AI/ML Project

Apple Quality

Team Members

Abhay sharma

Agrim

Ankush

Swastik

CONTENT

- Problem Statement
- Objective
- Tools Used
- Data Summary
- Features discription
- Exploratory Data Analysis
- Conclusions

PROBLEM STATEMENT

In the dataset 'apple,' encode the categorical column 'Quality' into numerical values using scikit-learn's LabelEncoder. This preprocessing step aims to convert qualitative data into a format suitable for machine learning models. Begin by importing LabelEncoder from scikit-learn. Define the categorical columns to be encoded, such as 'Quality.' Iterate through each specified column, applying LabelEncoder to transform categorical labels into numerical representations. This process ensures consistency and compatibility with algorithms that require numerical input. Verify that the encoding accurately reflects the original categorical values, preserving their meaning within the dataset. Display the modified dataset to confirm successful transformation. By converting categorical variables into numerical equivalents, the dataset becomes suitable for predictive modeling tasks. This encoding process maintains data integrity while enhancing the dataset's utility for machine learning algorithms. Ultimately, the goal is to prepare the data effectively for model training, enabling accurate predictions while retaining the interpretability of the original categorical information."

OBJECTIVE

The objective is to preprocess the 'apple' dataset by encoding the categorical column 'Quality' into numerical values using scikit-learn's LabelEncoder. This transformation facilitates the utilization of machine learning algorithms that require numerical input. The specific goals include ensuring consistency and compatibility of the dataset for predictive modeling tasks. By converting categorical variables into numerical equivalents, the dataset becomes suitable for training classification models. Furthermore, the objective encompasses verifying the accuracy of the encoding process to preserve the original meaning of categorical labels within the dataset. The ultimate aim is to prepare the data effectively for model training, enabling accurate predictions while retaining the interpretability of the original categorical information. This preprocessing step enhances the dataset's utility by making it more accessible to a broader range of machine learning algorithms, ultimately improving the efficiency and effectiveness of predictive modeling workflows."

Tools Used

- **Jupyter Notebook is used as IDE.**
- **Pandas and NumPy are used for Data Manipulation & Pre-processing and Mathematical functions respectively.**
- **Exploratory data analysis is automated by data prep.**
- **For visualization of the plots, Matplotlib, Seaborn, Plotty are used.**

DATA SUMMARY

Numerical Data

1.Size
2.Weight
3.Sweetness
4.Crunchiness
5.Juiciness
6.Ripeness
7.Acidity

Categorical Data

1.Quality

Unique Data

1.A_id

DATA SUMMARY

- **This is the Apple quality dataset. In the below table it shows the top 5 rows .**

	A_id	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Quality
0	0	-3.970049	-2.512336	5.346330	-1.012009	1.844900	0.329840	-0.491590	good
1	1	-1.195217	-2.839257	3.664059	1.588232	0.853286	0.867530	-0.722809	good
2	2	-0.292024	-1.351282	-1.738429	-0.342616	2.838636	-0.038033	2.621636	bad
3	3	-0.657196	-2.271627	1.324874	-0.097875	3.637970	-3.413761	0.790723	good
4	4	1.364217	-1.296612	-0.384658	-0.553006	3.030874	-1.303849	0.501984	good

FEATURES DESCRIPTION

id: Unique for each Apple

Size: size of the apple

Weight: weight of the apple

Sweetness: sweetness value of apple

Crunchiness: crunchiness value of apple

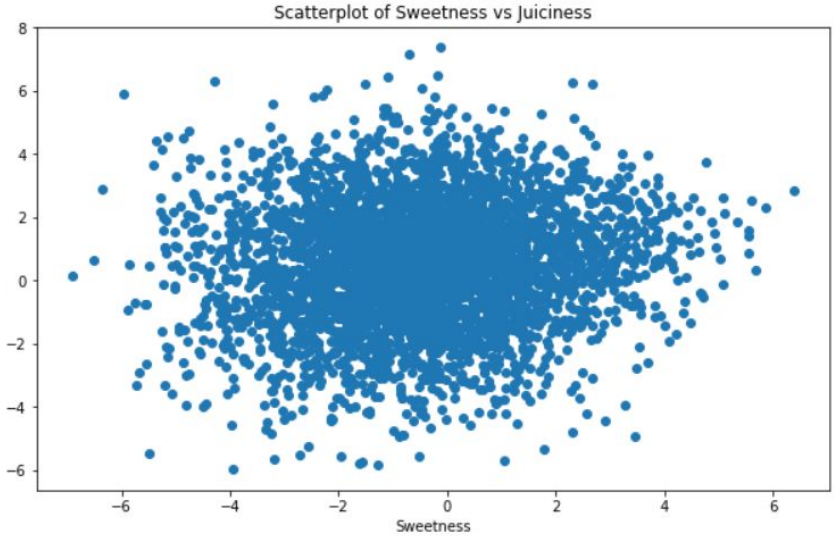
Juiciness: juiciness value of apple

Ripeness: ripeness value of apple

Acidity:acidity value of apple

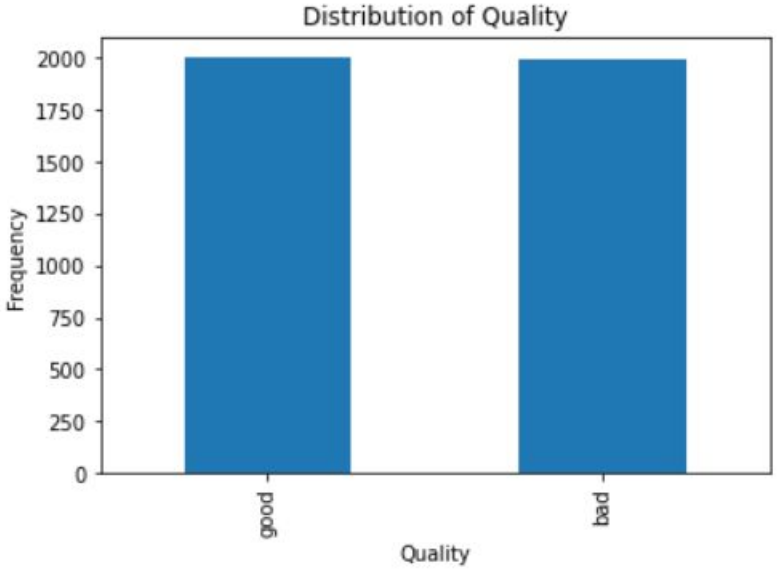
Relationship between sweetness and juiciness value

- The data points show an upward trend as you move from left to right. This aligns with the observation of a positive correlation between math and reading scores
- While there's a positive trend, the data points aren't perfectly aligned in a straight line. This indicates there's a spread of scores, meaning some students with high math scores might have lower reading scores (and vice versa)



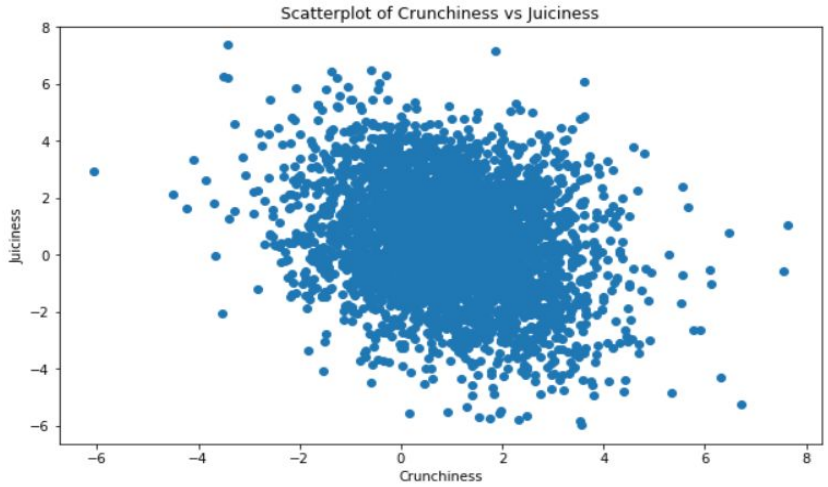
Distribution of quality

- The chart depicts how frequently observations fall into different quality categories (good, bad). This is similar to how a scatter plot shows the distribution of data points across a space
- The vertical axis labeled "Frequency" counts the number of observations that belong to each quality category. This is analogous to how scatter plots use position on the axes to show how many data points fall into each combination of variable values.



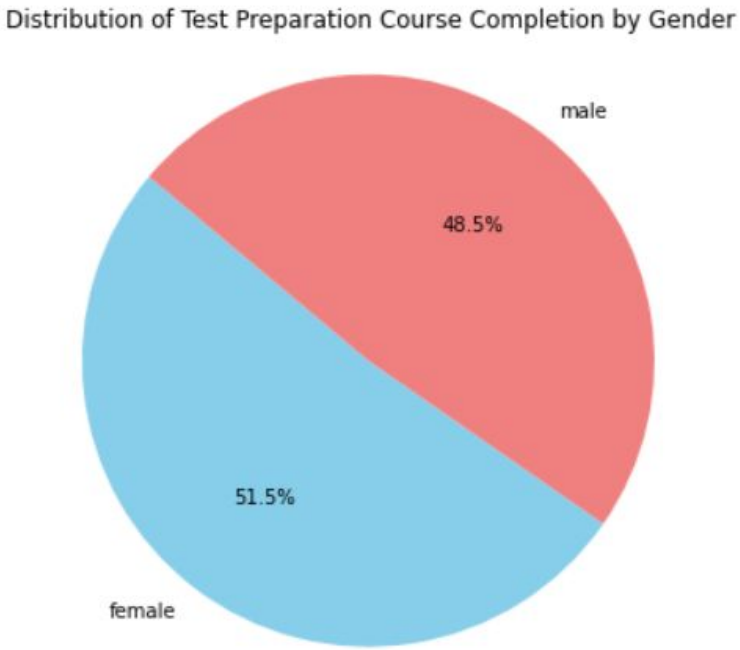
Relationship between sweetness and juiciness value

- There's a faint upward trend as you move from left to right. This suggests a possible positive correlation between churchiness and juiciness, though it's weak.
- The data points are not aligned in a straight line. There's a spread of scores, indicating that some very churchy foods can be rated as not juicy, and vice versa.



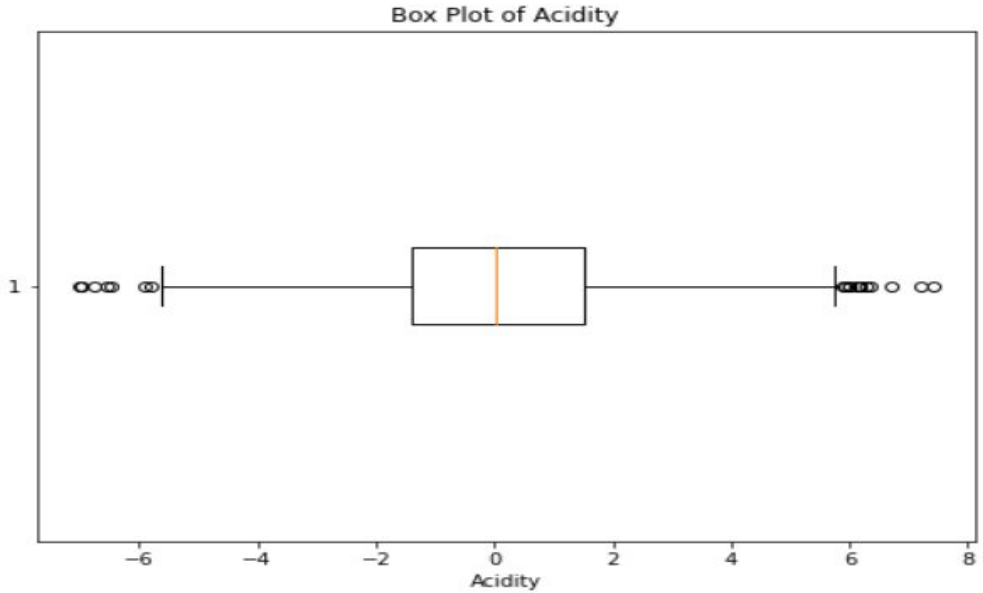
Distribution of quality

- **The pie chart divides the whole (100%) into two slices labeled "good" (50.1%) and "bad" (49.9%). This shows the proportion of items categorized as good quality versus bad quality.**



Acidity Value

- The box in the center of the plot contains the middle 50% of the data. The line in the middle of the box is the median, which splits the data in half with half the values lower than the median and half higher



Conclusions

- Overall quality of apples is good
- Most of the apples are good which have high sweetness value.
- Most of the apples are good which have high juiciness value.
- Maximum apples have high sweetness value.
- Maximum apples have high crunchiness value.