FORE SCHOOL OF MANAGEMENT

# Project Title: Classification of Consumer Data into Segments | Clusters | Classes

**Submitted by:**

**Naman Agarawal**

**(064030)**

**Submitted to:**

**Prof. Amarnath Mitra**

## Project Contents

## 1. Project Objectives | Problem Statements

1.1. PO1 | PS1: Classification of Consumer Data into Segments | Clusters | Classes using Supervised Learning Classification Algorithms

1.2. PO2 | PS2: Determination of an Appropriate Classification Model

1.3. PO3 | PS3: Identification of Important | Contributing | Significant Variables or Features and their Thresholds for Classification

## 2. Description of Data

### 2.1. Data Source, Size, Shape

2.1.1. Data Source (Website Link) https://www.kaggle.com

2.1.2. Data Size (in KB | MB | GB …) - 5.75 MB

2.1.3.1 Data Shape | 51289*17

2.1.3.2 Dimension: Number of Variables - 17 | Number of Records - 51289

### 2.2. Description of Variables

2.2.1. Index Variable(s): I1 - Order Date

I2 – Time

I3 – Customer Id

2.2.2. Outcome Variable or Feature: Cluster no.

2.2.3. Input Variables or Features having Categories | Input Categorical Variables or Features (ICV)

2.2.3.1. Input Variables or Features having Nominal Categories | Categorical Variables or Features - Nominal Type:

ICNV1 – Gender - Gender of customer.

ICNV2 – Device Type - The device the customer uses to actualize the transaction (Web/Mobile).

ICNV3 – Product Category - Product category

ICNV4 – Product – Product

ICNV5 – Payment Method - Payment method

ICNV6 – Customer login type - The type the customer logged in. Such as Member, Guest etc.

ICNV7 – Cluster id – The cluster which the record is belongs to ( from Project 1 )

2.2.3.2. Input Variables or Features having Ordinal Categories | Categorical Variables or Features - Ordinal Type:

ICOV1 – Order Priority - Order priority. Such as critical, high etc.

2.2.3. Input Non-Categorical Variables or Features:

INCV1 – Aging - The time from the day the product is ordered to the day it is delivered.

INCV2 – Sales - Total sales amount

INCV3 – Quantity - Unit amount of product

INCV4 – Discount - Percent discount rate

INCV5 – Profit - Profit

INCV6 – Shipping Cost - Shipping cost

## 2.3. Descriptive Statistics

2.3.1. Descriptive Statistics: Outcome Variable or Feature (Categorical)

2.3.1.1. Count | Frequency Statistics –

Cluster 0 – 13150

Cluster1 – 13095

Cluster2 – 3322

Cluster3 – 9046

Cluster4 - 12677

2.3.1.2. Proportion (Relative Frequency) Statistics – 5.89%

2.3.2. Descriptive Statistics: Input Categorical Variables or Features

2.3.2.1. Count | Frequency Statistics

**Gender** | Male – 28138 | female – 23152 |

**Device Type** | Mobile – 3658 | Web – 47632 |

**Product Category** | Auto & Accessories – 7505 | Electronic – 2701 | Fashion – 25646 | Home & Furniture – 15438 |

**Product** | Apple Laptop - 221 | Bedsheets - 1541 | Beds - 1542 | Bike tyres - 826 | Car & Bike Cares - 826 | Car Body Covers – 826 | Car Mat – 826 | Car Media Players - 826 | Car Pillow & Neck Rest - 829 | Car Seat Covers – 827 | Car Speakers - 826 | Casula Shoes - 2331 | Curtains - 1541 | Dinner Crockery - 1566 | Dinning Tables – 1542 | Fans - 221 | Formal Shoes - 2331 | Fossil Watch - 2332 | Iron - 221 | Jeans - 2332 | Keyboard - 221 | LCD - 224 | LED - 224 | Mixer/Juicer - 224 | Mouse - 221 | Running Shoes - 2331 | Samsung Mobile - 221 | Shirts - 2332 | Shoe Racks - 1542 | Sneakers - 2331 | Sofa Covers - 1539 | Sofas - 1542 | speakers – 261 | Sports Wear - 2331 | Suits – 2332 | T-shirts - 2332 | Tablet - 221 | Titak Watch - 2331 | Towels – 1541 | Tyre - 893 | Umbrella - 1542 | Watch - 221 |

**Payment Method |** Credit Card - 38137 | Debit card - 734 | e-wallet - 2789 | Money Order - 9629 | Not Defined – 1 |

**Customer Login Type - |** First Sign up – 173 | Guest – 1993 | Member - 49097 | New – 27 |

**Cluster Id - |** Cluster 0 – 13150 | Cluster1 – 13095 | Cluster2 – 3322 | Cluster3 – 9046 | Cluster4 - 12677

2.3.2.2. Proportion (Relative Frequency) Statistics – 47.05%

2.3.3. Descriptive Statistics: Input Non-Categorical Variables or Features

2.3.3.1. Measures of Central Tendency –

NCV1 – Aging | mean - 5.255 | mode – 1.0 | median - 5 |

NCV2 – Sales | mean - 152.341 | mode – 228 | median – 133 |

NCV3 – Quantity | mean - 2.503 | mode – 1.0 | median - 2 |

NCV4 – Discount | mean - 0.304 | mode – 0.3 | median - 0.3 |

NCV5 – Profit | mean - 70.407 | mode – 17.0 | median - 59.9 |

NCV6 – Shipping Cost | mean - 7.042 | mode – 1.7 | median - 6 |

2.3.3.2. Measures of Dispersion

NCV1 – Aging | std dev - 2.9599 | variance - 8.761 | skewness - 0.0656 |

NCV2 – Sales | std dev - 66.4954 | variance - 4421.641| skewness - -0.0878 |

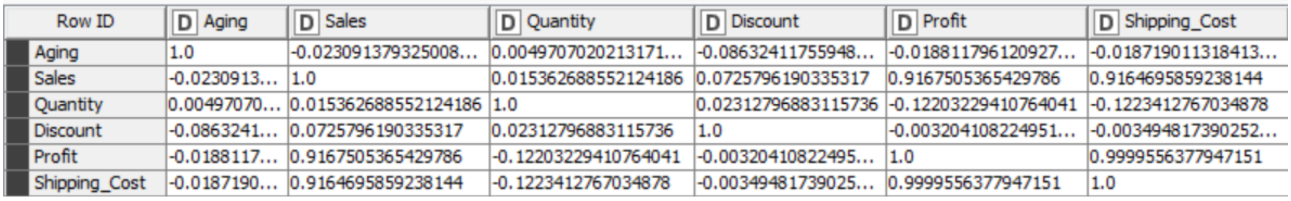
NCV3 – Quantity | std dev - 1.5119 | variance - 2.286 | skewness - 0.4642 |

NCV4 – Discount | std dev - 0.131 | variance - 0.017 | skewness - 0.0332 |

NCV5 – Profit | std dev - 48.7295 | variance - 2374.563 | skewness - 0.261 |

NCV6 – Shipping Cost | std dev - 4.8717 | variance - 23.734 | skewness - 0.2625 |

2.3.3.3. Correlation Statistics (with Test of Correlation)



| Row ID | D Aging | D Sales | D Quantity | D Discount | D Profit | D Shipping_Cost |
|---|---|---|---|---|---|---|
| Aging | 1.0 | -0.023091379325008... | 0.0049707020213171... | -0.08632411755948... | -0.018811796120927... | -0.018719011318413... |
| Sales | -0.0230913... | 1.0 | 0.015362688552124186 | 0.0725796190335317 | 0.9167505365429786 | 0.9164695859238144 |
| Quantity | 0.00497070... | 0.015362688552124186 | 1.0 | 0.02312796883115736 | -0.12203229410764041 | -0.1223412767034878 |
| Discount | -0.0863241... | 0.0725796190335317 | 0.02312796883115736 | 1.0 | -0.003204108224951... | -0.003494817390252... |
| Profit | -0.0188117... | 0.9167505365429786 | -0.12203229410764041 | -0.00320410822495... | 1.0 | 0.9999556377947151 |
| Shipping_Cost | -0.0187190... | 0.9164695859238144 | -0.1223412767034878 | -0.00349481739025... | 0.9999556377947151 | 1.0 |

## 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: Records = 6

3.1.1.1.2. Missing Data Treatment: Records

3.1.1.1.2.1. Removal of Records with More Than 50% Missing Data: None |

3.1.1.2.1. Missing Data Statistics: Categorical Variables or Features | none |

3.1.1.2.2. Missing Data Treatment: Categorical Variables or Features

3.1.1.2.2.1. Removal of Variables or Features with More Than 50% Missing Data: None |

3.1.1.2.2.2. Imputation of Missing Data using Descriptive Statistics: Mode

3.1.1.3.1. Missing Data Statistics: Non-Categorical Variables or Features

Aging – 1 | Sales – 1 | Quantity – 2 | Discount – 1 | Shipping Cost – 1 |

3.1.1.3.2. Missing Data Treatment: Non-Categorical Variables or Features

3.1.1.3.2.1. Removal of Variables or Features with More Than 50% Missing Data: None |

3.1.1.3.2.2. Imputation of Missing Data using Descriptive Statistics: Mean |

**3.1.2. Numerical Encoding of Categorical Variables or Features** (Encoding Schema - Alphanumeric Order)

**CV1 – Gender** | Female – 0 | Male – 1 |

**CV2 – Device Type** | web – 0 | Mobile – 1 |

**CV3 – Product Category** | Auto & Accessories – 0 | Fashion – 1 | Electronic – 2 | Home & Furniture – 3 |

**CV4 – Product** | Car Media Players - 0 | Car Speakers - 1 | Car Body Covers - 2 | Car & Bike Care - 3 | Tyre - 4 | Bike Tyres - 5 | Car Mat - 6 | Car Seat Covers - 7 | Car Pillow & Neck Rests - 8 | Shirts - 9 | jeans - 10 | Suits - 11 | Sports Wear - 12 | Casula shoes - 13 | Running Shoes - 14 | Formal Shoes - 15 | Sneakers - 16 | Tilak Watch - 17 | Fossil Watch - 18 | T-shirt - 19 | Samsung Mobile - 20 | Watch - 21 | Fans - 22 | Iron - 23 | Tablet - 24 | Mouse - 25 | Keyboard - 26 | Apple Laptop

- 27 | Mixer/Juicer - 28 | LED - 29 | LCD - 30 | Speakers - 31 - | Sofa Covers - 32 | Bed Sheets - 33 | Curtains - 34 | Towels - 35 | Sofas - 36 | Beds - 37 | Dining Tables - 38 | Shoe rack - 39 | Umbrellas - 40 | Dining Crockery - 41 |

**CV5 – Payment Method** | Credit card – 0 | Money Order – 1 | E-Wallet – 2 | Debit Card – 3 | Not Defined – 4 |

**CV6 – Customer login type** | Member – 0 | Guest – 1 | New – 2 | First Signup – 3 |

**CV7 – Order Priority** | Medium – 0 | Critical – 1 | High – 2 | Low – 3 |

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

3.1.3.1.1. Outlier Statistics: Non-Categorical Variables or Features | There is no Outlier present in any non – Categorical variables |

3.1.3.1.2. Outlier Treatment: Non-Categorical Variables or Features

3.1.3.1.2.1. Normalization using z score Normalization

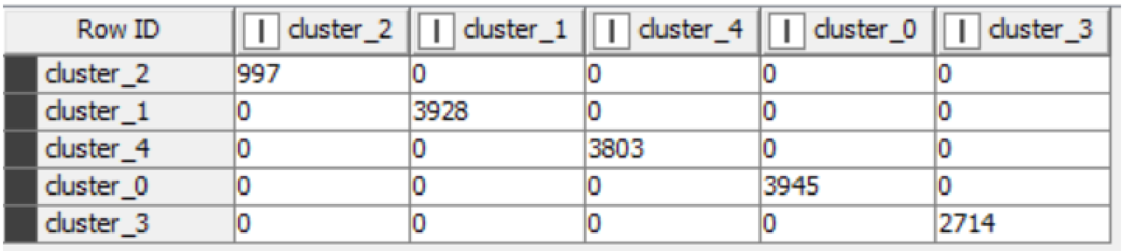
**3.1.4. Data Bifurcation: Training & Testing Sets** [Bifurcation Schema: Stratified Sampling (Based on Outcome Variable or Feature) with 70% Data in Training Set and 30% Data in Testing Set

**3.2. Data Analysis**

3.2.1.1. PO1 | PS1:: Supervised Machine Learning Classification Algorithm: Decision Tree (Base Model) | Metrics Used - Gini Coefficient, Entropy

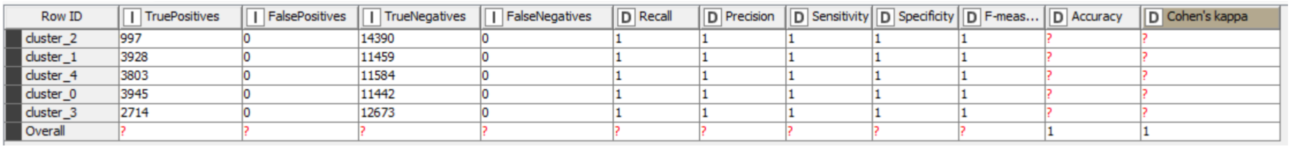
3.2.1.2. PO1 | PS1:: Supervised Machine Learning Classification Algorithms: {Logistic Regression | Support Vector Machine | K Nearest Neighbour} (Comparison Models) | Metrics Used

3.2.2.1.1. PO2 | PS2:: Classification Model Performance Evaluation: Base Model: **Decision Tree**| Confusion Matrix



| Row ID | cluster_2 | cluster_1 | cluster_4 | cluster_0 | cluster_3 |
|---|---|---|---|---|---|
| cluster_2 | 997 | 0 | 0 | 0 | 0 |
| cluster_1 | 0 | 3928 | 0 | 0 | 0 |
| cluster_4 | 0 | 0 | 3803 | 0 | 0 |
| cluster_0 | 0 | 0 | 0 | 3945 | 0 |
| cluster_3 | 0 | 0 | 0 | 0 | 2714 |

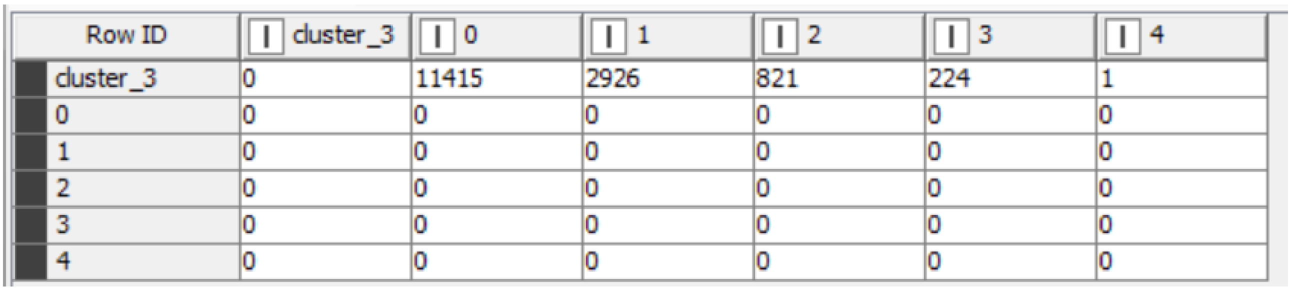
{Accuracy, Recall, Precision, F1-Score}



| Row ID | TruePositives | FalsePositives | TrueNegatives | FalseNegatives | Recall | Precision | Sensitivity | Specificity | F-meas... | Accuracy | Cohen's kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cluster_2 | 997 | 0 | 14390 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| cluster_1 | 3928 | 0 | 11459 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| cluster_4 | 3803 | 0 | 11584 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| cluster_0 | 3945 | 0 | 11442 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| cluster_3 | 2714 | 0 | 12673 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 1 | 1 |

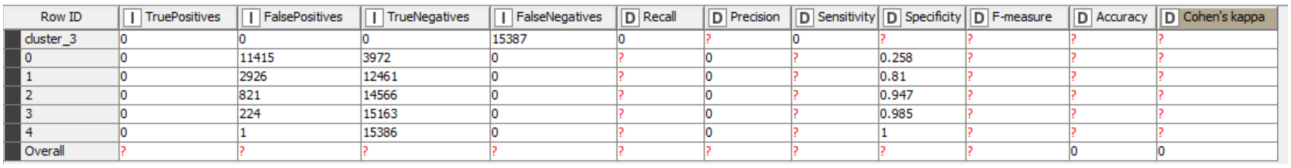
### 3.2.2.2.1. PO2 | PS2:: Classification Model Performance Evaluation:

**Logistic Regression - | Confusion Matrix**



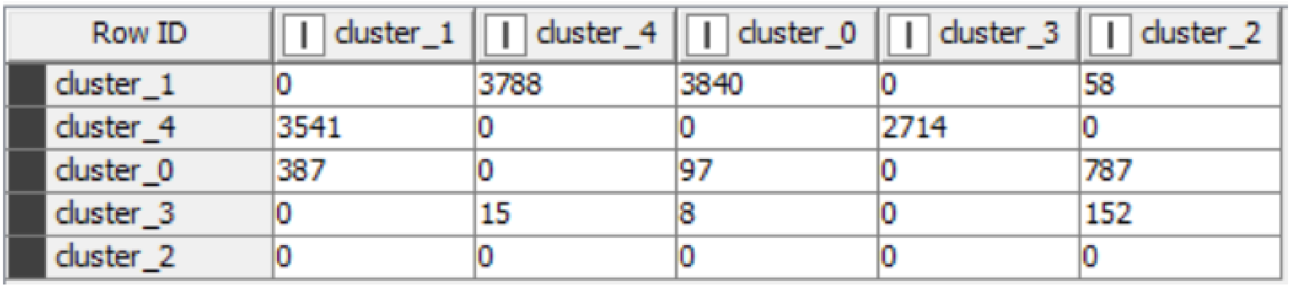
| Row ID | cluster_3 | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| cluster_3 | 0 | 11415 | 2926 | 821 | 224 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 |

{Accuracy, Recall, Precision, F1-Score}



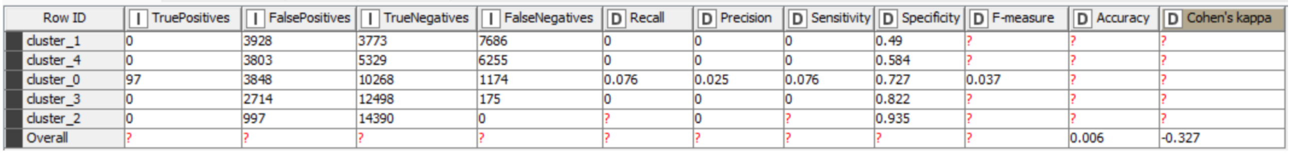
| Row ID | TruePositives | FalsePositives | TrueNegatives | FalseNegatives | Recall | Precision | Sensitivity | Specificity | F-measure | Accuracy | Cohen's kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cluster_3 | 0 | 0 | 0 | 15387 | 0 | ? | 0 | ? | ? | ? | ? |
| 0 | 0 | 11415 | 3972 | 0 | ? | 0 | ? | 0.258 | ? | ? | ? |
| 1 | 0 | 2926 | 12461 | 0 | ? | 0 | ? | 0.81 | ? | ? | ? |
| 2 | 0 | 821 | 14566 | 0 | ? | 0 | ? | 0.947 | ? | ? | ? |
| 3 | 0 | 224 | 15163 | 0 | ? | 0 | ? | 0.985 | ? | ? | ? |
| 4 | 0 | 1 | 15386 | 0 | ? | 0 | ? | 1 | ? | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 0 | 0 |

**Support Vector Machine - | Confusion Matrix**



| Row ID | cluster_1 | cluster_4 | cluster_0 | cluster_3 | cluster_2 |
|---|---|---|---|---|---|
| cluster_1 | 0 | 3788 | 3840 | 0 | 58 |
| cluster_4 | 3541 | 0 | 0 | 2714 | 0 |
| cluster_0 | 387 | 0 | 97 | 0 | 787 |
| cluster_3 | 0 | 15 | 8 | 0 | 152 |
| cluster_2 | 0 | 0 | 0 | 0 | 0 |

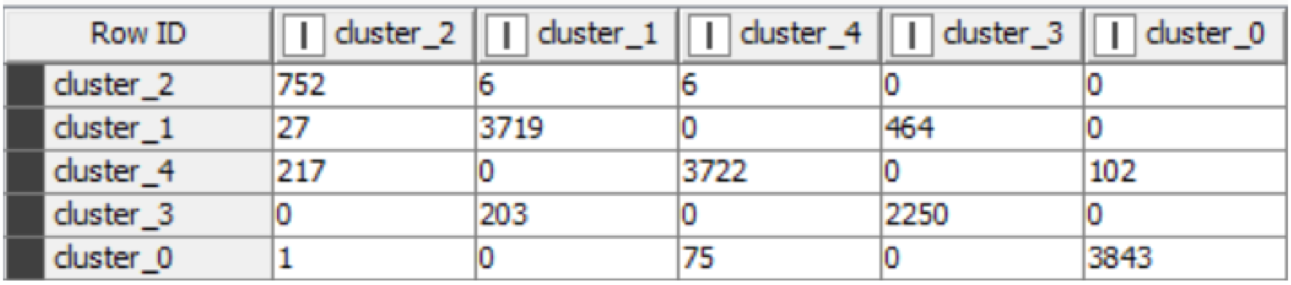
{Accuracy, Recall, Precision, F1-Score}



| 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_1 | 0 | 3928 | 3773 | 7686 | 0 | 0 | 0 | 0.49 | ? | ? | ? |
| cluster_4 | 0 | 3803 | 5329 | 6255 | 0 | 0 | 0 | 0.584 | ? | ? | ? |
| cluster_0 | 97 | 3848 | 10268 | 1174 | 0.076 | 0.025 | 0.076 | 0.727 | 0.037 | ? | ? |
| cluster_3 | 0 | 2714 | 12498 | 175 | 0 | 0 | 0 | 0.822 | ? | ? | ? |
| cluster_2 | 0 | 997 | 14390 | 0 | ? | 0 | ? | 0.935 | ? | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 0.006 | -0.327 |

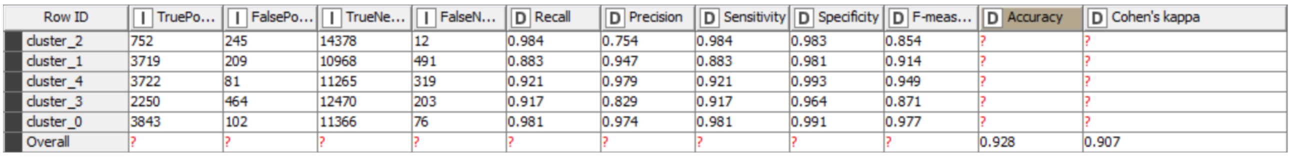
### K nearest Neighbour – Number of neighbours to consider – 11 |

Confusion Matrix



| Row ID | I cluster_2 | I cluster_1 | I cluster_4 | I cluster_3 | I cluster_0 |
|---|---|---|---|---|---|
| cluster_2 | 752 | 6 | 6 | 0 | 0 |
| cluster_1 | 27 | 3719 | 0 | 464 | 0 |
| cluster_4 | 217 | 0 | 3722 | 0 | 102 |
| cluster_3 | 0 | 203 | 0 | 2250 | 0 |
| cluster_0 | 1 | 0 | 75 | 0 | 3843 |

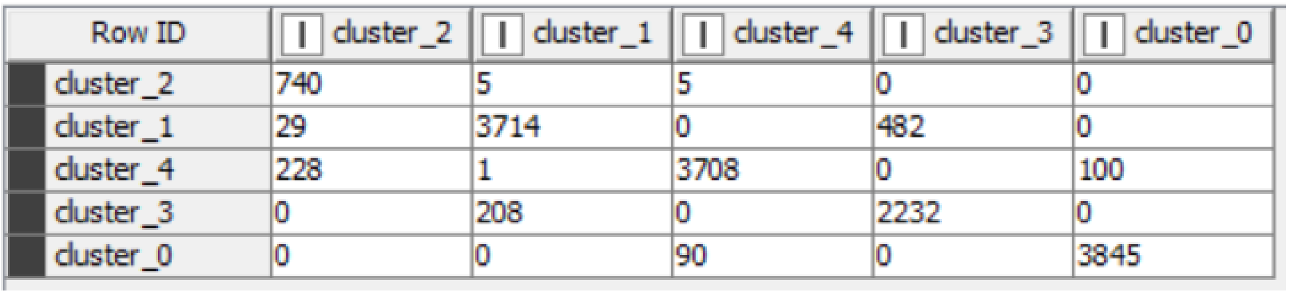
{Accuracy, Recall, Precision, F1-Score}



| 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's kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cluster_2 | 752 | 245 | 14378 | 12 | 0.984 | 0.754 | 0.984 | 0.983 | 0.854 | ? | ? |
| cluster_1 | 3719 | 209 | 10968 | 491 | 0.883 | 0.947 | 0.883 | 0.981 | 0.914 | ? | ? |
| cluster_4 | 3722 | 81 | 11265 | 319 | 0.921 | 0.979 | 0.921 | 0.993 | 0.949 | ? | ? |
| cluster_3 | 2250 | 464 | 12470 | 203 | 0.917 | 0.829 | 0.917 | 0.964 | 0.871 | ? | ? |
| cluster_0 | 3843 | 102 | 11366 | 76 | 0.981 | 0.974 | 0.981 | 0.991 | 0.977 | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 0.928 | 0.907 |

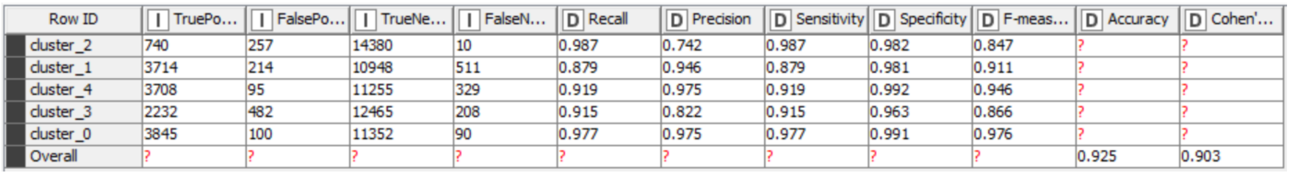
### K nearest Neighbour – Number of neighbours to consider – 13 |

Confusion Matrix



| Row ID | I cluster_2 | I cluster_1 | I cluster_4 | I cluster_3 | I cluster_0 |
|---|---|---|---|---|---|
| cluster_2 | 740 | 5 | 5 | 0 | 0 |
| cluster_1 | 29 | 3714 | 0 | 482 | 0 |
| cluster_4 | 228 | 1 | 3708 | 0 | 100 |
| cluster_3 | 0 | 208 | 0 | 2232 | 0 |
| cluster_0 | 0 | 0 | 90 | 0 | 3845 |

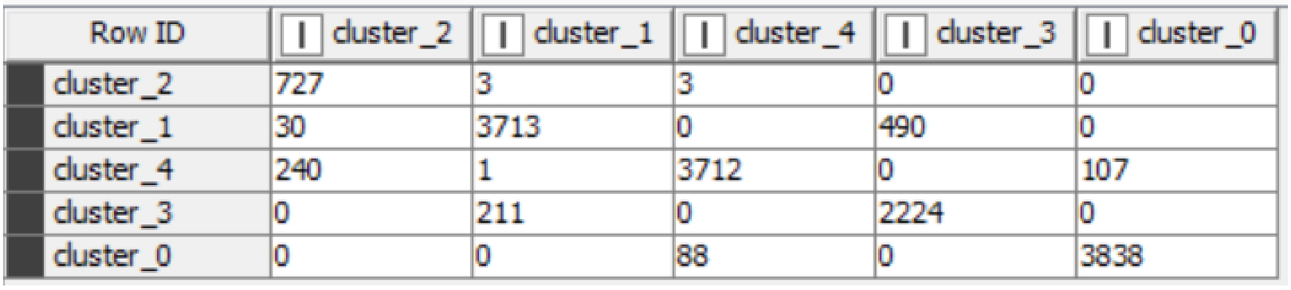
{Accuracy, Recall, Precision, F1-Score}



| 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_2 | 740 | 257 | 14380 | 10 | 0.987 | 0.742 | 0.987 | 0.982 | 0.847 | ? | ? |
| cluster_1 | 3714 | 214 | 10948 | 511 | 0.879 | 0.946 | 0.879 | 0.981 | 0.911 | ? | ? |
| cluster_4 | 3708 | 95 | 11255 | 329 | 0.919 | 0.975 | 0.919 | 0.992 | 0.946 | ? | ? |
| cluster_3 | 2232 | 482 | 12465 | 208 | 0.915 | 0.822 | 0.915 | 0.963 | 0.866 | ? | ? |
| cluster_0 | 3845 | 100 | 11352 | 90 | 0.977 | 0.975 | 0.977 | 0.991 | 0.976 | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 0.925 | 0.903 |

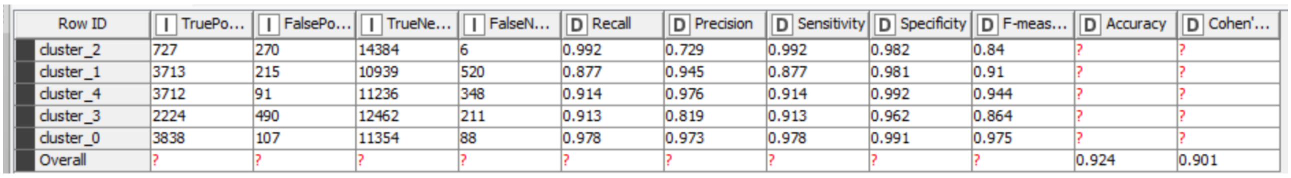
**K nearest Neighbour – Number of neighbours to consider – 15 |**

Confusion Matrix



| Row ID | I cluster_2 | I cluster_1 | I cluster_4 | I cluster_3 | I cluster_0 |
|---|---|---|---|---|---|
| cluster_2 | 727 | 3 | 3 | 0 | 0 |
| cluster_1 | 30 | 3713 | 0 | 490 | 0 |
| cluster_4 | 240 | 1 | 3712 | 0 | 107 |
| cluster_3 | 0 | 211 | 0 | 2224 | 0 |
| cluster_0 | 0 | 0 | 88 | 0 | 3838 |

{Accuracy, Recall, Precision, F1-Score}



| 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_2 | 727 | 270 | 14384 | 6 | 0.992 | 0.729 | 0.992 | 0.982 | 0.84 | ? | ? |
| cluster_1 | 3713 | 215 | 10939 | 520 | 0.877 | 0.945 | 0.877 | 0.981 | 0.91 | ? | ? |
| cluster_4 | 3712 | 91 | 11236 | 348 | 0.914 | 0.976 | 0.914 | 0.992 | 0.944 | ? | ? |
| cluster_3 | 2224 | 490 | 12462 | 211 | 0.913 | 0.819 | 0.913 | 0.962 | 0.864 | ? | ? |
| cluster_0 | 3838 | 107 | 11354 | 88 | 0.978 | 0.973 | 0.978 | 0.991 | 0.975 | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 0.924 | 0.901 |

3.2.3.1. PO3 | PS3:: Variable or Feature Analysis: Base Model (Decision Tree)

3.2.3.1.1. List of Relevant or Important Variables or Features and their Thresholds |

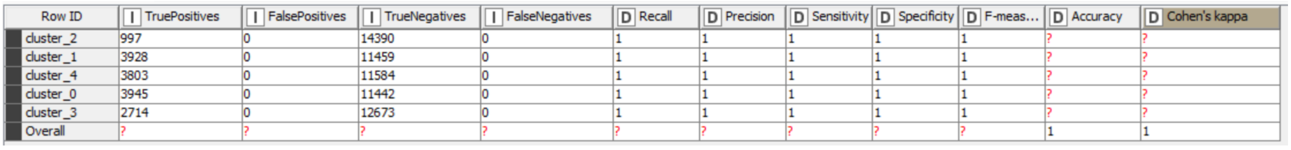
Sales 94.5, 136.5, 197.5, 226, 175.5, 221, 239.5, 217 | Shipping Cost 7.65, 12.55, 12.2, 10.7, 13.15 | Profit 128.95, 39.1, |

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

Gender | Aging | Product | Device Type | Product Category | Payment Method | Customer Login type | Quantity | Discount

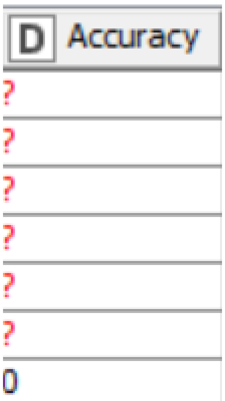
## 4. Results | Observations

4.1. Classification Model Parameters: Accuracy of the model | Base Model (Decision Tree)



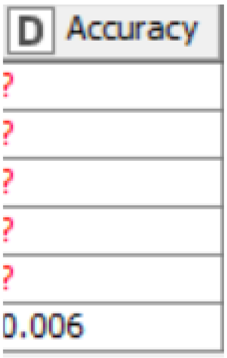
| Row ID | I TruePositives | I FalsePositives | I TrueNegatives | I FalseNegatives | D Recall | D Precision | D Sensitivity | D Specificity | D F-meas... | D Accuracy | D Cohen's kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|
| cluster_2 | 997 | 0 | 14390 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| cluster_1 | 3928 | 0 | 11459 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| cluster_4 | 3803 | 0 | 11584 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| cluster_0 | 3945 | 0 | 11442 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| cluster_3 | 2714 | 0 | 12673 | 0 | 1 | 1 | 1 | 1 | 1 | ? | ? |
| Overall | ? | ? | ? | ? | ? | ? | ? | ? | ? | 1 | 1 |

Logistic Regression |



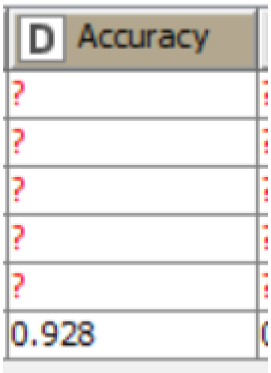
| D Accuracy |
|---|
| ? |
| ? |
| ? |
| ? |
| ? |
| ? |
| 0 |

Support Vector Machine |



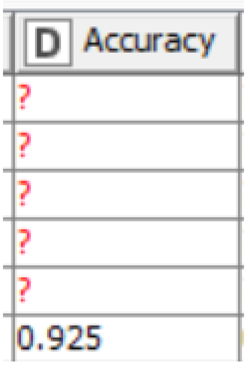
| D Accuracy |
|---|
| ? |
| ? |
| ? |
| ? |
| ? |
| 0.006 |

K Nearest Neighbour (11)



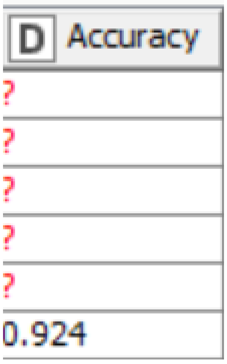
| D Accuracy |
|---|
| ? |
| ? |
| ? |
| ? |
| ? |
| 0.928 |

K Nearest Neighbour (13)



| D Accuracy |
|---|
| ? |
| ? |
| ? |
| ? |
| ? |
| 0.925 |

K Nearest Neighbour (15)



| D Accuracy |
|---|
| ? |
| ? |
| ? |
| ? |
| ? |
| 0.924 |

## 5. Managerial Insights

5.1. Appropriate Model – Decision Tree

It gives 100% accuracy for the given dataset. However the next best option would be the K nearest Neighbour with number of neighbours to be consider would be 11.

5.2. Relevant or Important Variables or Features

Of the Decision Tree Model, I observe that only Shipping Cost, Profit and sales is most important among all other variables to predict the cluster class.

Which means only these three i.e. Shipping Cost, Profit and Sales have large Impact for the outcome variable, however if I change the setting of partitioning from 70% to 75% it may change the Relevant and Important variables.