required to reduce the number of features.
- We used LDA to reduce the dimensionality of the dataset to 17 features, uniquely identified by numbering them as 0 to 16.

### C. K- Means Clustering

K-means clustering is a popular unsupervised machine learning algorithm used for grouping similar data points together in a dataset. The algorithm works by partitioning a ser of observations into a predefined number of clusters (k), where each observation belongs to the cluster with the nearest mean.

The algorithm begins by randomly selecting k points from the dataset as the initial centroids of the clusters. The remaining observations are then assigned to the closest centroid based on their distance to it. After all the observations have been assigned, the centroids are updated by computing the mean of all the observations in each cluster. This process of assigning observations and updating centroids is repeated until the centroids no longer change or a maximum number if iterations is reached.

It is used in image segmentation, customer segmentation, anomaly detection, and bioinformatics. It is useful for identifying patterns and relationships in large, complex data.

- We used K-Means Clustering on the dataset after applying PCA and LDA, dividing the dataset into 14 clusters.
- We added a column cluster to the dataset which denotes the cluster number to which that data belongs to.

### D. Grid Search CV

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. As the performance of a model significantly depends on the value of hyperparameters.

- We used GridSearchCV to find the best parameters to use in our model in order to get higher accuracy.

### E. Random Forest

Random Forest is a popular ensemble machine learning algorithm used for classification and regression tasks. It works by creating multiple decision trees on different subsets of the dataset and then combining their outputs to make a final prediction.

- We found out that the accuracy decreased after applying random forest, hence we removed it.

### F. Logistic Regression

Logistic Regression is a popular machine learning algorithm used for binary classification tasks. It models the probability of a binary outcome by using a logistic function to transform the linear regression output into a probability score between 0 and 1. The algorithm works by estimating the coefficients of the independent variables that best predict the binary outcome.

- We used logistic regression on the dataset after applying PCA, LDA and K-means clustering using the parameters we got from GridSearchCV algorithm.
- We stored the predict values we obtained from logistic regression for testing.

### G. K-Fold Cross Validation

K-fold cross-validation is a technique used to assess the performance of a machine learning model. It involves dividing the dataset into K equal parts for testing the model's performance. The goal is to evaluate the model's ability to generalize well to new, unseen data.

It is useful for preventing overfitting and providing a more accurate estimate of a model's performance.

- We used K-fold Cross-Validation on the dataset after applying logistic regression using K=15.

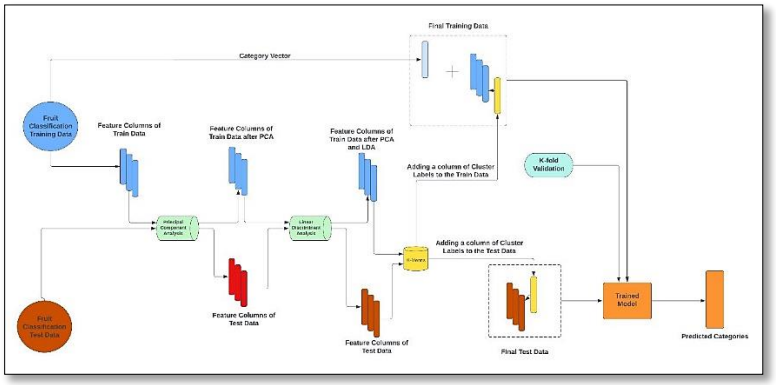


Fig. 1. Pipeline of Algorithm

## V. REFERENCES


[1] Christopher M. Bishop - Pattern Recognition and Machine Learning-Springer (2006)

[2] Sapir, Marina. (2022). Pragmatic Theory of Machine Learning. 10.13140/RG.2.2.26068.07043.

[3] Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd edition, ASIN: B07VBLX2WL