Apple Quality Prediction Modelling

Xintan Lin and Jay Verma
Prof. Fabio Crestani
16 April 2024

Overview

In this project, we will analyse apple dataset to build a model that will predict the quality of apple. This dataset contains information about various attributes of a set of fruits, providing insights into their characteristics. The dataset includes details such as fruit ID, size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and quality.

First, we preprocessed the data and then we explored it for finding relationships between attributes. We used histograms to see how the data was spread out. Then we removed outliers using z-score.

We calculated correlation values to understand how different attributes relate to each other. From these correlations, we identified size, sweetness, juiciness, ripeness, and acidity as key factors influencing apple quality. With these attributes in mind, we constructed a predictive model aimed at predicting apple quality.

We applied four different models—gradient boosting machines, random forest regressor, support vector machine, and artificial neural networks—to predict apple quality. After evaluating the models using Mean Squared Error (MSE) values, we determined that the support vector machine (SVM) exhibited the highest performance among them, making it the most suitable choice for predicting apple quality.

Introduction

The purpose of this report is to analyse the Apple Quality dataset available on Kaggle to predict the quality of apples. To achieve this, we will employ four distinct models. The dataset will be split evenly into training and testing sets facilitating a comprehensive evaluation of model performance.

Data Exploration

The Apple Quality dataset comprises 4001 entries and 9 columns, offering comprehensive insights into the characteristics of various fruits. It encompasses essential attributes such as fruit ID, size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and quality. There are two distinct categories for the quality attribute: "good" and “bad," enabling a clear distinction between fruit quality levels.

Data Preprocessing

Before training the model, the data was preprocessed to improve the model's performance. The preprocessing steps included:

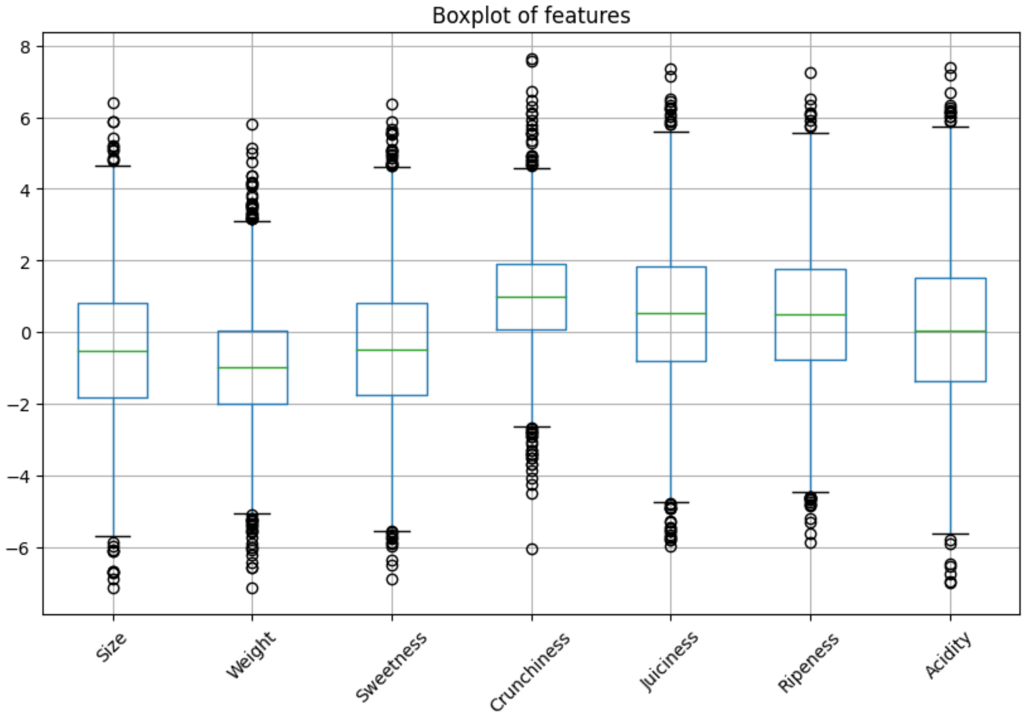
1. The last row was dropped as it contains NaN values.
2. ‘A_id’ attribute was dropped as it has no relevance for our analysis.
3. Data type of ‘Acidity’ attribute was converted from object to float to calculate correlation.

We visualised scatter plot of the data to find correlation between attributes. From the scatter plot matrix, we found that there is a strong positive correlation between the size, weight, and sweetness of the apples, and a slight negative correlation between

the juiciness and quality, size and quality. We also observe counterintuitive facts, that bigger size apple tend to have smaller weight, smaller size apples

could have bigger weight. It's the same effect
between ripeness and sweetness. More ripe apples could have
smaller sweetness, while less ripe apples, could lead to higher sweetness.

We found that the data have skewed distributions and/or outliers by plotting histograms. Based on the box plot, we can observe that some of the attributes have outliers beyond the whiskers of the box plot. Specifically, we can see that there are some extreme values in
the Size, Weight, Sweetness, Crunchiness, Juiciness, and Ripeness attributes. These outliers may affect the performance of predictive models, and we will remove them.Then we removed outliers using z-score. We calculated z-score for each value of the attributes and removed those values with a z-score greater than 3.



After removing outliers based on z-scores, the final dataset has 3902 rows. This represents a loss of 96 rows (or 2.4% of the original dataset). Based on this and summary statistics, we can conclude that the impact of removing outliers was minimal and did not result in a significant loss of data.

To compute correlations between various attributes and quality, we created a new attribute 'quality_numeric' in which we mapped 'good' to 1 and 'bad' to '0'. Correlations between various attributes and 'quality_numeric' are shows below:

Attributes	quality_numeric
Size	0.250592
Weight	-0.002243
Sweetness	0.250875
Crunchiness	-0.007663
Juiciness	0.255056
Ripeness	-0.263595
Acidity	-0.013298

Based on the correlation values, there was a strong positive correlation between each numerical feature (Size, Sweetness, Juiciness) and numerical attribute (quality_numeric) and a negative correlation between each numerical feature (Weight, Ripeness, Acidity, Crunchiness) and numerical attribute (quality_numeric).

Here are some variables that we can consider using to predict apple quality based on the correlations and domain knowledge: Size, Sweetness, Juiciness, Ripeness and Acidity.

Model Training and Evaluation

To build an effective prediction model for apple quality, we have tried four machine learning methods. We illustrate each one in below:

1) Gradient boosting machine. Gradient boosting machine is using multiple weak learning functions to accumulatively approximating output functions. We set the object error function as squared error, using 100

number of estimators, with max depth of 6 and learning rate of 0.1. These hyper parameters are chosen by past experiments as the optimal choice. We also went on several parameter experiments, find that the number of estimators can be reduced to 50 without changing the performance, while the number of training samples can have relatively high positive correlation impact on the MSE. By experiments, we find the minimal MSE is 0.09.

2) Random Forest. Random forest collects votes from a number of weak decision trees, form a strong classifier. There are two kinds of Random forest algorithms involved in sklearn: Random forest regressor or random forest classifier. We run both of them to see their performance. In random forest regressor, the number of estimators have brought strong effects to its regression accuracy. We found since number 15 upwards, the MSE value has been stabilised around 0.09. But if we use random forest classifier, with 'gini' as criterion and 10 above estimators, we can achieve MSE error as 0.03. Thus, random forest classifier archives a nice result.

3) Support Vector Machine. Support vector machine is a strong classifier for wider arrange of data. We import SVM model from sklearn, and build it with linear kernel, polynomial kernel, radial basis function kernel, and sigmoid kernel to test its prediction accuracy. By experiment, we found, with linear kernel, we have MSE error as 0.06. In polynomial kernel, we have MSE as 0.03. Using radical basis kernel, we have MSE as 0.02. Sadly with sigmoid kernel, we got 0.15 MSE error. Thus, we obtain a good apple quality prediction model using SVM radical basis kernel.

4) Artificial neural net. In modern trends, neural nets have been proven as a very efficient tool for complex classification tasks. Thus in this task, we also implement a neural net for predicting apple qualities. To do that we carried following steps:

I. Separate training and testing datasets from the task. We use a ratio of 80% training data, and 20% testing data.

II. Scale datasets to a standard distribution. We archieve that using standard scaler with fit transform.

III. Build an ANN model with three layers of fully connected dense layer (linear layer in pytorch). The Input dimension is our attributes number, the output dimensions are 64, 32 and 1. We also use 'relu' as the activation function. Generally without loss of efficiency, we use Adam gradient stepper with 0.001 learning rate.

IV. Training the model with 100 epochs with 32 as batch size.

After above four steps, we gained an effective apple quality predictor. At test stage, we induct test result by using model evaluation. The final MSE result is 0.06.

Conclusion

To sum up above machine learning methods, we find that scikit learn library has provide most efficient and powerful tools for gradient boosting method, random forest decision trees, and support vector machine. While in neural net machine learning, we use keras as a high level neural nets builder. In terms of predictor quality, we find support vector machine with radial basis function kernels have reached the best performance. That means MSE as 0.02, meaning in 100 apples, there will be 2 false positive or true negatives apples. Thus we complete apple quality prediction modelling.