



FORE SCHOOL OF MANAGEMENT, NEW DELHI

Academic Session 2023-2025

Project-2

**TOPIC: To Determine the Segmentation of Pizza Sales at The
Store According to Various Sizes**

Machine Learning for Managers 2

FMG 32 Section C

Submitted to:

Prof. Amarnath Mitra

Submitted by:

321146 – Kshitij Tiwari

TABLE OF CONTENTS

CONTENTS	PAGE NO.
Acknowledgement	3
Objective of the project	4
Data descriptions	5
Analysis	8
Observation	18
Managerial Insights	19

ACKNOWLEDGEMENT

I express my sincere appreciation to **Professor Amarnath Mitra** for enlightening us about the captivating realm of supervised learning during this course on big data analysis. Under his tutelage, we delved deep into the intricacies of machine learning, and his mentorship, encouragement, and profound knowledge played a pivotal role in guiding us through the challenges of our project, ensuring its triumphant conclusion. Professor Mitra's unwavering dedication to nurturing our comprehension and implementation of machine learning principles has been a constant source of inspiration. I am profoundly grateful for his steadfast commitment and for bestowing upon us this invaluable learning experience.

Furthermore, Professor Mitra's passion for the subject matter was evident in every lecture and discussion, igniting our curiosity and driving us to explore the limitless possibilities of big data analysis. His ability to simplify complex concepts and provide real-world examples made the learning process both engaging and accessible. Moreover, his encouragement to think critically and creatively challenged us to push the boundaries of our knowledge and skills.

Beyond the classroom, Professor Mitra's mentorship extended to providing valuable insights and advice, nurturing our growth not only as students but also as aspiring data analysts. His willingness to invest time and effort in our development demonstrates his genuine commitment to our success.

As we move forward in our academic and professional endeavors, we carry with us the invaluable lessons and experiences gained under Professor Mitra's guidance. His impact on our education and career trajectories will undoubtedly be enduring, and for that, we are truly grateful.

1. PROJECT OBJECTIVES

- 1.1 The first objective is to segment the pizza data 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 – **8 MB**

2.1.3. Data Shape | Dimension:

Number of Variables - **10**

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

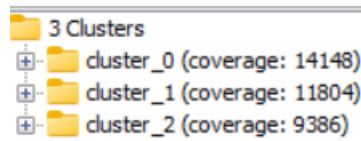
Order_time: Time at which the pizza was ordered

Order_date: date at which the pizza was ordered

2.3. Descriptive Statistics

2.3.1. Descriptive Statistics of Outcome Categorical Variables

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



3 Clusters	
+	cluster_0 (coverage: 14148)
+	cluster_1 (coverage: 11804)
+	cluster_2 (coverage: 9386)

2.3.2. Descriptive Statistics: Categorical Variables or Features

2.3.2.1. Count | Frequency Statistics

Pizza Size count

Row ID	count
L	19243
M	15971
S	14678
XL	563
XXL	28

Pizza Category

Row ID	count
Chicken	11236
Classic	15130
Supreme	12211
Veggie	11906

Pizza Name

Row ID	count
The Barbecue...	2454
The Big Meat ...	1878
The Brie Carr...	504
The Calabres...	958
The California...	2387
The Chicken ...	1010
The Chicken P...	1007
The Classic D...	2517
The Five Che...	1418
The Four Che...	1920
The Greek Pizza	1457
The Green Ga...	1028
The Hawaiian ...	2459
The Italian Ca...	1453
The Italian Su...	1920
The Italian Ve...	1014
The Mediterra...	967
The Mexicana...	1508
The Napolitan...	1501
The Pepper S...	1481
The Pepperon...	2462
The Pepperon...	1403
The Prosciutt...	1481
The Pepperon...	1403
The Prosciutt...	1481
The Sicilian Pizza	1951
The Soppress...	989
The Southwe...	1966
The Spicy Itali...	1958
The Spinach P...	1002
The Spinach S...	969
The Spinach a...	1485
The Thai Chic...	2412
The Vegetabl...	1564

2.3.3.2. Measures of Dispersion

Name	Mean	Standard Deviation	Variance	Skewness	Kurtosis
quantity	1.019	0.142	0.02	7.687	-95.495
unit_price	16.495	3.621	13.109	0.124	0.871
total_price	16.819	4.429	19.615	1.719	-12.832

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:

Name	# Missing values	50% Quantile (Median)
Cluster	22213	②

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

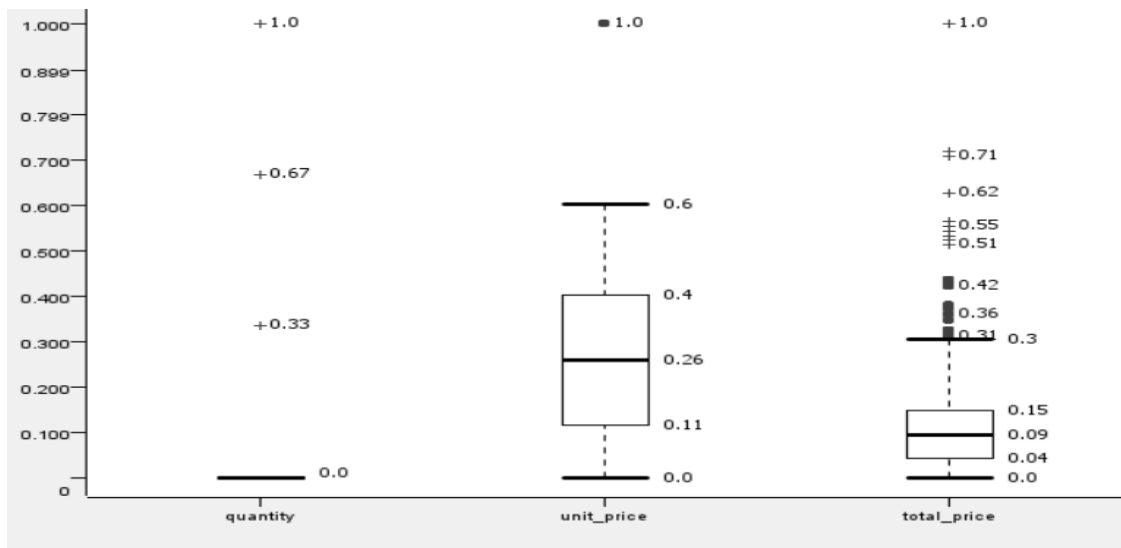
- Alphanumeric Order)

- In this case, category to number node will be used to encode the categorical variable

S	pizza_name_id
S	order_date
S	order_time
S	pizza_size
S	pizza_category
S	pizza_name
S	Cluster

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

Row ID	S Column	D Min	D Max	D Mean	D Std. de...	D Variance	D Skewness	D Kurtosis	D Overall ...	I	I	I	I	D M...	I Row count	Histogram
quantity	quantity	0	1	0.006	0.047	0.002	7.687	65.456	327.667	0	0	0	0	?	50483	
unit_price	unit_price	0	1	0.257	0.138	0.019	0.124	-0.597	12,995.813	0	0	0	0	?	50483	
total_price	total_price	0	1	0.097	0.06	0.004	1.719	8.795	4,872.124	0	0	0	0	?	50483	

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:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

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

Prediction \ Actual	cluster_0	cluster_1	cluster_2
cluster_0	6609	0	0
cluster_1	0	1900	0
cluster_2	0	0	1588

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
cluster_0	6609	0	3488	0	1	1	1	1	1	?	?
cluster_1	1900	0	8197	0	1	1	1	1	1	?	?
cluster_2	1588	0	8509	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

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

Comparing decision trees with and without pruning typically shows that pruning enhances precision and specificity, albeit at the cost of slightly reducing recall and sensitivity. Pruning entails the removal of unnecessary branches from the tree, simplifying the model and mitigating overfitting, thereby promoting better generalization and potentially improving performance on unseen data.

In our case, we did not detect a significant disparity between pruned and non-pruned decision trees. This could be attributed to several factors:

1. **Dataset Complexity:** The dataset utilized might be relatively straightforward, with the decision tree lacking substantial overfitting even without pruning.
2. **Pruning Configuration:** The pruning settings within the KNIME's Decision Tree Learner node might have been set conservatively, resulting in minimal branch removal.
3. **Random Variability:** Decision tree generation can involve stochastic elements. Repeating the experiment with both pruned and non-pruned trees might reveal slight discrepancies in another iteration.

The decision to prune the decision tree hinges on the specific needs of the problem and the balance between precision and recall. If minimizing false positives is paramount, such as in risk assessment scenarios, pruning may be preferred. Conversely, if maximizing true positives, as in customer retention strategies, is crucial, avoiding pruning may be advisable.

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

Prediction ...	cluster_0	cluster_1	cluster_2
cluster_0	6609	1900	1588
cluster_1	0	0	0
cluster_2	0	0	0

Accuracy statistics

Row ID	I TruePositives	I FalsePositives	I TrueNegatives	I FalseNegatives	D Recall	D Precision	D Sensitivity	D Specificity	D F-measure	D Accuracy	D Cohen's kappa
cluster_0	6609	0	0	3488	0.655	1	0.655	?	0.791	?	?
cluster_1	0	1900	8197	0	?	0	?	0.812	?	?	?
cluster_2	0	1588	8509	0	?	0	?	0.843	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.655	0

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

Row ID	I cluster_0	I cluster_1	I cluster_2
cluster_0	4859	718	585
cluster_1	962	1182	0
cluster_2	788	0	1003

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificty	D F-meas...	D Accuracy	D Cohen'...
cluster_0	4859	1750	2185	1303	0.789	0.735	0.789	0.555	0.761	?	?
cluster_1	1182	718	7235	962	0.551	0.622	0.551	0.91	0.585	?	?
cluster_2	1003	585	7721	788	0.56	0.632	0.56	0.93	0.594	?	?
Overall	?	?	?	?	?	?	?	?	?	0.698	0.432

K=9

Row ID	I cluster_0	I cluster_1	I cluster_2
cluster_0	4831	694	574
cluster_1	976	1206	0
cluster_2	802	0	1014

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificty	D F-meas...	D Accuracy	D Cohen'...
cluster_0	4831	1778	2220	1268	0.792	0.731	0.792	0.555	0.76	?	?
cluster_1	1206	694	7221	976	0.553	0.635	0.553	0.912	0.591	?	?
cluster_2	1014	574	7707	802	0.558	0.639	0.558	0.931	0.596	?	?
Overall	?	?	?	?	?	?	?	?	?	0.698	0.437

K=19

Row ID	cluster_0	cluster_1	cluster_2
cluster_0	4674	564	496
cluster_1	1057	1336	0
cluster_2	878	0	1092


Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_0	4674	1935	2428	1060	0.815	0.707	0.815	0.556	0.757	?	?
cluster_1	1336	564	7140	1057	0.558	0.703	0.558	0.927	0.622	?	?
cluster_2	1092	496	7631	878	0.554	0.688	0.554	0.939	0.614	?	?
Overall	?	?	?	?	?	?	?	?	?	0.703	0.464

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

 Confusion Matrix - 6:88 - Scorer

File Hilite

Cluster \ P...	cluster_0	cluster_1	cluster_2
cluster_0	2698	0	0
cluster_1	2371	0	0
cluster_2	1999	0	0

 Accuracy statistics - 6:88 - Scorer

File Edit Hilite Navigation View

Table "default" - Rows: 4 Spec - Columns: 11 Properties Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_0	2698	4370	0	0	1	0.382	1	0	0.553	?	?
cluster_1	0	0	4697	2371	0	?	0	1	?	?	?
cluster_2	0	0	5069	1999	0	?	0	1	?	?	?
Overall	?	?	?	?	?	?	?	?	?	0.382	0

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, pizza_name, order_id, total_price, unit_price

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

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: -

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 state, color, interior and transmission are most significant variables in cluster 0.

3.2.3.2.2. List of Non-Important Variables

Car_id, odometer, vin, saledate, condition and trim are insignificant variables.

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

File	Hilite			
Clusters \ ...	cluster_1	cluster_0	cluster_2	
cluster_1	17895	0	0	
cluster_0	0	7804	0	
cluster_2	0	0	7832	
Correct classified: 33,531				
Wrong classified: 0				
Accuracy: 100%				
Error: 0%				
Cohen's kappa (κ): 1%				

Logistic Regression

Clusters \ ...	cluster_1	cluster_0	cluster_2	
cluster_1	11347	3337	3211	
cluster_0	274	7528	2	
cluster_2	303	3	7526	
Correct classified: 26,401				
Wrong classified: 7,130				
Accuracy: 78.736%				
Error: 21.264%				
Cohen's kappa (κ): 0.678%				

KNN=19

Clusters \ ...	cluster_1	cluster_0	cluster_2	
cluster_1	11930	0	0	
cluster_0	0	5202	0	
cluster_2	0	0	5222	
Correct classified: 22,354				
Wrong classified: 0				
Accuracy: 100%				
Error: 0%				
Cohen's kappa (κ): 1%				

SVM

Clusters \ ...	cluster_1	cluster_0	cluster_2	
cluster_1	11970	3029	2896	
cluster_0	981	6820	3	
cluster_2	927	2	6903	
Correct classified: 25,693				
Wrong classified: 7,838				
Accuracy: 76.625%				
Error: 23.375%				
Cohen's kappa (κ): 0.636%				

5. MANAGERIAL INSIGHTS

5.1. Appropriate Model

METRICS	Decision tree	Logistic Regression	KNN	KNN
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 78.74% and 76.63% 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.

MANAGERIAL IMPLICATIONS DERIVED FROM THE SUITABLE MODEL (DECISION TREE)

Managerial implications drawn from the Decision Tree model for car prices:

Market segmentation: The model has the capability to recognize unique customer segments distinguished by preferences for Pizza Size (S, M, L, XL, XXL) and Pizza category. This information can guide tailored marketing initiatives aimed at each cluster.

Price determination: Utilizing the decision tree, price boundaries can be established based on the identified clusters.

Inventory optimization: Leveraging the model's predictive abilities, demand for specific Pizza types (clusters) can be anticipated. This insight can inform decisions regarding inventory management, ensuring the stocking of appropriate Pizza categories and varieties to align with customer preference

