# Multi-class classification problem on fruits dataset

Vishesh, Divya raj Singh
*B.Tech in Computer Science and Engineering, B.Tech in Computer Science and Artificial Intelligence*
*Indraprastha Institute of Information Technology*
Delhi, India
vishesh21114@iiitd.ac.in, divya21528@iiitd.ac.in

## Abstract

We try to solve the multi-class classification problem on a fruits dataset by preprocessing the dataset and extracting important features from the dataset with techniques such as LDA and PCA. After preprocessing, we fit an appropriate algorithm and record its performance in terms of accuracy evaluation metrics.

**Keywords**: Classification, LDA, PCA, Validation, Logistic regression

## 1. Introduction

In machine learning, classification is a predictive modeling problem where the class label is anticipated for a specific example of input data. A classification problem is used to identify specific categories of new observations based on one or more independent variables. It is a supervised machine learning problem. The classification problem in this case was to train a ML algorithm which can precisely classify a new data point based on the fruits dataset which it is was previously trained on.

## 2. Preprocessing steps

The project starts with importing essential libraries in python such as numpy, pandas, etc. and loading the dataset in form of a pandas dataframe which can be used later. The dataframe consists of independent variables X and target variable y. Both needs to be extracted from the dataframe in the next step. Upon analysis we find our dataset consists of 1216 rows and 4098 columns. Since the number of features are much larger than the number of samples we need to apply dimensionality reduction to our feature vector X before applying any ML algorithm.

For dimensionality reduction, we first proceed with PCA to get 100 principal components and then perform LDA to get 19 features which transforms our dataset into 1216 rows and 19 columns.

Since, clustering id of a sample can be useful information while trying to classifying a point, we perform K-means on X to get cluster ids of samples.

In the next step, we do the 80-20 train-test split to train our model on training set and validating our model on other 20% of data.

We then apply StandardScaler function to our X vector to scale our features vector. This is done to normalize the range of independent variables of data, because some distance-based algorithms such as gradient descent are heavily impacted by range of features.

## 3. Fitting appropriate ML algorithm

After preprocessing step our dataset is ready, so that we can start training variety of algorithms on our dataset. We train a lot of algorithms such as Logisitic Regression, SVM, LDA, Random Forest, Decision Tree, Bagging classifier with Logistic Regression and SVM, Voting Classifier with Logisitc Regression and SVM, Neural Networks, Boosting algorithms such as ADA Boost and XGBoost. We used GridSearchCV for hyperparameter tuning for above mentioned algorithms.
After training our model we calculated accuracy scores on both training and validation set to finalize the algorithm which works best with our dataset. Logistic Regression gives the best score on validation set and does not overfit on training set. So we proceed with Logistic Regresion for prediciting values on the test set.

## 4. Prediction on test set

We then import our test set and perform the same preprocessing steps as we did on training and validation set. Then we predict the labels for each sample using predict function of the Logistic Regression model we trained above.

## 5. Exporting the predictions

We then convert the prediction labels from numpy array into a pandas dataframe and then export a “submission.csv” file to a local folder.

## References

Aized Amin Soofi (2017), *Classification Techniques in Machine Learning: Applications and Issues*

Javier Fernandez (2022), *Learn to use Bagging in your models*

Sckit-learn.org - LDA, PCA, StandardScaler