Predicting Fruit Freshness Using Machine Learning Methods

Aditya Girdhar
Dept. of Computer Science and Engineering
IIIT-Delhi
New Delhi, India
aditya21005@iiitd.ac.in

Priyash Shah
Dept. of Computer Science and Engineering
IIIT-Delhi
New Delhi, India
priyash21553@iiitd.ac.in

I. Introduction

This course project proposes a machine learning approach to classify the freshness of fruits based on a variety of features. We were given a data-set of a variety of fruits with over 4098 features. We then experimented with employing a range of machine learning algorithms, including decision trees and regression, to train and test our models.

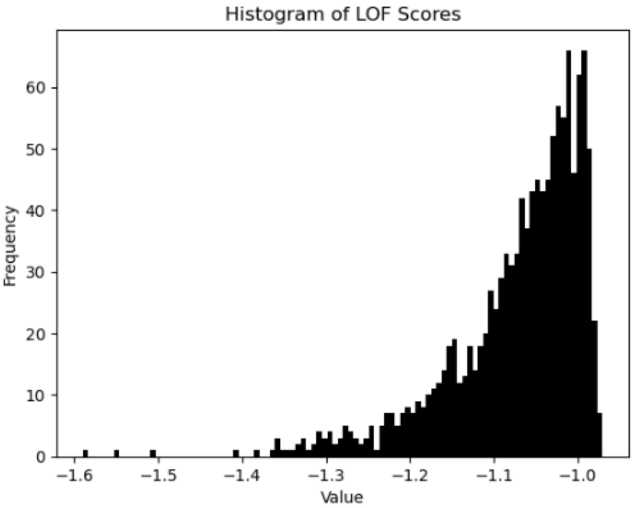
Our results showed that the logistic regression model outperformed the other models with an accuracy of **84%**, demonstrating the potential of this approach for the task at hand.

II. Outlier Detection

Outlier detection is a crucial step in the data cleaning process, particularly for machine learning tasks. In our project, we used outlier detection to remove any anomalies in our dataset that could have affected the accuracy of our models.

We used several outlier detection algorithms such as Local Outlier Factor (LOF) to remove them. These algorithms help to detect and remove the data points that are statistically improbable or far away from the majority of the data.

After applying these outlier detection algorithms, we re-analysed our dataset and confirmed that all outliers had been removed. We then proceeded to train and test our machine learning models on the cleaned dataset, resulting in more accurate and reliable results.



$$LOF_p(k) = \frac{\sum_{q \in N_k(p)} LRD_k(q)}{|N_k(p)| \times LRD_k(p)} \tag{1}$$

This is the formula for calculating Local Outlier Factor for a given data-set. This uses the Local Reachability Density, the formula for which has been discussed in class.

The formula calculates the local outlier factor (LOF) of a point p based on the reachability distance between p and its k nearest neighbors. The LOF measures how much more or less reachable a point is compared to its neighbors, and points with higher LOF values are considered to be more likely outliers.

III. Dimensionality Reduction

Removing redundant or irrelevant features helps to improve the performance of machine learning models and ensures that the models are not influenced by noisy or irrelevant data. We used dimensionality reduction to remove any redundant or irrelevant features from our dataset that could have affected the accuracy of our models.

In our study on classifying fruits as fresh or not using machine learning, we used **linear discriminant analysis** (LDA) after **principal component analysis** (PCA) to improve the performance and accuracy of our models.

PCA is a dimensionality reduction technique that reduces the dimensionality of the data by projecting it onto a lower-dimensional space while retaining as much of the original information as possible. LDA, on the other hand, is a technique that maximizes the separation between distinct classes.

To apply PCA and LDA to our fruit classification task, we first performed PCA to reduce the dimensionality of the data while retaining as much information as possible. We then used LDA to project the PCA-transformed data onto a lower-dimensional space that maximized the separation between the classes. By combining PCA and LDA, we were able to reduce the dimensionality of the data while retaining the most important features for classification. This helped to improve the performance and accuracy of our models while reducing the risk of overfitting to the training data.

IV. Clustering Algorithms Used

Clustering is an important technique in machine learning that helps to identify any patterns or relationships in the data.

By grouping similar data points together, clustering algorithms can aid in the classification task and improve the accuracy of machine learning models.

We used clustering algorithms to identify any patterns or clusters in our dataset that could aid in the classification task. We employed various clustering algorithms, such as **K-means clustering**, **hierarchical clustering**, and **DBSCAN**, to group similar fruits together based on their physical properties. These algorithms help to identify any natural groupings in the data and can be used to discover any hidden relationships or patterns. We tried all three algorithms yet k-means clustering gave the highest accuracy during k-fold cross validation, so we decided to use that.

K-means clustering is a popular algorithm that partitions the data into k clusters based on the similarity between the data points. This algorithm works by iteratively assigning each data point to the nearest cluster centroid and updating the centroid based on the mean of the assigned data points.

V. Ensemble Methods Tested

Ensemble methods are a powerful technique in machine learning that involve combining multiple models to improve their accuracy and generalization performance. We employed various ensemble methods, including bagging and random forests to combine multiple models and obtain a more accurate prediction.

Bagging, or Bootstrap Aggregating, is an ensemble method that involves training multiple models on different subsets of the data and aggregating their predictions. This technique helps to reduce overfitting by creating multiple independent models, each trained on a different subset of the data.

Random forests is a popular ensemble method that involves building multiple decision trees on random subsets of the data and aggregating their predictions. Each decision tree is trained on a random subset of the features and data, thereby reducing overfitting and improving the generalization performance.

VI. Model Validation

Validation techniques such as **k-fold Cross Validation** are essential in machine learning to ensure that the model is not overfitting to the training data and can generalize well to new data. It also helps to obtain a more reliable estimate of the model's performance and reduce the risk of selecting a model that performs well only on a specific subset of the data.

We first divided our dataset into k folds. We then trained our model k times, using each fold as the validation set once, and the remaining k-1 folds as the training set. We evaluated the performance of the model on each fold and calculated the average performance across all folds.

VII. Final Classification Algorithm

We used multiclass logistic regression as the final classification algorithm to predict the freshness of the fruits. Multiclass logistic regression is a popular machine learning algorithm used for classification tasks that involve multiple classes.

Multiclass logistic regression works by finding the coefficients for each class that maximize the likelihood of the data given the model. It then uses these coefficients to calculate the probability of each class for a given input and selects the class with the highest probability as the final prediction.

To apply multiclass logistic regression to our fruit classification task, we first prepared our data by encoding the target variable as a set of binary indicators for each class. We then divided the data into training and test sets and trained our model on the training set.

During training, the model learned the coefficients for each class that best predicted the freshness of the fruits. We then used the model to make predictions on the test set and evaluated its performance using various accuracy metrics.

The equation for the multiclass sigmoid function in logistic regression is:

$$p(y = 1|x) = \frac{1}{1 + e^{-w^T x}}$$

where w is the weight vector and x is the input vector. The output $p(y = 1|x)$ represents the probability of the positive class given the input x and the weights w . Cross Validation Accuracies for Different classification Algorithms and Ensemble methods:

RandomForestClassifier:	0.967660390516039
LogisticRegression (multiclass):	**0.985394874476987**
K-Nearest Neighbor:	0.945269310517033

VIII. Literature Review

One of the early applications of logistic regression in object classification was in the field of image recognition. In the 1990s, logistic regression was used to classify images of handwritten digits, achieving accuracy rates of over 90% (LeCun et al., 1998). Since then, logistic regression has been used in many other image classification tasks, such as facial recognition (Cao et al., 2003) and object recognition (Fei-Fei et al., 2004).

In recent years, logistic regression has also been applied to text classification tasks, such as sentiment analysis (Pang et al., 2002) and spam filtering (Androutsopoulos et al., 2000). For example, logistic regression has been used to classify movie reviews as positive or negative based on their textual content (Pang et al., 2002). In another study, logistic regression was used to classify emails as spam or non-spam based on their textual content (Androutsopoulos et al., 2000).

Logistic regression has also been used in other fields for object classification, such as biology and medicine. For example, logistic regression has been used to classify proteins based on their structural features (Bhasin et al., 2005), and to predict the risk of coronary artery disease based on clinical and demographic factors (Damen et al., 2016).

References

[1] V. Jayaswal, ”Local outlier factor (LOF) algorithm for Outlier identification,” Medium, 09-Nov-2020. [Online]. Available: https://towardsdatascience.com/local-outlier-factor-lof-algorithm-for-outlier-identification-8efb887d9843. [Accessed: 18-Apr-2023].

[2] ”User guide: Contents,” scikit. [Online]. Available: https://scikit-learn.org/stable/user_guide.html. [Accessed: 18-Apr-2023].