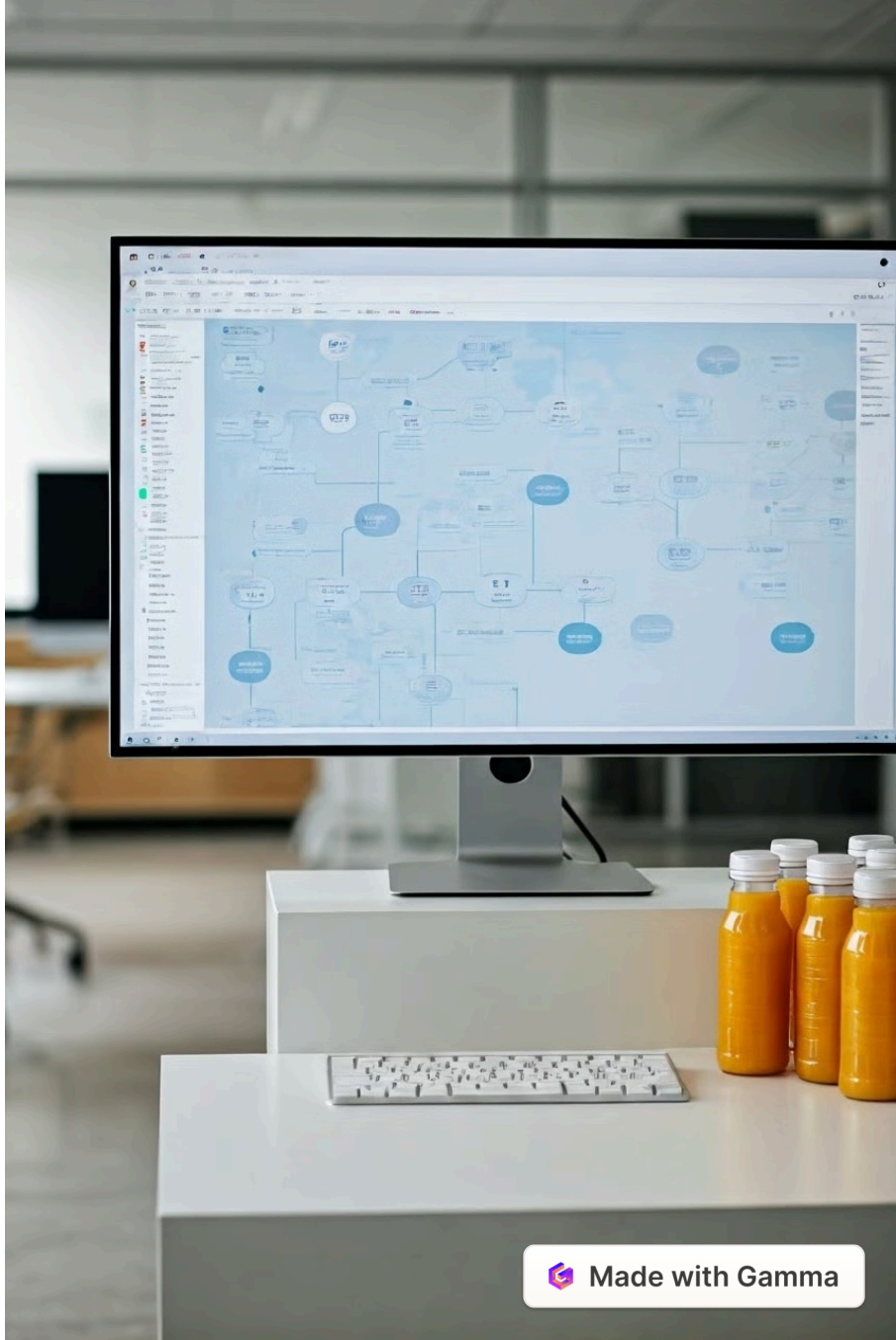
Decision Tree Analysis on OJ Dataset

This presentation explores the factors impacting customer purchase decisions between Citrus Hill and Minute Maid orange juice brands. We'll analyze the OJ dataset using decision tree methods, evaluating model performance and key predictors.

H by Harry Depina



Dataset Overview

1 Sample Size

1,070 observations of orange juice purchases, providing a robust dataset for analysis.

2 Prediction Goal

Determine whether a customer chooses Citrus Hill or Minute Maid based on various factors.

3 Key Features

Price, Brand Loyalty, and Store Type serve as primary predictors in our model.

Data Splitting Strategy

1

Training Set

70% of data used to build and train the decision tree model.

2

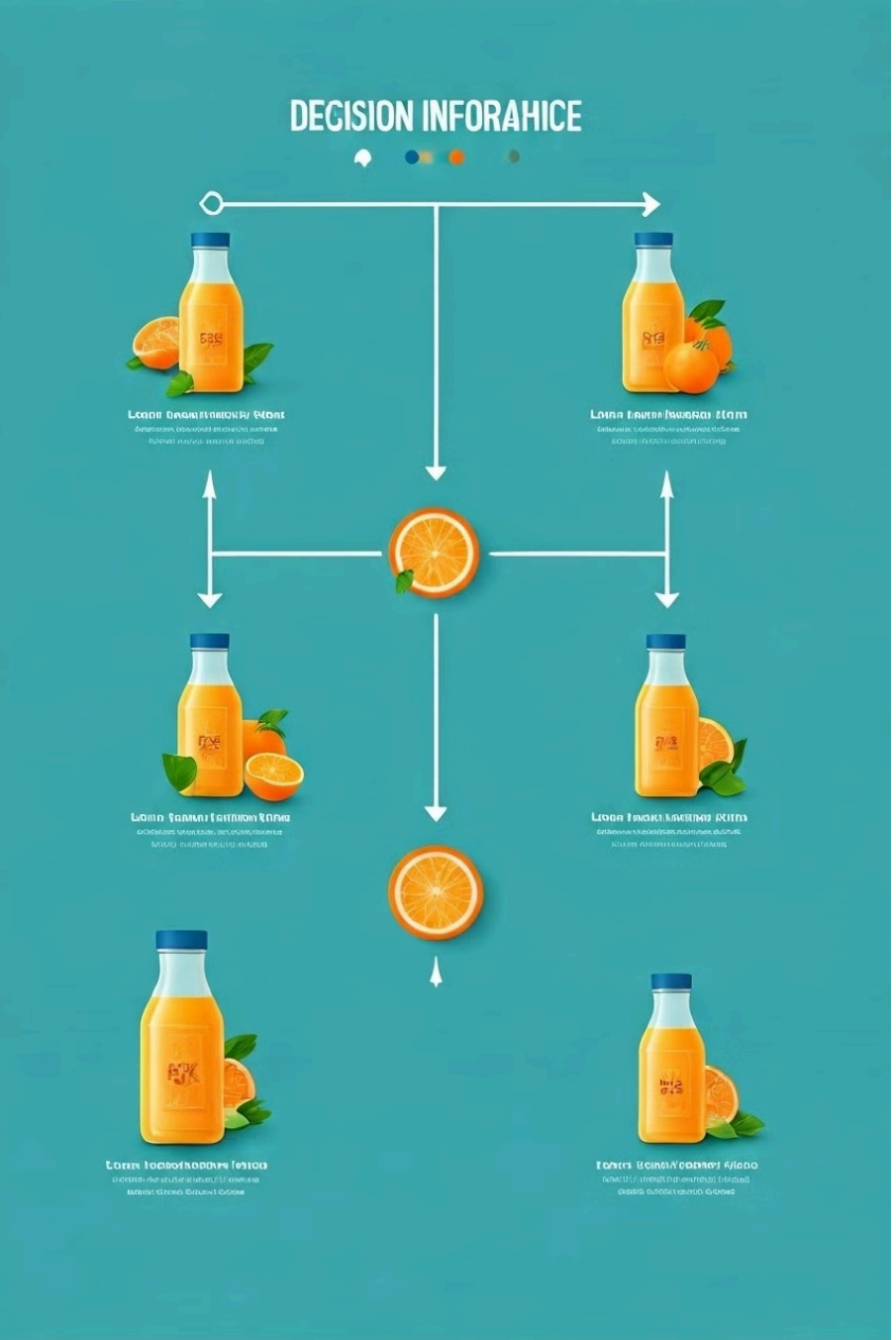
Testing Set

30% reserved for evaluating model performance and generalization capabilities.

3

Cross-Validation

Implemented to fine-tune model parameters and prevent overfitting.



Decision Tree Visualization

1

Root Node: Brand Loyalty

The most influential factor in predicting orange juice brand choice.

2

Secondary Splits

Price differences and store-related variables further refine predictions.

3

Terminal Nodes

Represent predicted classes (CH or MM) with associated probability estimates.

Model Evaluation

Accuracy

Measured by correct prediction rate on training data. Indicates model's basic performance.

ROC Curve

Visualizes trade-off between true positive and false positive rates across thresholds.

AUC Score

0.839 indicates good model performance in differentiating between the two brands.

Key Purchase Factors

Brand Loyalty

Most critical predictor, influencing initial splits in the decision tree.

Price

Affects decisions, even among loyal customers, when significant price differences exist.

Store Environment

Certain retail settings impact customer choices, potentially due to promotions or availability.

Model Performance Summary

Metric	Value	Interpretation
AUC Score	0.839	Good discrimination
Accuracy	84.2%	Solid predictive power
F1 Score	0.83	Balanced precision and recall

Conclusion and Future Steps

Key Findings

Brand loyalty, pricing, and store characteristics are main drivers behind purchase decisions.

Model Effectiveness

AUC score of 0.839 indicates good performance, but there's room for improvement.

Refinement Strategies

Consider pruning or different splitting criteria to enhance model accuracy.

Feature Expansion

Incorporate customer demographics or seasonal factors for more nuanced predictions.