Report On

**UNSUPERVISED CLASSIFICATION ML MODEL**

**CUSTOMER SEGMENTATION DATA SET**

**Table of Contents**

[**Introduction** 3](#_Toc59401420)

[**Importing libraries** 3](#_Toc59401421)

[**Data Pre-processing** 3](#_Toc59401422)

[**Loading Dataset** 4](#_Toc59401423)

[**Missing value Handling** 4](#_Toc59401424)

[**Feature Selection** 4](#_Toc59401425)

[**Outlier detection and removal** 5](#_Toc59401426)

[**Unsupervised Learning using K-means clustering** 5](#_Toc59401427)

[**Supervised Learning** 6](#_Toc59401428)

[**Train Test split** 6](#_Toc59401429)

[**Feature Scaling** 6](#_Toc59401430)

[**Classification using Naive Bayes** 6](#_Toc59401431)

[**Classification using SVM classifier** 7](#_Toc59401432)

[**Feature Transformation – PCA** 8](#_Toc59401433)

[**K-means clustering with PCA** 9](#_Toc59401434)

[**Naïve Bayes classification with PCA** 9](#_Toc59401435)

[**SVM classifier with PCA** 9](#_Toc59401436)

[**Results** 10](#_Toc59401437)

[**Conclusion** 10](#_Toc59401438)

[**References** 11](#_Toc59401439)

# **Introduction**

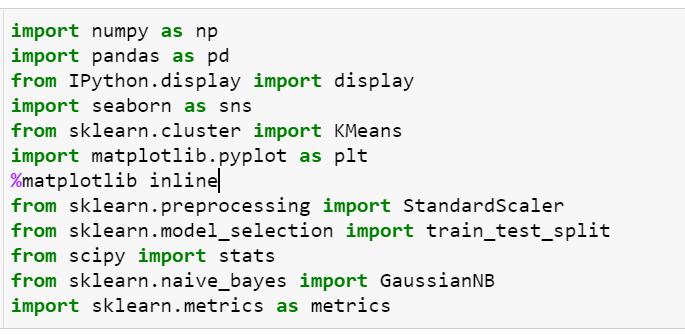
Customer Segmentation identify customer needs and helps Segmenting your customers can be a powerful means to focusing marketing efforts, so that businesses can increase profits and overall customer satisfaction. This is done by aggregating customers with similar features into one category. The dataset is ‘Wholesale customers Data Set’ obtained from UCI Dataset Repository and is created by Margarida G. M. S. Cardoso. Wholesale Customer dataset contains data about clients of a wholesale distributor and our objective in this project is to segment different customers based on their spending on different caregories. The following are the major characteristics of this dataset.

* Dataset Characteristics: Multivariate
* Attribute Characteristics: Integer
* Number of Instances: 440
* Number of Attributes: 8
* Missing Attribute Values: None

Both supervised and unsupervised learning algorithms are used in the evaluation of this dataset, where classification methods are Gaussian Naïve Bayes and SVM classification and clustering method of K-means clustering.

# **Importing libraries**

To utilize the functions in a specific module, importing the corresponding libraries are necessary. The python libraries used in this model is numpy, pandas, matplotlib, seaborn, scipy and scikitlearn.

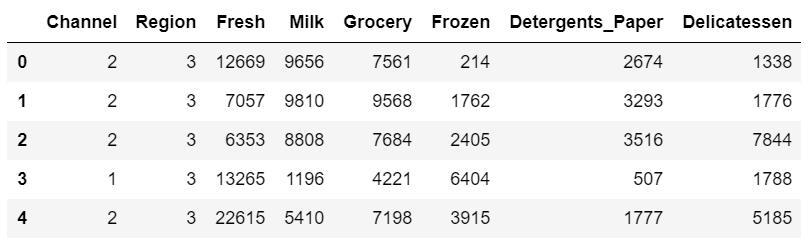


# **Data Pre-processing**

The most important process in machine learning pipeline and involves the extraction of useful data from raw data, which includes missing value handling and formatting of our dataset if necessary. Data preparation can improve the accuracy of our model significantly. Data preparation comprises of loading dataset, missing value handling, feature selection, feature scaling and dimensionality reduction.

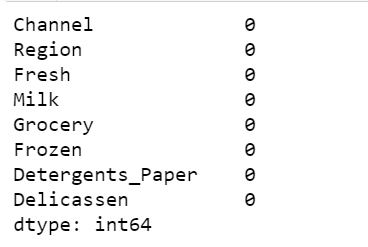
## **Loading Dataset**

After importing libraries, we need to load data using read\_csv function of pandas. The dataset consists of 8 features and 440 instances.



## **Missing value Handling**

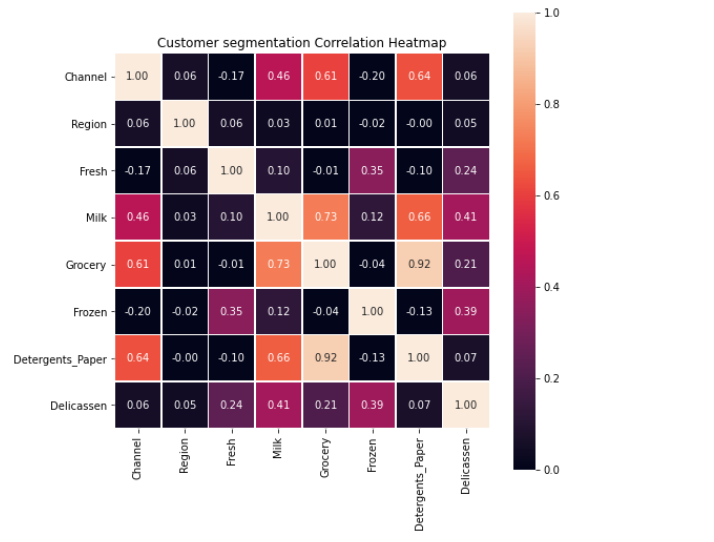
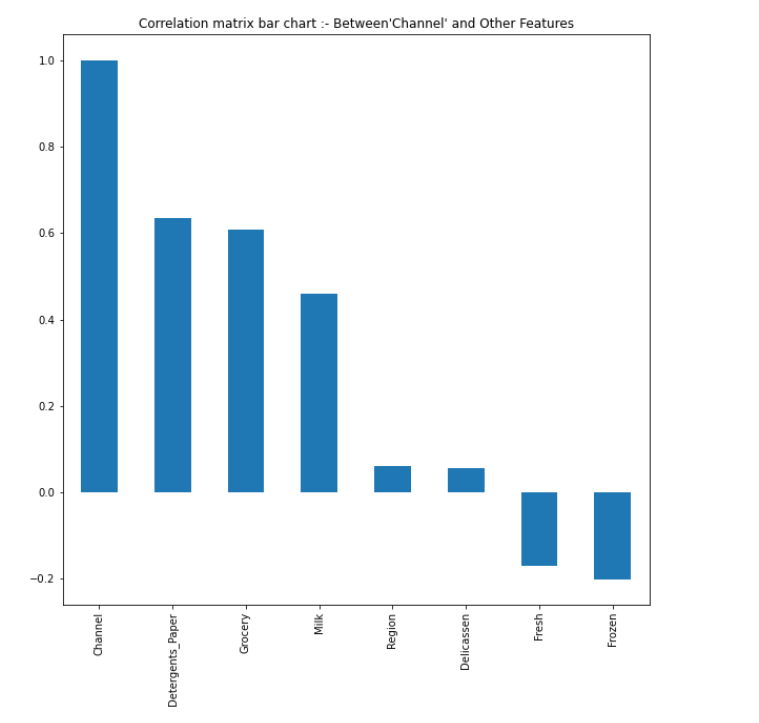
Missing values are found in many datasets, which may impact the machine learning model efficiency. Missing values are often imparted in the dataset as NaN or blanks. Missing value can be analyzed using pandas isnull() function. In this dataset, we could find no missing values.



## **Feature Selection**

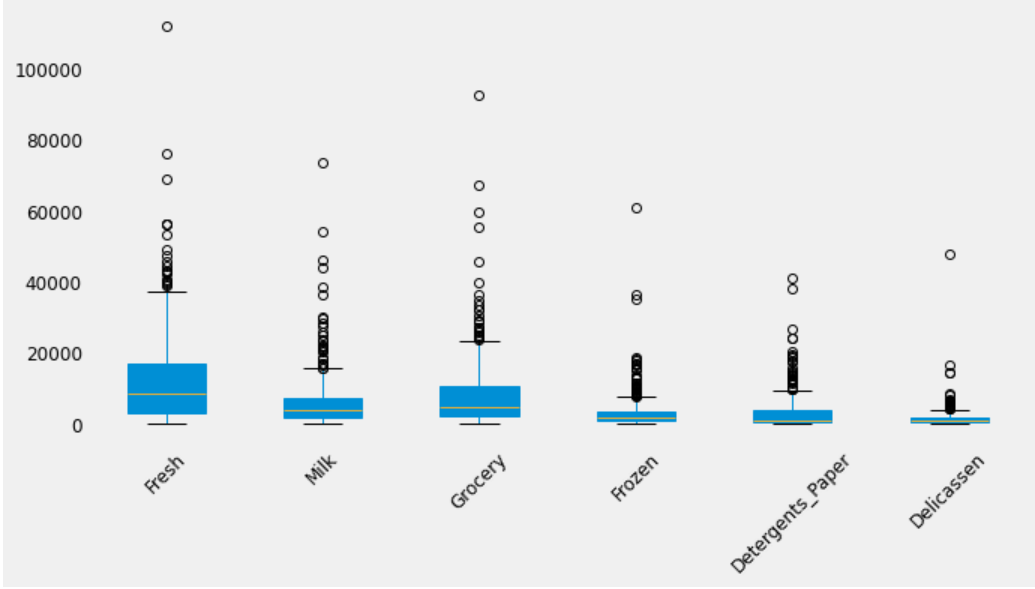
A correlation heatmap of each feature are visualized to get a better understanding of the dataset. If the feature is highly relevant, it shows no correlation with others in the dataset.

Higher correlation was found for detergents\_paper with grocery and milk, which implies that detergents\_paper is a weak feature and can be excluded. Here, we have not removed any feature, not to make our model very simple and to address underfitting problem.

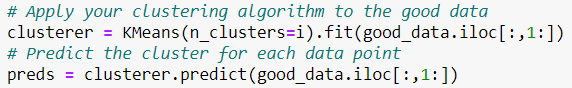
## **Outlier detection and removal**

Detecting outliers in the data is extremely important in the data preprocessing step of any analysis. Presence of outliers will lead to falsified predictions. An outlier step is calculated as 1.5 times the interquartile range (IQR). A data point with a feature that is beyond an outlier step outside of the IQR for that feature is considered abnormal. The dataset instances reduced to 332 after outlier removal process.



# **Unsupervised Learning using K-means clustering**

K-means clustering refers to an unsupervised machine learning algorithm widely used for data cluster analysis. When applied to the ‘Wholesale customers dataset’, this algorithm identifies the hidden customer segments that exist in our dataset.



The validation or goodness of a cluster is analyzed by measuring each data point's silhouette coefficient with a distance metric called the Euclidian distance. This coefficient's value ranges from -1 to +1, where a high value denotes the closeness of data points to its clusters compared to neighbouring clusters.



Plotting the average silhouette score for a different number of clusters is useful in finding the optimum number of clusters. In our case, we got a high silhouette score of 0.399 for k=2.

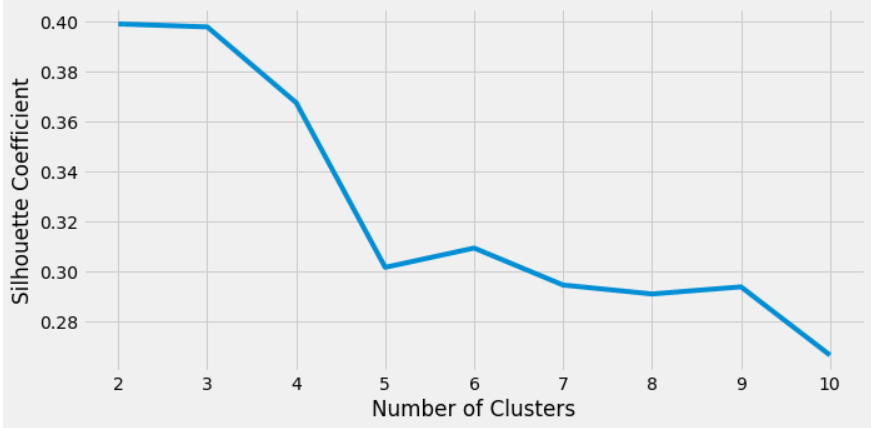


Figure : Plot showing silhouette score for different values of k

As the number of cluster increases, the silhouette coefficient tend to decrease. A low value for this coefficient indicates that the distance between different clusters is not significant, making it difficult to distinguish them.

# **Supervised Learning**

## **Train Test split**

Train Test split is the process of dividing the dataset into a training set and a test set. We performed a random split to use 70% of our data to train the model and the remaining 30% to test the model.

## **Feature Scaling**

Feature scaling is a crucial data pre-processing step used to bring all the features to a standard scale to avoid highly varying magnitudes or feature values. The most commonly used methods for feature scaling are Min-Max Normalization and Standardization. Among these methods, we have used the Standardization technique to re-scale the feature values so that the overall distribution has zero mean and unity variance.

## **Classification using Naive Bayes**

Fitting data into a model involves running a machine learning algorithm on the training set for which the target variable is known to produce a machine learning model. A good model fit is essential for creating more accurate predictions. Naive Bayes algorithm based on Bayes theorem is a supervised machine learning algorithm suitable for binary classification problems.

The **GaussianNB classifier is used**to fit into the training dataset using the fit () method. Naive Bayes classifier is one of the most simple and highly accurate algorithms that help make fast predictions.

Model fitting is followed by predicting the test set results. The accuracy of the predictions made are analyzed using the Confusion matrix which determines the number of times the model gives correct and wrong predictions. Below shown is the obtained confusion matrix.

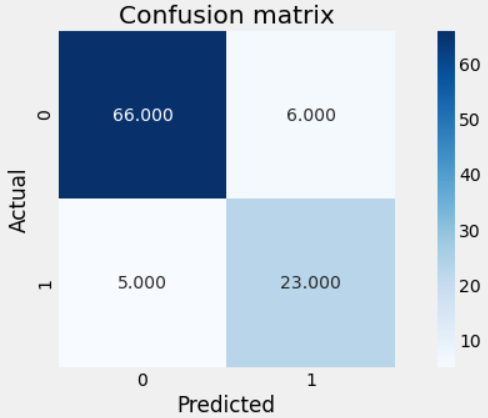
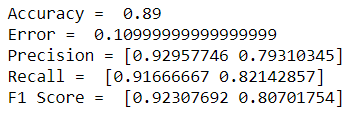


Figure : Naive Bayes confusion matrix

Confusion matrix output shows that there is a total of 66+23 correct predictions and 5+6 incorrect predictions which accounts for an accuracy of 89%

The main metrics of the classification report are precision, recall, f1-score, and support. The f1 score is the harmonic mean of precision and recall. It shows that the classifier is 92% accurate in classifying data points in one class and 80% accurate for other class.



## **Classification using SVM classifier**

Support Vector Machine (SVM) is an ML algorithm widely used for classification tasks. To classify data points, SVM finds a hyperplane in an N-dimensional space where N indicates the number of features. The hyperplanes which act as decision boundaries are chosen such that the distance between data points (margin distance) of two classes is maximum.

SVM classifier is built with SVC (Support Vector Classifier) from Scikit-Learn. We have trained SVM with Radial Basis Function (RBF) kernel, which helps to create complex decision boundaries. After training the model, it is essential to see the performance of the model. The model performance is analyzed by creating a confusion matrix, which gives the number of true positives, true negatives, false positives and false negatives. Below shown is the obtained confusion matrix.

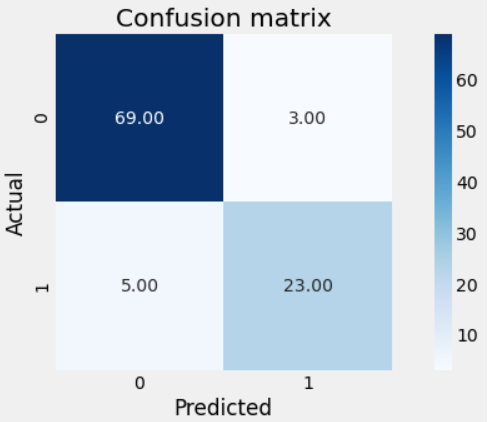
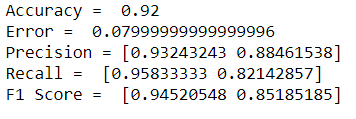


Figure : SVM Confusion Matrix

Confusion matrix output shows that there is a total of 69+23 correct predictions and 5+3 incorrect predictions which accounts for an accuracy of 92%

The obtained f1 score which is the harmonic mean of precision and recall shows that the classifier is 94% accurate in classifying data points in one class and 85% accurate for other class.



# **Feature Transformation – PCA**

Performing Principal Component Analysis (PCA) helps find dimensions that give maximum variance and determine which features best describe the customers when merged together. PCA is one of the most commonly used dimensionality reduction techniques in which highly correlated features combine to form new features, preserving majority of the relevant information.

## **K-means clustering with PCA**

When we initially performed K-means clustering without PCA, the highest silhouette score was 0.399 for K=2. Upon performing PCA, the maximum silhouette score increased to 0.45, and the optimum number of clusters has changed from two to three.

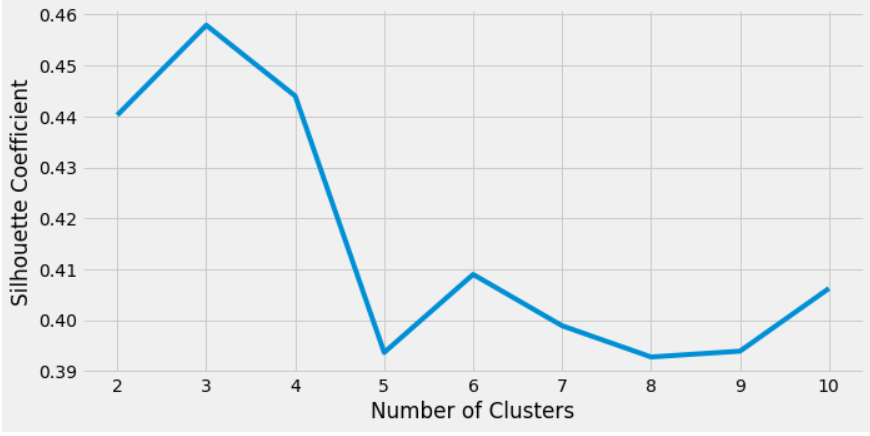


Figure : Plot of silhouette score for different K after PCA

## **Naïve Bayes classification with PCA**

The confusion matrix obtained after performing PCA is given below. The model gives 91 correct predictions and 9 incorrect predictions.

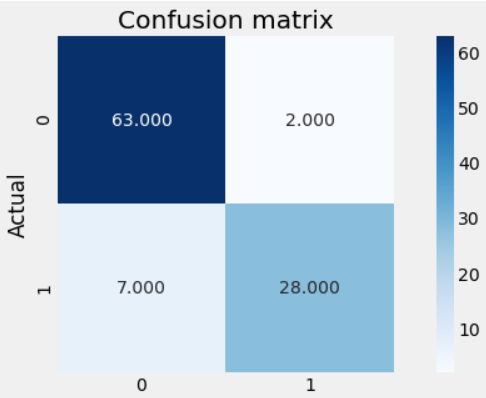


Figure : Naive Bayes confusion matrix with PCA

The obtained accuracy score with PCA is 91% and without PCA is 89%. There exist 2% increase in accuracy with PCA.

## **SVM classifier with PCA**

The obtained confusion matrix gives an accuracy of 91% which is one per cent less than that obtained without PCA.

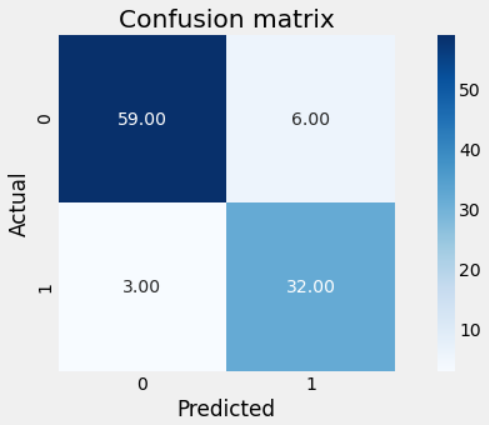
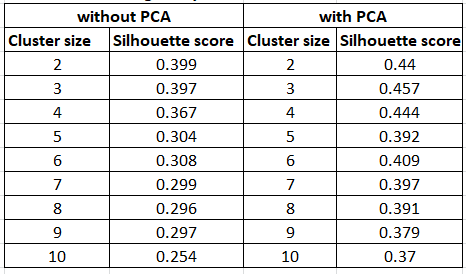


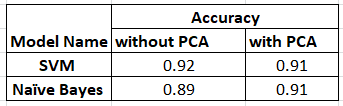
Figure : SVM confusion matrix with PCA

# **Results**

**K-means clustering analysis**



**SVM and Naïve Bayes result comparison**



# **Conclusion**

Predictive models were built on the 'Wholesale customers dataset' using supervised and unsupervised machine learning algorithms. K-means clustering, the unsupervised algorithm gave a maximum silhouette score of 0.399 for a cluster count of two. The score changes to 0.45 for a K value of 3 upon performing PCA. The supervised ML algorithms Gaussian Naive Bayes and SVM classification had an accuracy of 89% and 92%, respectively, before applying the PCA method. So we can conclude that SVM is better than Naïve Bayes algorithm. While the accuracy of SVM classification get reduced by 1% after PCA, the Naive Bayes algorithm's accuracy showed a slight increase of 2%.

# **References**

* <https://archive.ics.uci.edu/ml/datasets/wholesale+customers>
* <http://rstudio-pubs-static.s3.amazonaws.com/398634_b572cb3916f742e2865014328b38820d.html>
* <https://towardsdatascience.com/tagged/unsupervised-learning>