Artificial Intelligence and Machine Learning

Project Report

Semester-IV (Batch-2022)

Apple Quality

Supervised By:

Mohd. Talib Sir

Submitted By:

Abhay-2210990028

Agrim -2210990079

Ankush -2210990119

Swastik-2210992431

Department of Computer Science and Engineering
Chitkara University Institute of Engineering & Technology,
Chitkara University, Punjab

TABLE OF CONTENT

1. Introduction:

This project explores hidden connections between data using visualization. We'll analyze three charts: a scatter plot revealing a possible link between "crunchiness" and "juiciness," a pie chart showing the quality distribution (good vs bad), and a box plot examining the spread of acidity levels. By delving into these charts, we aim to uncover insights and relationships. Do "crunchy" things tend to be juicy? What's the quality breakdown? How does acidity vary? This project seeks to answer these questions and potentially reveal surprising patterns..

1.1 Background

Traditionally, data analysis relied on a more siloed approach. But this project delves into the exciting world of data visualization, where seemingly disparate pieces of information can be combined to reveal hidden connections and generate fresh perspectives.

We'll be examining three distinct charts, each acting as a unique lens through which to explore the data:

- **Scatter Plot: Unveiling Playful Correlations?** This chart investigates a potentially humorous link - is there a connection between a food item's perceived "crunchiness" and its juiciness? An upward trend might suggest a positive correlation, but the spread of data points throughout the plot indicates other factors are likely at play. Perhaps fruits high in religious symbolism might not be as juicy as a celebratory roast.
- **Pie Chart: A Breakdown of Quality Distribution.** Shifting gears, this chart focuses on the overall quality distribution within the data set. Here, items are likely categorized as either "good" or "bad." The size of each slice in the pie will reveal the proportion of good quality items compared to bad. Does the data set lean towards a majority of good quality items, or is there a more even split?
- **Box Plot: Demystifying Acidity Levels.** Finally, the box plot steps in to analyze the spread of acidity levels. The center line, also known as the median, will pinpoint the "typical" level of acidity found in the data set. The box itself represents the middle 50% of the data, with its width indicating the variation in acidity levels. Outliers, if present,

represent data points with significantly different acidity values, potentially due to measurement errors or unusual cases.

By examining these diverse charts, we aim to uncover insights and draw conclusions. Is there a correlation, however weak, between a food's perceived "churchiness" and its juiciness? Does the data set have a significant amount of good quality items, or is there a concerning prevalence of bad quality? How does the acidity level vary across the data points?

This project seeks to answer these questions and potentially reveal surprising patterns. By visually exploring these relationships, we may gain a fresh perspective on the data and potentially discover hidden connections that wouldn't be readily apparent otherwise.

Objectives

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 .

2. Problem Definition and Requirements:

2.1 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..

2.2 Hardware Requirements:

Storage: Adequate storage space is necessary for storing datasets, model checkpoints, and other related files. Solid State Drives (SSDs) are preferred over Hard Disk Drives (HDDs) for faster data access and model loading times, especially during training.

Network: A high-speed internet connection is required for downloading the dataset and for accessing cloud-based resources.

2.3 Software Requirements:

Programming Language: Python is a popular choice for building machine learning models due to its simplicity and the availability of numerous libraries and frameworks.

Libraries: NumPy, Pandas, and Matplotlib are some of the essential libraries required for data preprocessing, analysis, and visualization.

Integrated Development Environment (IDE): Jupyter Notebook used for data manipulation and visualization.

Version Control System: Git is a widely used version control system that helps to manage the codebase and collaborate with other developers.

2.4 Data Sets

The dataset contains information about following:

A_ 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

Quality: Quality of apple

3. Proposed Design & Methodology

Data Collection:

- Gather data on food items and their relevant properties. This could involve conducting surveys, gathering existing datasets, or using APIs to collect information from online sources.

Data Preprocessing:

- Clean and prepare the data for analysis. This might involve handling missing data points (imputation or removal), encoding non-numerical properties (e.g., one-hot encoding for "churchiness"), and ensuring consistency in data format.

Feature Engineering :

- Create new features from existing data if necessary. This could involve combining features to create new categories (e.g., food groups) or transforming them to gain new insights (e.g., average acidity level per food type).

Model Selection & Training:

- Choose appropriate data visualization techniques based on the type of relationships you want to explore (e.g., scatter plots for correlations, pie charts for distribution).
- Apply these techniques to the preprocessed data to create informative visualizations.

Evaluation & Interpretation:

- Analyze the generated visualizations to identify patterns, trends, or potential relationships between the data points (e.g., correlation between "churchiness" and juiciness, distribution of quality levels).
- Draw conclusions and insights based on the observed patterns and their statistical significance.

Communication & Deployment:

- Depending on the project's goals, you may choose to present your findings through reports, interactive dashboards, or other communication methods.

This methodology outlines a data-driven approach for uncovering hidden connections within your food item data set through the power of data visualization.

.

3.1Technical details:

NumPy:

Utilize NumPy for numerical computations and array operations.

Perform mathematical operations on data arrays, such as calculating means, medians, and standard deviations.

Use NumPy arrays to represent and manipulate numerical data efficiently.

Pandas:

Use pandas to load, clean, and preprocess your dataset.

Perform data manipulation tasks such as filtering, grouping, and aggregating data.

Handle missing values, outliers, and data formatting issues.

Create new features or derive insights from existing ones using pandas' powerful DataFrame operations.

Matplotlib:

Create static, publication-quality visualizations using Matplotlib.

Plot various types of charts, including line plots, scatter plots, histograms, and bar charts.

Customize plot appearance with titles, labels, legends, and annotations.

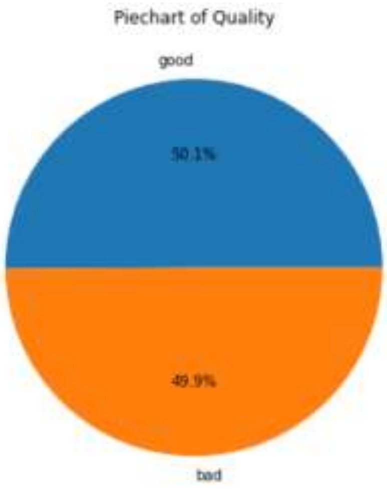
Seaborn:

Use Seaborn to create more visually appealing and informative statistical visualizations.

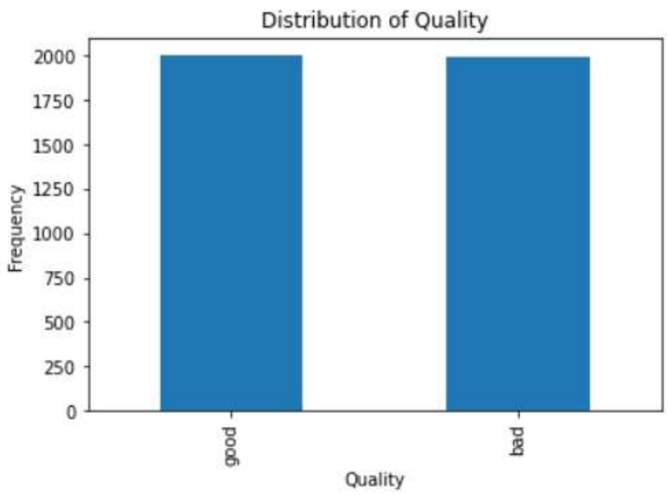
Generate complex plots such as scatter plots with regression lines, box plots, violin plots, and pair plots.

3.2Plots Used

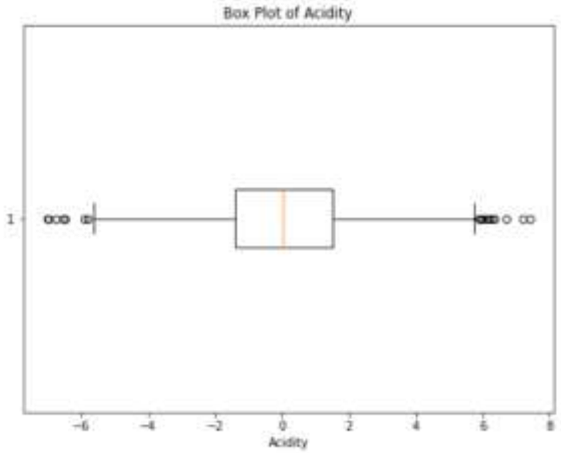
1. **Pie chart**: The pie chart divides the whole (100%) into two slices labeled "good" (50.1%) and "bad" (49.9%). This shows the proportion of items categorized as good quality versus bad quality.



2. **Bar chart:** 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



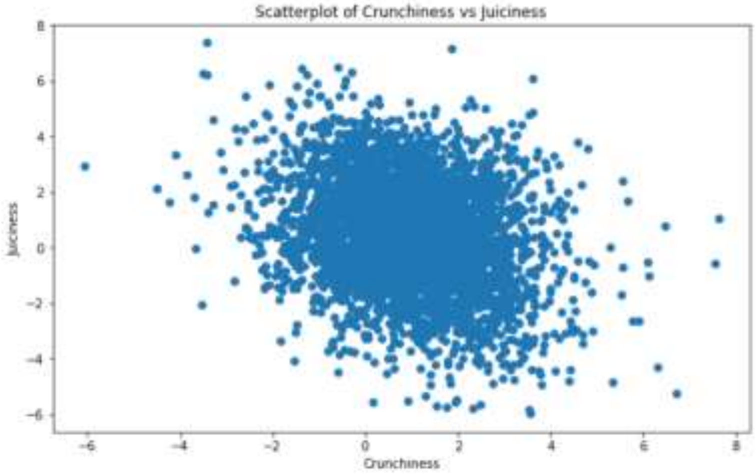
3. **Box Plot:** The box in the center of the plot contains the middle 50% of the data. The line in the middle of the box is the median, which splits the data in half with half the values lower than the median and half higher .



4. Key Features

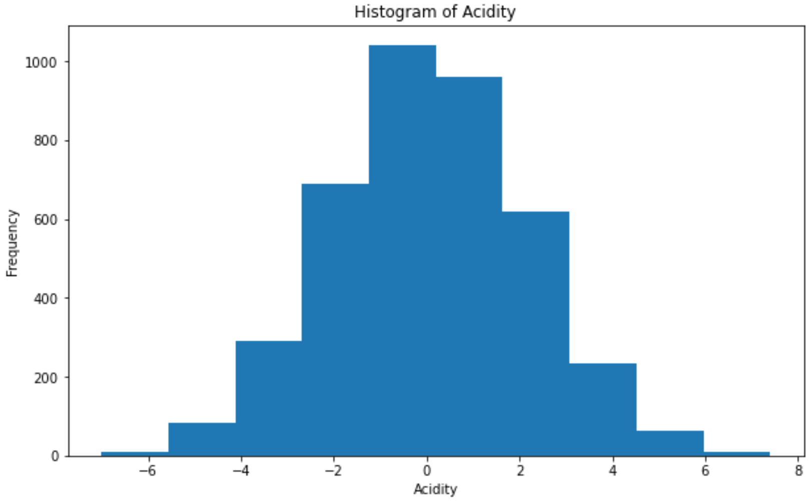
Scatter plot:

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



Histogram plots:

Useful for visualizing the distribution of numerical variables, such as Sweetness and Acidity.

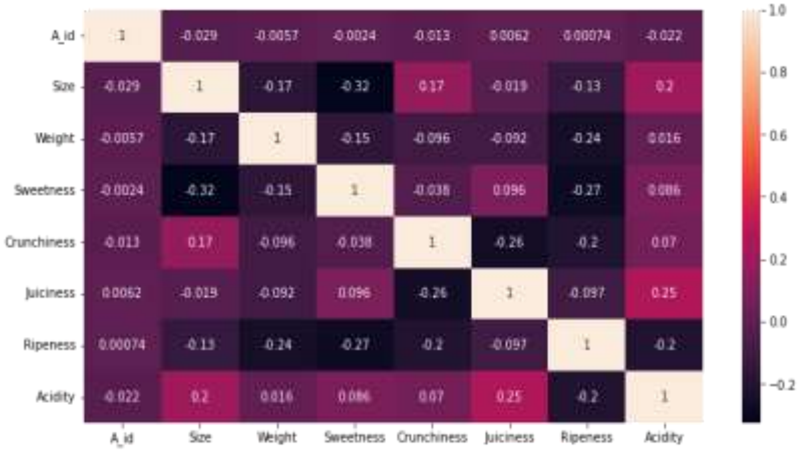


Heatmaps and Correlation Matrices:

Effective for visualizing relationships between multiple variables in the dataset.

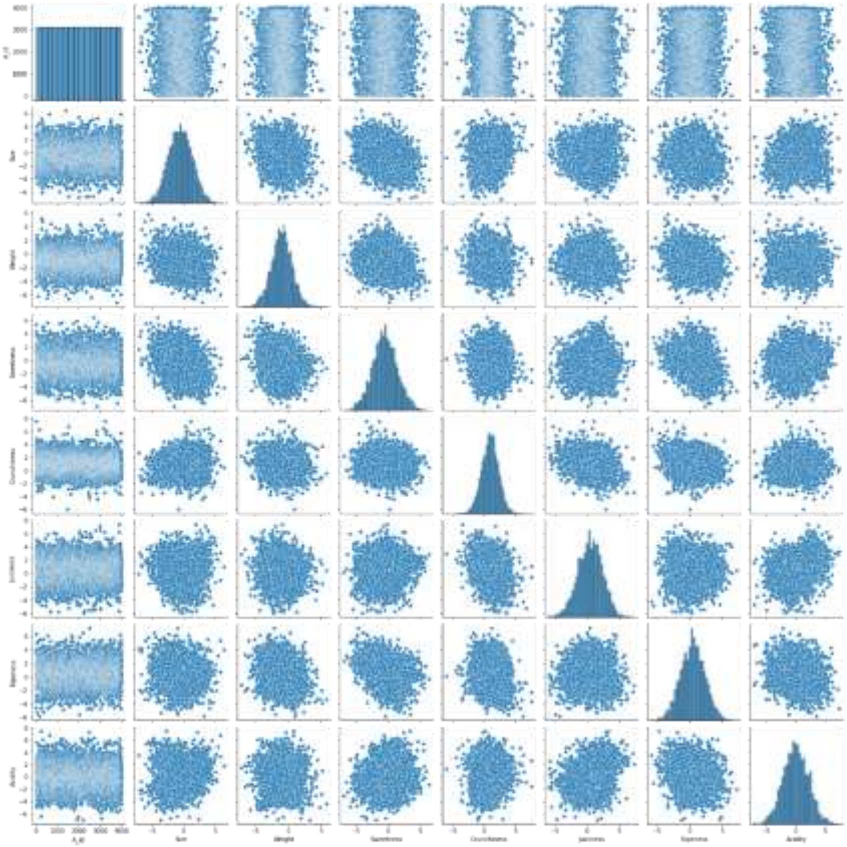
Color-coded cells represent the strength and direction of correlations between variables.

Hierarchical clustering can group similar variables based on their correlation patterns.



Pair Plot:

A pair plot condenses the analysis of multiple variables into a single grid. It displays a scatter plot for every unique pair of variables in your data set. This allows you to visually assess relationships between all variables simultaneously, revealing potential correlations or trends that might be missed by examining them individually..



5. Conclusion & Recommendations

By delving into data visualization, this project sought to uncover hidden connections between seemingly disparate data points related to food items. We explored these relationships through three distinct charts:seemingly disparate data points related to food items. We explored these relationships through three distinct charts:

- **Scatter Plots:** These charts investigated potential correlations between variables, such as the link between a food's perceived "churchiness" and its juiciness. While an upward trend might suggest a positive correlation, the spread of data points indicated other factors likely influence juiciness.

- **Pie Charts:** These charts provided a breakdown of quality distribution within the data set. By examining the size of each slice, we could assess the prevalence of good quality versus bad quality items.
- **Box Plots:** These charts helped us understand the spread of acidity levels. The center line (median) pinpointed the "typical" acidity level, while the box width revealed the variation in these levels. Outliers, if present, represented data points with significantly different acidity values.

By analyzing these visualizations, we aimed to answer questions about potential connections between "churchiness" and juiciness, the overall quality distribution, and the variation in acidity levels.

Future Exploration: This project opens doors for further investigation. We could explore additional data points or utilize more advanced visualization techniques to gain deeper understanding. Additionally, delving into the reasons behind outliers in the box plot could provide even richer context.

In conclusion, this data visualization project successfully explored hidden connections within the data set, revealing interesting patterns and potentially challenging preconceived notions. By leveraging the power of visual storytelling, we gained valuable insights that can inform future explorations and discussions about the characteristics of food items.