Apple Quantity visualization

By Jing Xue

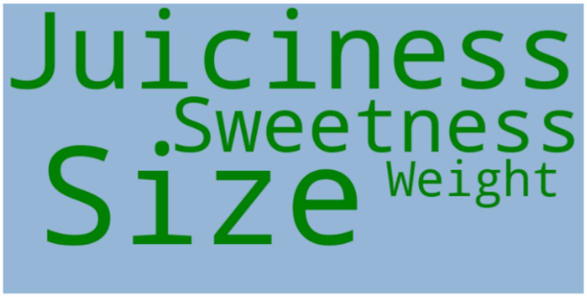

Figure 1 Feature Importance Word Cloud for Apple Quality Prediction

Figure 1 shows the most important features in the decision tree model and their relative importance in analyzing apple quality predictions. These features are presented in the form of word clouds, where a larger font size indicates a higher importance of the feature for prediction. For example, the most important feature is "Juiciness", followed by "Size" and "Sweetness".

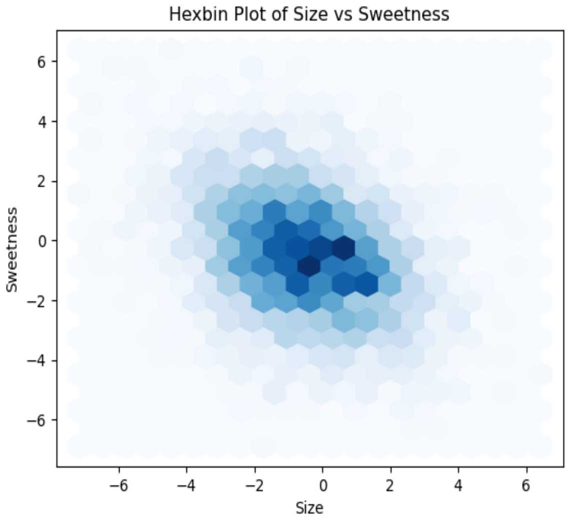

Figure 2 Hexbin Plot of Apple Size and Sweetness

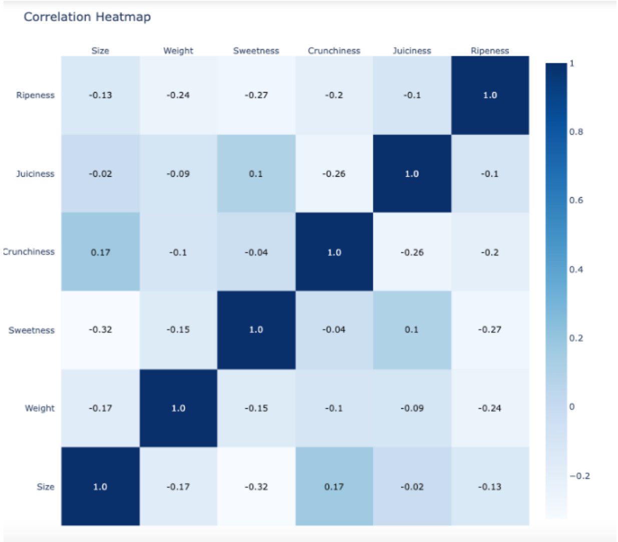

Figure 3 Correlation Heatmap of Apple Attribute

Figures 2 and 3 show the distribution of apple characteristics and their correlation, respectively. Figure 2 shows the relationship between apple size and sweetness using a hexagonal heat map, with dark blue indicating areas of dense distribution. Most of the apples have size and sweetness clustered around the mean value, with no clear linear relationship between the two. Figure 3 shows a heat map of the correlation between the various characteristics of apples: size was positively correlated with weight (0.17), there was a slight positive correlation between sweetness and juiciness (0.10), and ripeness was negatively correlated with sweetness (-0.27) and size (-0.13). Other variables such as weight, acidity and crispness were weakly correlated with the main characteristics.

To further analyze the data, I decided to plot the correlation between the columns on a violin plot, where the colors represent the quality of the apples, with high quality apples being red and poor quality apples being green.

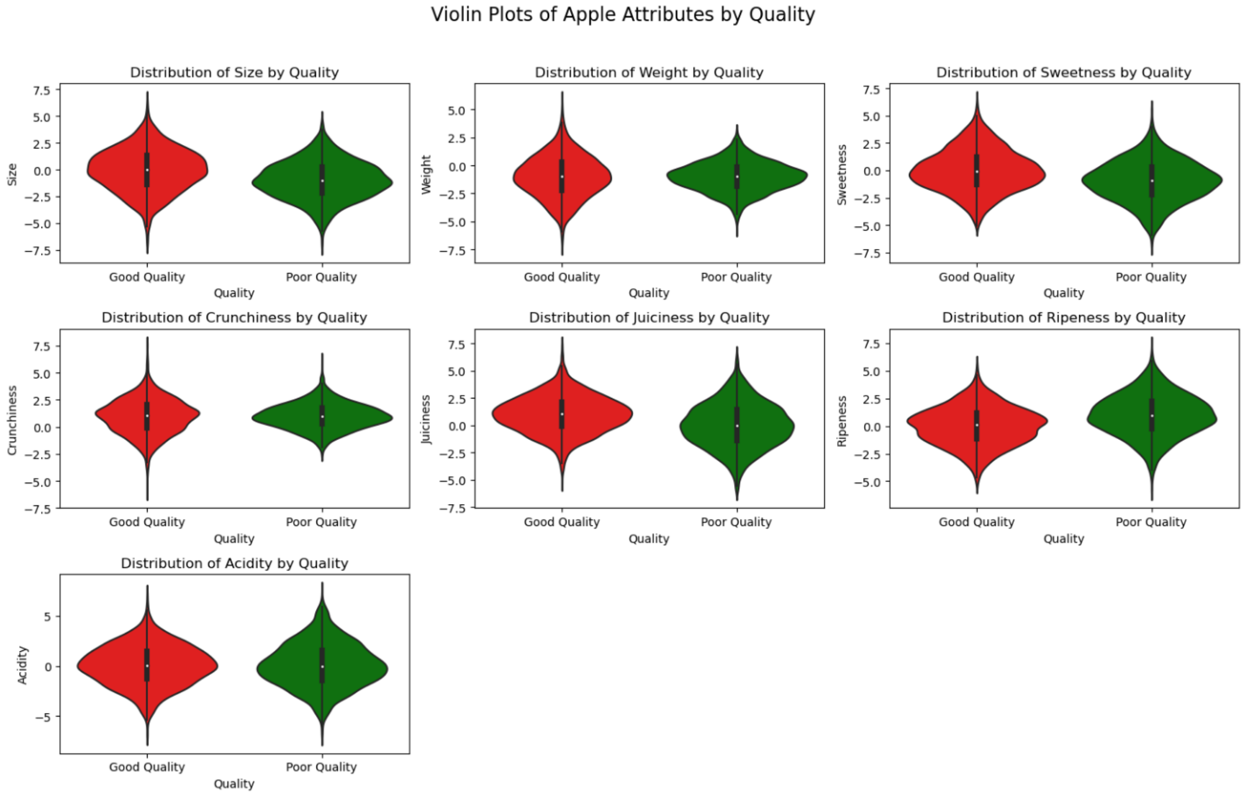

Figure 4 Violin Plots of Apple Attributes by Quality

Figure 4 is a violin plot matrix showing the distribution of different apple characteristics across quality categories. The horizontal axis represents the quality categories of apples (“high quality” and “low quality”) and the vertical axis represents the values of the characteristics, including size, weight, sweetness, crispness, juiciness, ripeness and acidity. The white dots indicate the median values for the characteristics. The thick line in the center indicates the interquartile range (IQR), which runs from Q1 (lower quartile, 25%), Q2 (median, 50%) to Q3 (upper quartile, 75%).

The wide part of the violin plot indicates a high probability distribution of the characteristic value, while the narrow part indicates a lower probability. Therefore, it can be observed that characteristics such as sweetness and juiciness are more prominent in “high quality” apples, while “low quality” apples have slightly higher ripeness values. Other characteristics, such as size, weight, crispness and acidity, differed less between the two quality categories.

Figure 5 shows the decision tree model used to categorize apple quality as “good” or “bad”. The tree splits the dataset by key characteristics (e.g., juiciness, sweetness, size, and weight) to make classification predictions. The root node starts with juiciness, and for apples with juiciness values less than or equal to -0.416, the model classifies them as “bad”. If the juiciness value is high, the model further evaluates the sweetness and categorizes the data into subgroups based on the sweetness value. The decision-making process continues until the leaf node is reached, where the final classification is done.

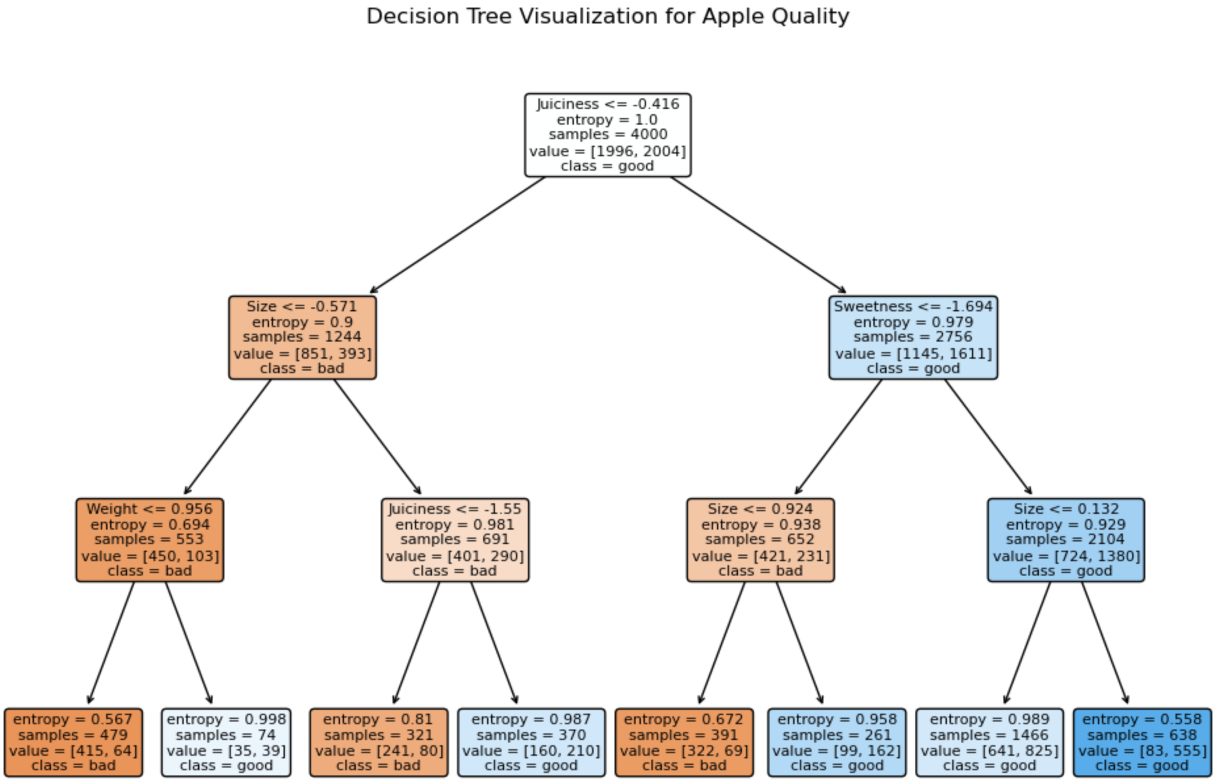
Decision Tree Visualization for Apple Quality

Figure 5 Apple Quality Decision Tree

Important: In my code, I chose five charts to demonstrate the key characteristics and classification model logic of Apple's quality data. With multiple chart presentations, we can check whether different analytics methods convey consistent information. For example, the important characteristics in the word cloud diagram (Figure 1) also occupy key positions in the decision tree (Figure 5). Figure 1 provides the basis for feature selection, from which Figures 2 and 3 allow for further analysis of the relationships between key features (e.g., size and sweetness), and finally the classification results are refined in Figure 4, which integrates the findings from the previous charts to produce the decision tree (Figure 5). This gives us an overall picture of the impact of key characteristics (e.g. size, weight) on apple quality.

Data and methord: The data I used in this assignment comes from the apple quantity's dataset on the open source kaggle. You can also download it directly from my Github repository. All charts were generated using python. Depending on the type of chart, python packages such as wordcloud and graphviz are installed. More detailed information on the use of python packages can also be found in the repositories.

Github Link:
https://github.com/Jiny-Xue/apple-quantity-visualization