SML Project Report

Lakshya Agrawal
Indraprastha Institute of Information Technology
Delhi
Team Name on Kaggle: Lakshya
Email: lakshya21535@iiitd.ac.in

***Abstract*— The goal of this project is to develop a machine learning model that can accurately classify data samples from a given dataset. Using machine learning algorithms such as clustering, dimensionality reduction, outlier detection, and classification algorithms the model is built. The dataset used for this project was given by the instructor.**

Keywords— Machine learning; kMeans; Outlier detection; Clustering; Classification; Logistic regression; Dimensionality Reduction; Validation; Ensemble methods.

I. INTRODUCTION

Fruit classification helps to identify and sort different fruits based on their ripeness. It is a very important step for many industries especially the agricultural sector. Accurate fruit classification has a lot of direct benefits for the farmers as well. It can help farmers in the harvesting stage of fruits, where the farmer can get to know if the fruit is ripe or raw by its many features such as colour, size, weight, etc.
In this project, we aim to classify fruits as either ripe or raw as well identify the fruit using a dataset containing over 4000 features. The task is to develop a classification model that can accurately predict the ripeness of the fruits based on these features. The next section briefs upon the methodology adapted for the project.

II. METHODOLOGY

The first step is loading the dataset. The dataset was obtained from Kaggle, a dataset competition platform where the instructor had uploaded the dataset. First the dataset was uploaded to the google drive and then the google drive was mounted on google colab to access the dataset. The dataset used in this project contained various features related to the data instances. The dataset was divided into two parts: the training set and the test set. The training set contained the target variable, while the test set did not.

Characteristic of the dataset – The shape of the training set: 1216 rows × 4098 columns. It includes the ID as well as the target variable of the data. Test dataset has the shape:415 rows × 4097 columns. The next step is to extract the category column of the training dataset and assign it as y_train. Then the following steps were performed to train the model.

Dimensionality Reduction: Dimensionality reduction is the process of reducing the number of features in a dataset while retaining as much information as possible. In this project, feature selection method named Variance Thresholding. Here, those features were removed which had a variance lesser than a particular threshold(0.01). Here, the X_train is fit and transformed by the variance thresholding object created which is followed by removing those corresponding features in the testing dataset which were removed from the training dataset. It is very helpful in removing noise from the dataset and utilize those features which would help in better classification.

It is followed by applying PCA(Principal Component Analysis) to reduce the number of features to 400. It helps in increasing the efficiency of the model and helps represent the data in lesser features effectively.

Clustering: Clustering is a technique used to group similar data points together. In this project, KMeans clustering algorithm is used to cluster the data. The number of clusters(here) is set to 25. Cluster labels are created for both training data as well as the testing data. The cluster labels are then added as additional features to both datasets correspondingly. This is done to improve the performance of the classification model by providing it with more information about the data.

Standardization of data: The fit_transform() method is used to standardize the training data, X_train, by computing the mean and standard deviation of each feature in X_train and then scaling the features based on these statistics.

Outlier Detection: Outlier detection is the process of identifying and removing data points that are significantly different from the rest of the data. In this project, Isolation Forest algorithm is used for outlier detection. Isolation forest object is made which is fitted on the training data followed by predicting the outliers, and then finally removing the outliers from the training dataset. Isolation Forest is a tree-based algorithm that isolates outliers by recursively partitioning the data into smaller subsets. The contamination parameter is set to 0.1 to remove 10% of the data points that are identified as outliers.

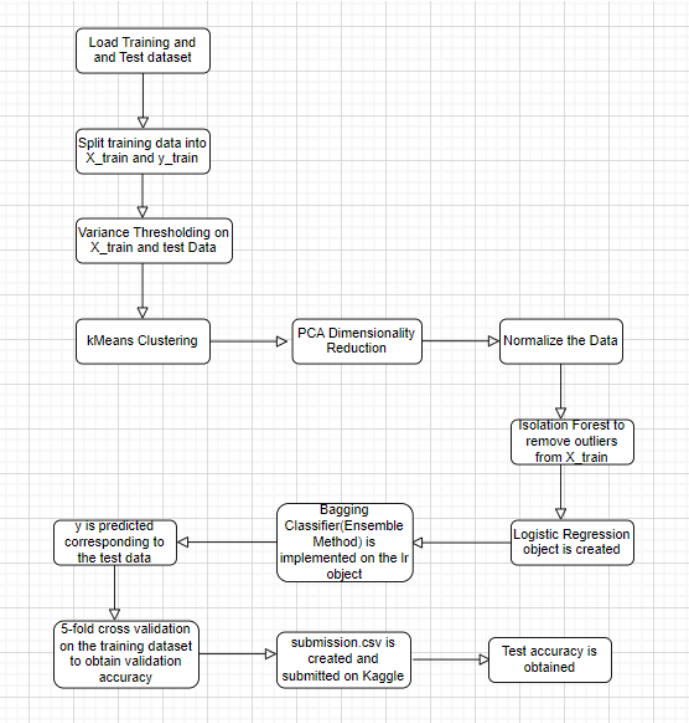
Classification: In this project, Logistic Regression algorithm is used for classification. Logistic Regression is a linear model that predicts the probability of a data point belonging to a particular class. The random_state parameter is set to 42, which ensures that the results are reproducible. The LogisticRegression object is then created and stored in the lr variable. The model is trained on the preprocessed data via ensemble methods and finally cross-validation is used to estimate the accuracy of the model.

Ensemble Method: Ensemble methods are techniques that combine multiple models to improve the performance of the classification model. In this project, Bagging Classifier is used as an ensemble method. The base_estimator parameter is set to lr, which means that logistic regression will be used as the base estimator for the bagging classifier. The BaggingClassifier object is then trained on the training data (X_train and y_train) using the fit() method. Bagging Classifier is an estimator that fits multiple base classifiers on different subsets of the dataset and then aggregates their predictions. Bagging Classifier is used to reduce the variance of the Logistic Regression model and improve its accuracy.

K-fold Cross-Validation:

K-fold Cross-Validation is a technique used to evaluate the performance of a classification model by dividing the data into k equally sized subsets. The model is trained on k-1 subsets and tested on the remaining subset. This process is repeated k times, with each subset serving as the test set once. In this project, 5-fold cross-validation is used to estimate the accuracy of the Logistic Regression model. This means that the training data is divided into 5 equal parts, and the bagging classifier is trained and evaluated 5 times, each time using a different fold as the validation set.

Finally, the bagging classifier is used to predict the target variable for the test data, and the results are saved to a CSV file called submission.csv .



The overall procedure has been shown in the following figure:

II. LITERATURE REVIEW

In the recent years, the usage of machine learning techniques have gradually increased especially in fields such as healthcare, finance, agriculture and image processing. The techniques widely used includes clustering, dimensionality reduction, outlier detection and classification.

Clustering algorithms are used in various applications such as image clustering,agriculture,etc. It is used for tasks such as crop yield prediction and disease detection. For example, a study used clustering algorithms to perform the task of customer segmentation[1].After analysis of data and classifying customers with features annual income and spending score, the author got clusters of customers.

Outlier detection techniques are used in various applications such as fraud detection[2] and medical diagnosis. In the domain of agriculture, outlier detection techniques have been used for various tasks such as detecting diseased leaves in tomato plants.

Logistic regression is used in various applications such as credit scoring, medical diagnosis. In the case of medical diagnosis, logistic regression can be used to predict the likelihood of a patient having a certain disease based on their symptoms, medical history, and other relevant factors.

Ensemble methods such as random forests and gradient boosting is used in many applications, such as credit scoring, medical diagnosis, and image classification.

Dimensionality reduction techniques is used in tasks including image and signal processing, finance, and healthcare. In the domain of agriculture, dimensionality reduction techniques is mainly used for food quality inspection.

I. RESULTS

We applied a combination of machine learning techniques to the task of classifying fruits as either ripe or raw while also identifying the fruit type.

After applying all the steps stated in Methodology, an accuracy of 74.6 was achieved in the cross validation step. The submission.csv file was created which contained the predicted labels of the test data. Upon submission to the platform, the accuracy of the model on the test data came out to be 80.67%.

II. Conclusion

In conclusion, in this project we have shown the use of several machine learning algorithms for the task of fruit classification. We showed that clustering algorithms can be used to separate ripe and raw fruits, and outlier detection techniques can be used to identify samples that do not fit into these clusters. We also showed that logistic regression and ensemble methods can be used to classify the fruit types.
Our results suggest that these techniques can be applied in real-world scenarios for fruit classification and ripeness inspection.

REFERENCES

[1]N. Patankar,S.Dixit,A.Darpel, “Customer Segmentation Using Machine Learning”: Recent Trends in Intensive Computing

[2]Samaneh Sorournejad, Zojah, Atani, “A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective”, 2016