

NEW DELHI

FORE SCHOOL OF MANAGEMENT

**Academic Year 2023-25**

Machine Learning for Managers

TOPIC: To find out the pizza sales segmentation of the store based on different sizes

Submitted to: Prof. Amarnath Mitra

Submitted by:

Name- Arindam Chakraborty

Roll No.-321128 PGDM 32 Section C

*Acknowledgement*

I would like to extend my heartfelt gratitude to our prof. Amarnath Mitra Sir for introducing us to the fascinating world of supervised learning through this project. Their guidance, support, and expertise were invaluable in helping us navigate the complexities of the project and ensuring its successful completion. Their dedication to fostering our understanding and application of machine learning concepts has been truly inspiring. Thank you for your unwavering commitment and for providing us with this invaluable learning opportunity.

|  |  |
| --- | --- |
| Contents | Page no. |
| **Objective of the project** | 4 |
| **Data descriptions** | 5 |
| **Analysis** | 8 |
| **Observation** | 18 |
| **Managerial Insights** | 19 |

Table of contents:

1. **Objectives**

1.1 The first objective is to segment the pizza data of the store using supervised learning algorithms using Decision tree.

1.2 The second objective is to determine the number of appropriate classification model by comparing and contrast using logistic regression, KNN (k-nearest neighbour) and SVM (support vector machine).

1.3 The third objective is to identify significant variables or features and their thresholds for classification.

2. Description of Data

**2.1. Data Source, Size, Shape**

2.1.1. Data Source – https://www.kaggle.com/code/azamatjonkhasanzoda/pizza-dataset-analysis-clustering-time-series/input

2.1.2. Data Size (in KB | MB | GB …) – **10 MB**

2.1.3. Data Shape | Dimension:

Number of Variables - **12**

Number of Records – **50483**

**2.2. Description of Variables**

2.2.1. Index Variable(s): pizza\_id, Order\_id

2.2.2. Variables or Features having Categories | Categorical Variables or Features (CV)

2.2.2.1. Variables or Features having Nominal Categories | Categorical Variables or Features - **Nominal Type**:

pizza\_name\_id, , pizza\_size, pizza\_category, pizza\_ingredients, pizza\_name

2.2.2.2. Variables or Features having Ordinal Categories | Categorical |

**Ordinal Type**- order\_time, order\_Date

2.2.3. Non-Categorical Variables or Features: Quantity, Unit\_price, total\_price Pizza\_ID: Unique identifier for each Pizza

Order\_id: Unique identifier for each identifier

pizza\_name\_id: Code name of the pizza name

pizza\_size: Size of pizza, S, M, L, XL, XXL

Pizza\_category: Variant or version of the Pizza

Pizza Ingredients: Ingredients of the Pizza

Pizza\_name: Name of the pizza

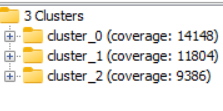
Order\_time: Time at which the pizza was ordered

Order\_date: date at which the pizza was ordered

**2.3. Descriptive Statistics**

2.3.1. Descriptive Statistics of Outcome Categorical Variable

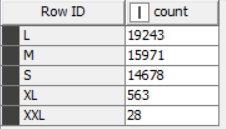
It provides the statistics of cluster variable (categorical variable) by giving frequency as well as relative frequency (in %).



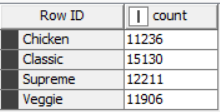
2.3.2. Descriptive Statistics: Categorical Variables or Features

2.3.2.1. Count | Frequency Statistics

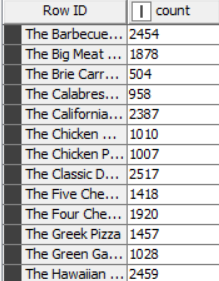
**Pizza Size count**

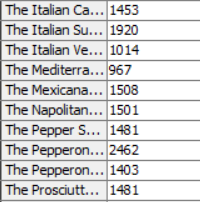


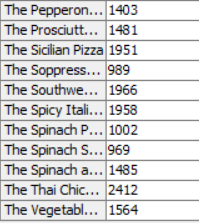
**Pizza Category**



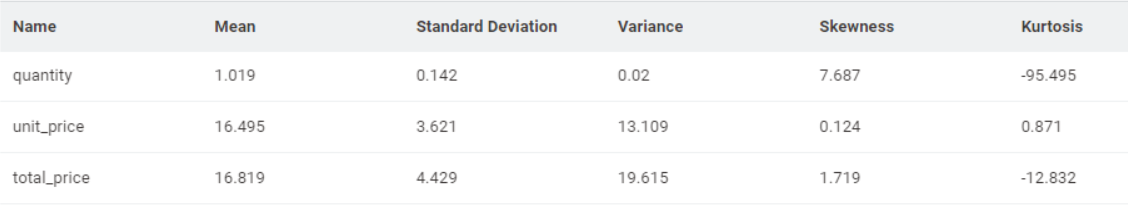
**Pizza Name**





****

2.3.3.2. Measures of Dispersion



**Source of data-**

https://www.kaggle.com/code/azamatjonkhasanzoda/pizza-dataset-analysis-clustering-time-series/input

3. Analysis of Data

**3.1. Data Pre-Processing**

**3.1.1. Missing Data Statistics and Treatment**

3.1.1.1.1. Missing Data Statistics: 16

3.1.1.1.2. Missing Data Treatment:

3.1.1.1.2.1. Removal of Records with More Than 50% Missing Data

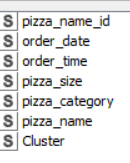
3.1.1.2.1. Missing Data Statistics: Categorical Variables or Feature

3.1.1.2.2. Missing Data Treatment: Categorical Variables or Features - 10

3.1.1.2.2.1. Removal of Variables or Features with More Than 50% Missing

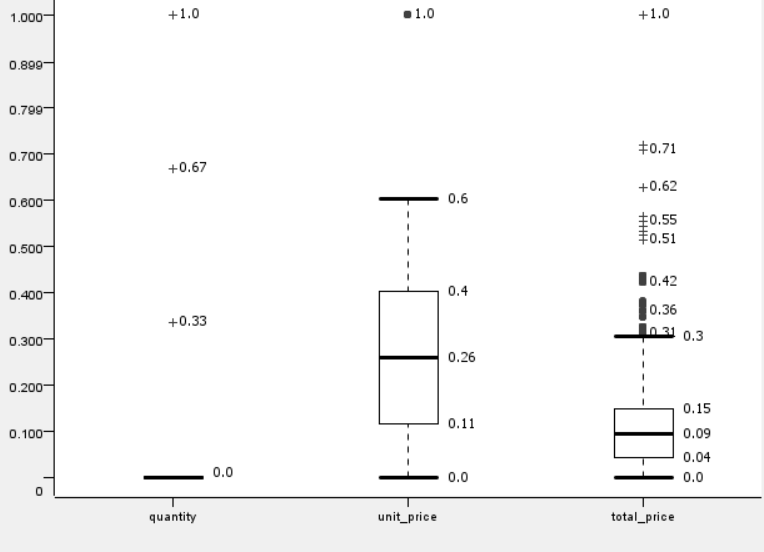
**3.1.2. Numerical Encoding of Categorical Variables or Features** (Encoding Schema

* Alphanumeric Order)
* In this case, category to number node will be used to encode the categorical variable



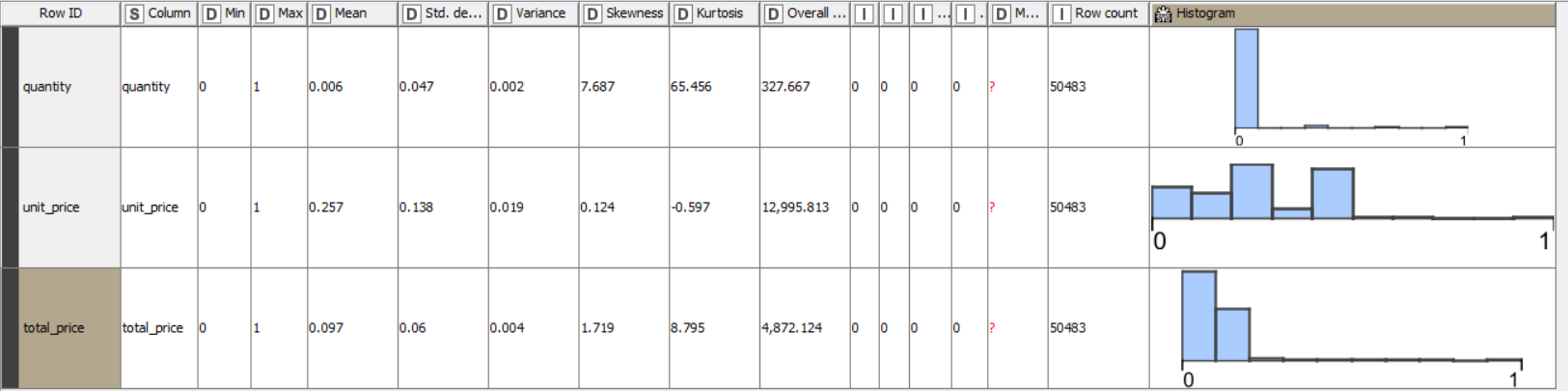
**3.1.3. Outlier Statistics and Treatment** (Scaling | Transformation)

3.1.3.1.1. Outlier Statistics: Non-Categorical Variables or Features



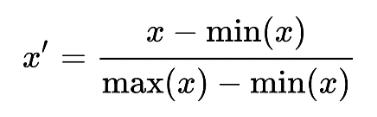
3.1.3.1.2. Outlier Treatment: Non-Categorical Variables or Features

3.1.3.1.2.1. Standardization

3.1.3.1.2.2. Normalization using Min-Max Sca

Min-max Scalar technique is a normalizer technique used in data pre-processing to scale numerical features to a specific range, typically between 0 and 1.

The formula for min-max normalization is:



3.1.3.1.2.3. Log Transformation

**3.1.4. Data Bifurcation: Training & Testing Sets**

The training and testing data have been bifurcated into 80% and 20% respectively.

**3.2. Data Analysis**

**3.2.1. Supervised Machine Learning Classification Algorithm: Decision Tree**

* A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the input space into smaller regions based on feature values, creating a tree-like structure of decisions. At each node of the tree a decision is made based on the value of a specific feature, and the data is split into subsets. This process continues until a stopping criterion is met, such as reaching a maximum depth or no further improvement in impurity reduction.
* In this project, decision tree will be the classification algorithm used for unsupervised learning. The metrics used in decision tree is Gini coefficient.
* When using decision tree, we will be also seeing comparison when no pruning method is used and when pruning method is used.

**3.2.2. Supervised Machine Learning Classification: Other Methods Logistic Regression**

It is a supervised learning algorithm used for binary classification tasks. It models the probability of the input belonging to a particular class using the logistic function. The algorithm learns the relationship between input features and the probability of the binary outcome, making it suitable for predicting categorical outcomes.

In this project, logistic regression will be used and the metric used in logistic regression is iteratively reweighted least squares (solver method).

**K-Nearest Neighbours**

K-Nearest Neighbours (KNN) is a supervised learning algorithm that is also used for both classification and regression tasks. It predicts the classification of a data point by finding the majority class among its k nearest neighbours in the feature space. KNN's performance heavily depends on the choice of distance metric and the value of k, making it sensitive to the dataset's characteristics.

In this project, KNN will be used and the metric used is Euclidean distance. For comparison, we will be using k =7 till k=19 in steps of 2 i.e. k=7,9,11,13,15,17 and 19.

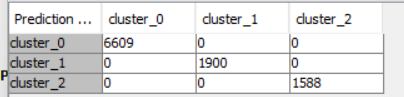
**Support Vector Machines**

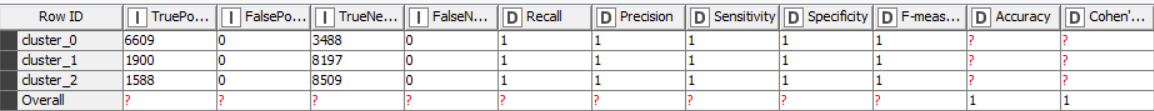
Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space, maximizing the margin between them. SVM can handle high-dimensional data and is effective even in cases where the number of features exceeds the number of samples.

In this project, the kernel used will be polynomial and the parameters are power = 1, bias = 1 and gamma = 1.

**3.2.2.1. Classification Model Performance Evaluation of Decision Tree by using Confusion Matrix**

Without Pruning





**Cluster 0**

* This cluster has a high number of true positives and true negatives indicating that the model correctly classified most instances within this cluster.
* The precision and recall scores are both very high suggesting that the model effectively identifies true positives while also minimizing false positives.

**Cluster 1**

* This cluster has a lower recall and precision compared to cluster 0, indicating that the model's performance is not as strong for this segment.
* The number of false positives is relatively high, suggesting that the model may misclassify some instances within this cluster.

**Cluster 2**

* This cluster has a relatively high recall and precision, indicating that the model performs well.
* The number of false positives is relatively low suggesting that the model effectively minimizes misclassifications within this cluster.
* Both sensitivity and specificity scores are high indicating that the model correctly identifies both true positives and true negatives within this cluster.

Small pizzas: The model seems to perform well for small pizzas, with 1914 correctly classified and only a small number misclassified as medium (2112) and large (10).

Medium pizzas: The model also performs well for medium pizzas, with 10035 correctly classified and a relatively small number misclassified as large (467) and extra large (18).

Large pizzas: The model seems to have more difficulty with large pizzas. While 10,035 were correctly classified, there were also substantial misclassifications as small (1223) and extra large (632).

Extra Large (XL) pizzas: The model performs well for extra large pizzas, with 16,300 correctly classified and only a small number misclassified as medium (467) and large (632).

XXL pizzas: There were only 10 XXL pizzas, and all were correctly classified.

In conclusion, the decision tree model seems to perform well for small, medium, and extra large pizzas, but has more difficulty accurately classifying large pizzas

Despite the lower performance metrics, the specificity is very high indicating that the model correctly identifies true negatives within this cluster.

**Comparative analysis of decision tree with and without pruning**

Pruning generally improves precision and specificity while slightly reducing recall and sensitivity. Pruning removes unnecessary branches from the tree, simplifying the model and reducing overfitting. This can lead to better generalization and potentially improved performance on unseen data.

We didn't observe a significant difference between a pruned and non-pruned decision tree in our case. It may be because:

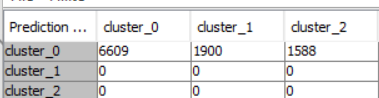
1. **Dataset characteristics:** The data we used might be relatively simple, and the decision tree without pruning may not have overfit considerably.
2. **Pruning settings:** The pruning settings in KNIME's Decision Tree Learner node might have been configured in a way that resulted in minimal removal of branches.
3. **Randomness:** There can be an element of randomness in decision tree generation. Rerunning the experiment with both pruned and non-pruned trees might yield a slight difference on another iteration.

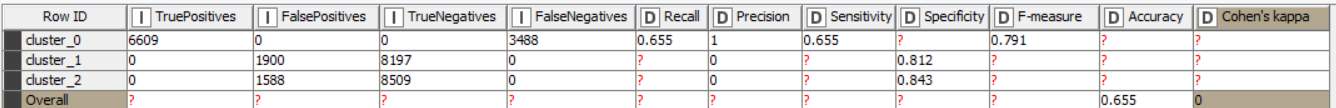
The choice of whether to prune the decision tree depends on the specific requirements of the problem and the trade-off between precision and recall. If minimizing false positives is crucial (can be used for risk assessment) pruning may be preferred. If capturing as many true positives as possible is more important (can be used for customer retention) pruning may be avoided.

**3.2.2.2. Classification Model Performance Evaluation of Other Supervised Learning methods by using confusion matrix**

**Logistic Regression**

Confusion Matrix



Accuracy statistics

Overall Accuracy: The accuracy for "Overall" is 0.655, which means the model classified 65.5% of the pizzas correctly.

Cluster\_0: It seems the model performed well for cluster\_0 with a precision of 1.0, which means all pizzas classified as cluster\_0 were actually from that cluster (no false positives). However, recall is not provided (denoted by "D"), so we can't say how many actual cluster\_0 pizzas were correctly identified.

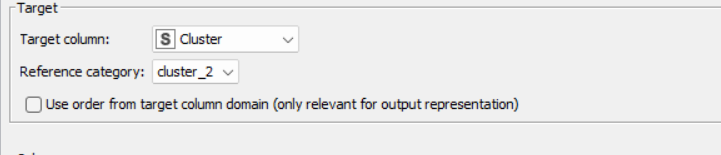
Cluster\_1 & Cluster\_2: There are no values for these clusters, possibly indicating that the model did not predict any pizzas into these clusters, or there were none in the data used to evaluate the model.

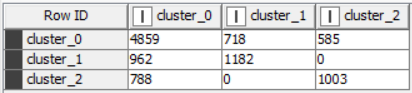
In conclusion, the model seems to have achieved good accuracy (65.5%) overall. It performed well in terms of precision for cluster\_0, but more information is needed about the clusters and recall metrics to definitively assess the model's performance for each cluster.

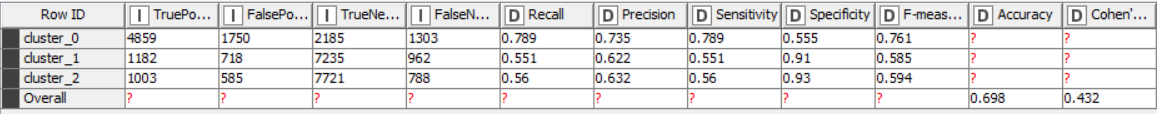
**Cluster\_2** was used as the reference category

**K-Nearest Neighbour**

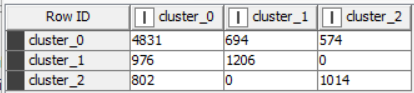
**K=7**

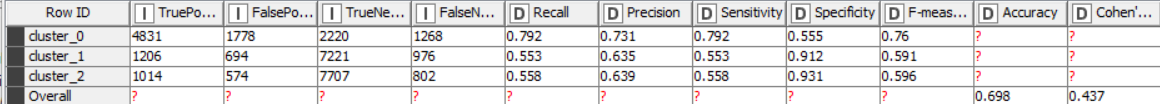




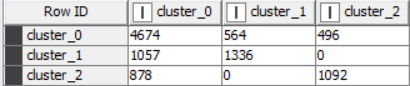


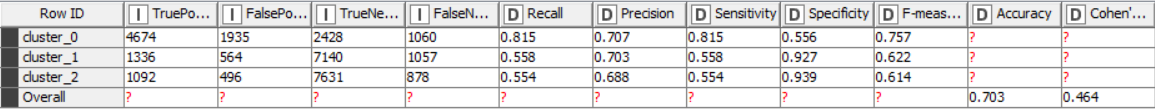
**K=9**





**K=19**



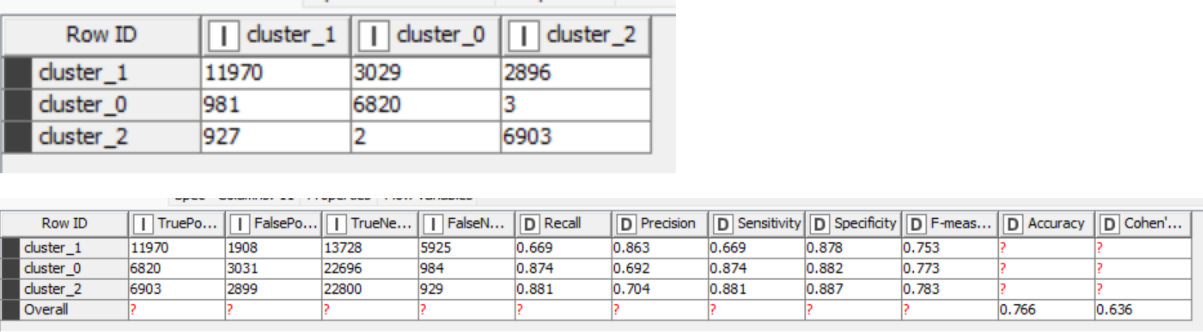


**Similarly, we have applied k nearest neighbour for K=11, 13,15,17 and observed that –**

In KNN, the number of neighbours to be considered are from k=7 to 19. From the images, it is seen that as the number of k increases the accuracy also increases. For k=19, as the accuracy is the highest from all the other k’s, this cluster will be considered.

The overall accuracy of the KNN model is moderate showing mixed performance across different clusters. Cohen's Kappa coefficient also suggests moderate agreement beyond chance among the predicted and actual cluster labels.

**Support Vector Machines**



The overall performance of the SVM model is very poor with extremely low recall, precision and accuracy metrics.

**3.2.3.1. Variable or Feature Analysis for Decision Tree**

3.2.3.1.1. List of Relevant or Important Variables.

In the decision tree analysis, we see that these were the important variables that contributed in the supervised learning algorithm which are: -

Pizza\_size, pizza\_category, Unit\_price ,Total price, Quantity, Cluster, Pizza\_ID

3.2.3.1.2. List of Non-Relevant or Non-Important Variables

Ingredients, Pizza\_name\_id

In the decision tree analysis, we see that these were the non-important variables that did not contribute in the supervised learning algorithm which are: -

Ingredients, Pizza\_name\_id, Order\_id, Order\_date, Order\_time,

**3.2.3.2. Variable or Feature Analysis for Logistic Regression, K-Nearest Neighbour and Support Vector Machine**

3.2.3.2.1. List of Relevant Variables

We have observed that Unit\_price ,Total price, Quantity, Cluster, are most significant variables in cluster 0.

3.2.3.2.2. List of Non-Important Variables

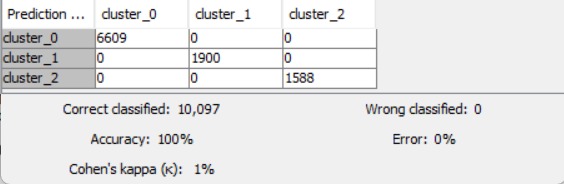
Ingredients, Pizza\_name\_id, Order\_id, Order\_date, Order\_time

The above variables have value of p>0.05 which suggests potentially negligible impact on loan outcomes.

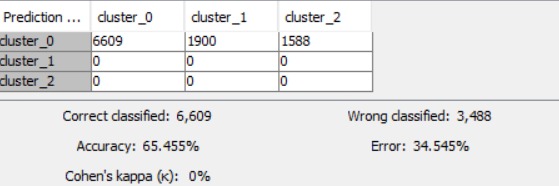
**4. Results and Observations**

4.1. Comparing Supervised Learning models: Decision Tree VS Logistic Regression, KNN and SVM

**Decision tree**

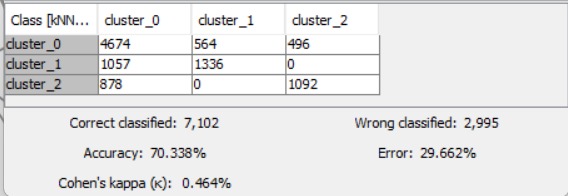


**Logistic Regression**

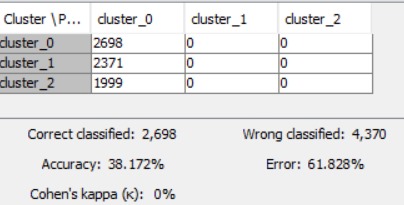


**KNN**

**K=19**



**SVM**



**4. Observations**

**4.1.** Appropriate Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metrics** | **Decision tree** | **Logistic Regression** | **Knn** | **Svm** |
| Accuracy(in%) | 100% | 65% | 70.8% | 38.172 |

The decision tree and logistic regression has the highest accuracy (100%). KNN and SVM have significantly lower accuracies of 70.8% and 38.172% respectively.

Decision tree provides the highest accuracy of all the models according to the data and will be the appropriate model for the customer classification. Decision tree is able to handle both numerical and categorical which does benefit in this data as the data contains a combination of variables which are categorical and continuous in nature.

**5. Managerial insights according to the appropriate model (Decision Tree)**

Managerial insights according to the Decision Tree model for car prices:

**Market differentiation:** The model could identify distinct customer segments based on preferences for Pizza Size (S, M, L, XL, XXL) and Pizza category. This can inform targeted marketing campaigns for each cluster.

**Pricing strategy:** The decision tree can help establish price ranges based on the clusters.

**Inventory management:** The model can help predict demand for specific Pizza types (clusters). This can guide decisions on stocking the right Pizza category and trims to meet customer preferences.