|  |
| --- |
| HEPSİBURADA |
| Related Products Recommendation – Machine Learning Model Report |
| Data Scientist Job Assignment |

|  |
| --- |
| Ali GÜVEN  9-21-2022 |

Contents

[1. Abstract 2](#_Toc114742356)

[2. Related Documents 2](#_Toc114742357)

[3. Introduction 2](#_Toc114742358)

[4. Algorithm 2](#_Toc114742359)

[4.1. K-Means++ 2](#_Toc114742360)

[4.2. Recommendation Scores 2](#_Toc114742361)

[5. Results 3](#_Toc114742362)

[6. Conclusion 3](#_Toc114742363)

[7. Improvements 3](#_Toc114742364)

Table of Figures

[Figure 1: Number of clusters vs Silhouette Coefficient score 3](#_Toc114742365)

# Abstract

Model was chosen as K-Means++ machine learning clustering algorithm. In the problem, the dataset was provided as without labels so it is an unsupervised data. K-Means++ is the one of the most popular algorithm for clustering. After preparing the data, “sessionid” column was separated from the dataset, and put the dataset into Standard Scaler. Then, a pipeline was made for both scaler and machine learning model (K-Means).

# Related Documents

|  |  |
| --- | --- |
| **Document Number** | DOC-001 |

# Introduction

The column of “sessionid” shows the unique carts, therefore, after predicting the cluster number of the test cart, related recommendations are done with looking the carts in the predicted cluster. Standard Scaler is necessary to make balance amongs to features. Some features’ values are bigger than the rest. K parameter was decided to 5 according to the trials doing with changing K in a range. Silhouette Coefficient score was used for describing the model performance.

# Algorithm

# K-Means++

Instead of classical K-Means, K-Means++ choses the center points at initialization as select random one then chose the other centroids according to the samples and previously chosen centroid. Hence, poor performance of random initialization is enhanced. Next, update the centroids by checking the sum of the squared error between centroids and the samples.

# Recommendation Scores

When a test cart is come, firstly transformed like described in DOC-001. Then scaled and predict the label of the cart according to the machine learning model. Then the products of the carts in the cluster are found. They are put in an order according to their frequency in the cluster. Most bought product has highest score.

# Results

Chosen number of cluster is determined as 5 by looking the silhouette scores shown in Figure 1.

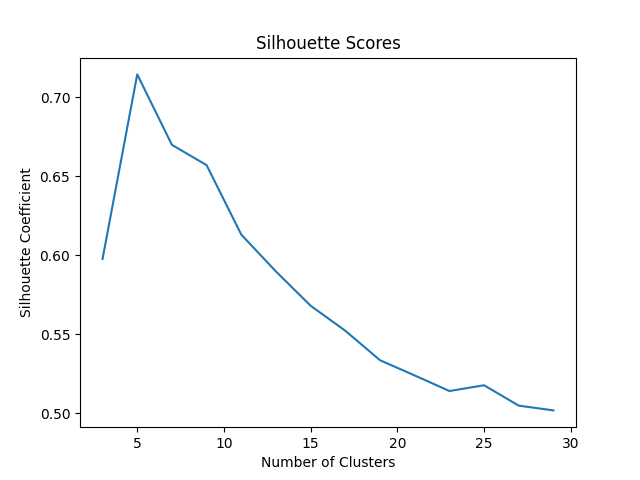


Figure 1: Number of clusters vs Silhouette Coefficient score

# Conclusion

K-means performed well with around 0.71 silhouette coefficient score. By clustering the carts, model was made independent from product number.

# Improvements

* Deep learning models can be done.
* Model is unsupervised, so supervised models can be outperforming the unsupervised models. Dataset may be gathered with labels.
* Different machine learning / deep learning models (Auto encoders, PCA) can be applied to see which one is more sufficient.