APPLE QUALITY ML PREDICTIONS

DATA ANALYTICS BOOTCAMP PROJECT 4

PRESENTED BY KENTO NAKAJIMA

JANUARY 2024

DATA SOURCE
TOOLS
MODELS

Kaggle.com

Python Pandas, Python Matplotlib, Python Seaborn, Sqlite3, Scikit-Learn

Logistic Regression

Random Forest

ETL

EXTRACT DATA:

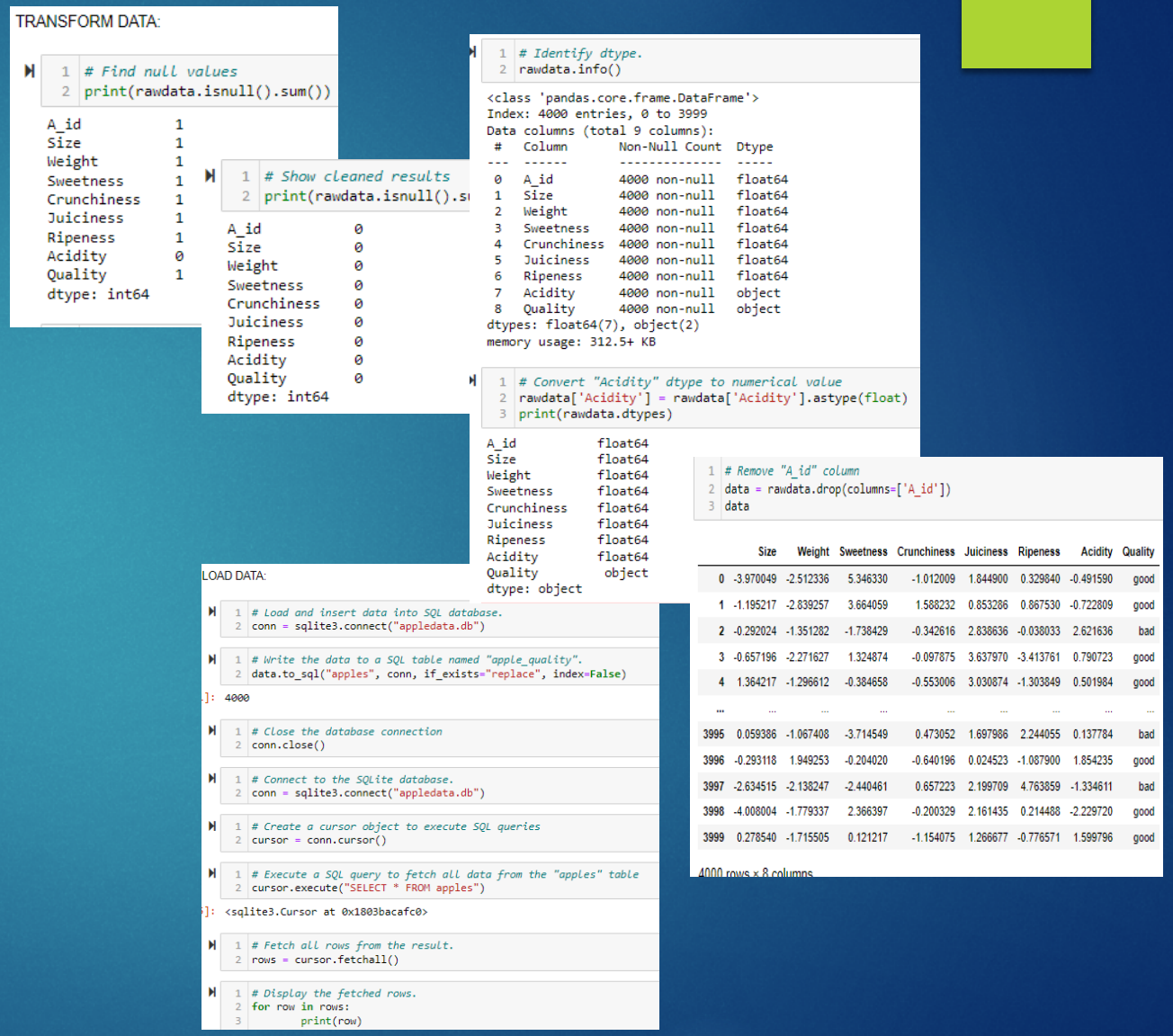
- Apple Quality CSV from Kaggle.com
- 4000+ rows
- 9 columns ; A_id, Size, Weight, Sweetness, Crunchiness, Juiciness, Ripeness, Acidity, and Quality

TRANSFORM DATA:

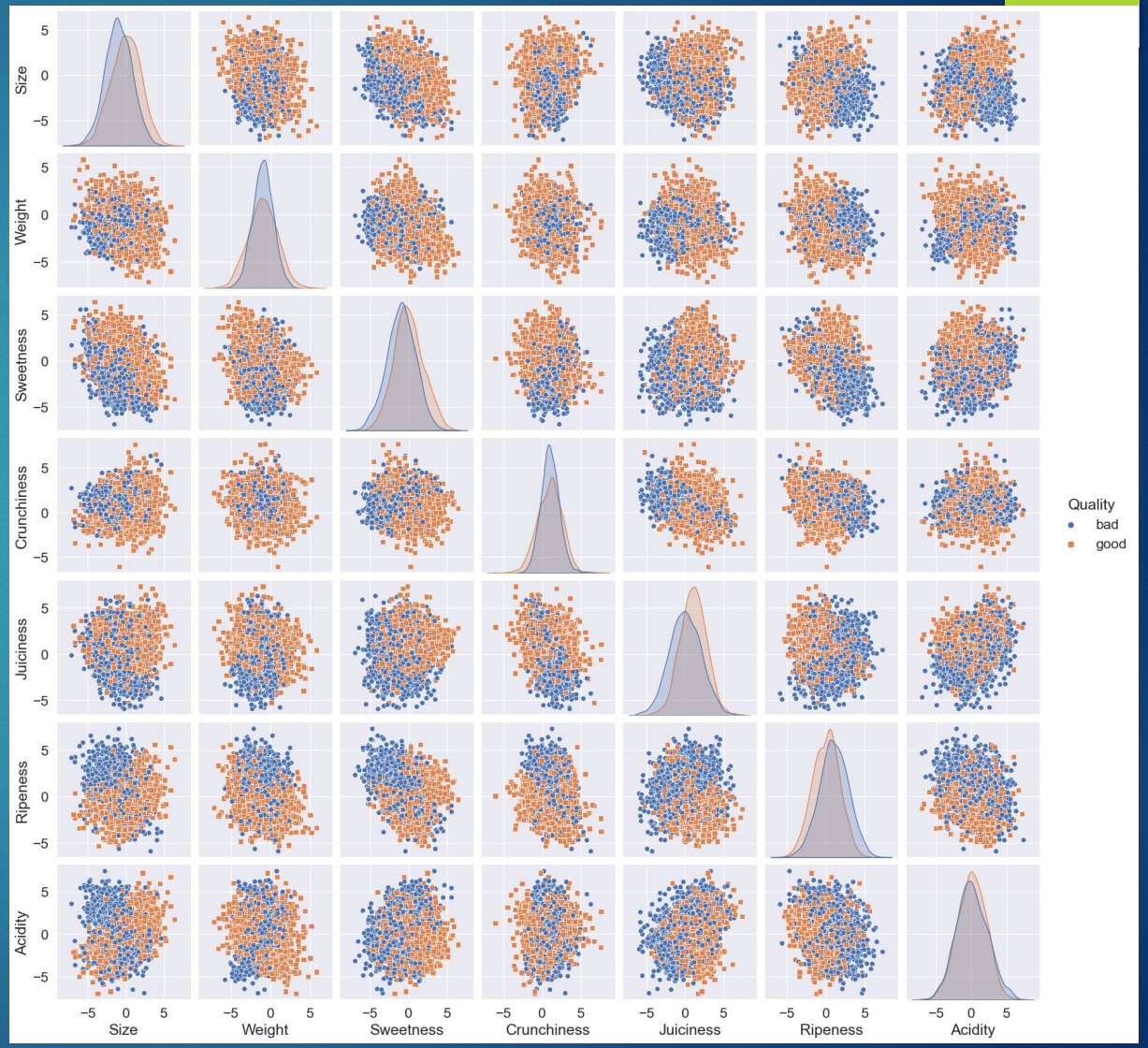
- Cleaning – Missing Values
- Conversion – Data Types
- Selection – Column Removal

LOAD DATA:

- Load SQLite Database
- Execute SQL Query
- Retrieve SQLite Database to DataFrame.
- Visualization - Seaborn's Pairplot.



DATA VISUALIZATION



LOGISTIC REGRESSION

Relatively balanced in brecision, recall, and F1-score – Fair class treatment

Close precision and recall values – No class imbalance issues

Balanced false positives and false negatives – No material bias in types of errors – Room for further investigation

RANDOM FOREST

High accuracy

Few errors

Well balanced precision, recall, and F1-score in both classes

A few false positives and negatives

*** LOGISTIC REGRESSION MODEL EVALUATIONS ***

All Features Included

ACCURACY SCORE: 0.754

CONFUSION MATRIX:

	Predicted 0	Predicted 1
Actual 0	368	131
Actual 1	115	386

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0.0	0.76	0.74	0.75	499
1.0	0.75	0.77	0.76	501
accuracy			0.75	1000
macro avg	0.75	0.75	0.75	1000
weighted avg	0.75	0.75	0.75	1000

*** RANDOM FOREST MODEL EVALUATIONS ***

All Features Included

ACCURACY SCORE: 0.886

CONFUSION MATRIX:

	Predicted 0	Predicted 1
Actual 0	450	62
Actual 1	52	436

Feature Importance Ranking:

	Feature	Importance
4	Juiciness	0.165536
5	Ripeness	0.164077
0	Size	0.160593
2	Sweetness	0.147335
6	Acidity	0.128319
1	Weight	0.121169
3	Crunchiness	0.112971

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0.0	0.90	0.88	0.89	512
1.0	0.88	0.89	0.88	488
accuracy			0.89	1000
macro avg	0.89	0.89	0.89	1000
weighted avg	0.89	0.89	0.89	1000

LOGISTIC REGRESSION MODEL OPTIMIZATION

*** LOGISTIC REGRESSION MODEL EVALUATIONS ***

All Features Included

ACCURACY SCORE: 0.754

CONFUSION MATRIX:

	Predicted 0	Predicted 1
Actual 0	368	131
Actual 1	115	386

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
0.0	0.76	0.74	0.75	499
1.0	0.75	0.77	0.76	501
accuracy			0.75	1000
macro avg	0.75	0.75	0.75	1000
weighted avg	0.75	0.75	0.75	1000

*** LOGISTIC REGRESSION MODEL #2 EVALUATIONS ***

Less Features - Size and Weight Removed

ACCURACY SCORE: 0.67

CONFUSION MATRIX:

	Predicted 0	Predicted 1
Actual 0	368	131
Actual 1	115	386

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
bad	0.67	0.68	0.67	499
good	0.67	0.66	0.67	501
accuracy			0.67	1000
macro avg	0.67	0.67	0.67	1000
weighted avg	0.67	0.67	0.67	1000

*** LOGISTIC REGRESSION MODEL #3 EVALUATIONS ***

Less Features - Crunchiness Removed

ACCURACY SCORE: 0.755

CONFUSION MATRIX:

	Predicted 0	Predicted 1
Actual 0	368	131
Actual 1	115	386

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
bad	0.76	0.74	0.75	499
good	0.75	0.77	0.76	501
accuracy			0.76	1000
macro avg	0.76	0.75	0.75	1000
weighted avg	0.76	0.76	0.75	1000

FINAL THOUGHTS

Thank You, Drew!