MEF UNIVERSITY

EMPLOYEE CLASSIFICATION

Capstone Project

Gökhan Şahin

İSTANBUL, 2018

MEF UNIVERSITY

EMPLOYEE CLASSIFICATION

Capstone Project

Gökhan Şahin

Advisor: Asst. Prof. Alper Yıkıcı

İSTANBUL, 2018

MEF UNIVERSITY

Name of the project: Employee Classification
Name/Last Name of the Student: Gökhan Şahin
Date of Thesis Defense:

I hereby state that the graduation project prepared by Gökhan Şahin has been completed under my supervision. I accept this work as a "Graduation Project".

Asst. Prof. Alper Yıkıcı

I hereby state that I have examined this graduation project by Gökhan Şahin which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

Prof. Dr. Özgür Özlük

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member		Signature
1.	Asst. Prof. Alper Yıkıcı	……..
2.	Prof. Dr. Özgür Özlük	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Gökhan Şahin		

EXECUTIVE SUMMARY

EMPLOYEE CLASSIFICATION

Gökhan ŞAHİN

Advisor: Asst. Prof. ALPER YIKICI

DECEMBER, 2018, 25

Employee is an important element of the organization. The success or failure of the organization depends on its performance. In this study, there are solutions that affect the theoretical framework and models of employee development and employee performance. A hybrid procedure based on data clustering can be used by the authority to estimate the performance of employees in the next month. This study shows how to implement the data clustering method to evaluate the performance of the work in the decision-making process. This study is done to eliminate the effectiveness of inefficient work, inefficiency, and effective lack of work.

The clustering analysis itself is not a specific algorithm, but a general task that needs to be solved. It can be achieved with various algorithms that differ significantly in their understanding of what constitutes a cluster and how it can be found effectively. Popular concepts of clusters include small distances between cluster members, dense areas of the data field, intervals, or certain statistical distributions. Clustering can therefore be formulated as a multipurpose optimization problem. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to be used, the density threshold, or the number of expected clusters) depend on the individual data set and the intended use of the results. Cluster analysis is not an automated task, but a recursive knowledge discovery process involving trial and failure, or an interactive multipurpose optimization process. As a result, it is usually necessary to modify the data preprocessing and model parameters until the desired properties are obtained.

The intent of this study is to provide an assessment methodology for human resource professionals. Employee profiles play a crucial role in the evaluation process in order to improve the performance of the training process. In this article, we focus on clustering in specific categories that represent the characteristics of employees according to their profiles. In this study, it is used that employee project number, production time, customer note and quality ratio. As a result of the study, the most valuable data objects are project number and production time.

Key Words: K Means, Employee Classification, Clustering

ÖZET

ÇALIŞAN SINIFLANDIRILMASI

Gökhan Şahin

Tez Danışmanı: Asst. Prof. Alper YIKICI

ARALIK, 2018

Çalışan, organizasyonun kilit unsurudur. Organizasyonun başarısı veya başarısızlığı çalışan performansına bağlıdır. Bu çalışma, çalışan gelişimi ve çalışan performansı üzerindeki etkisi ile ilgili teorik çerçeve ve modelleri analiz etmektedir. Veri Kümeleme yöntemine dayanan karma prosedür, çalışanların bir sonraki ay için performanslarını tahmin etme yetkisi tarafından kullanılabilir. Bu çalışma, karar verme sürecinde çalışan performansını değerlendirmek için veri kümeleme yönteminin nasıl uygulanabileceğini göstermektedir ve verimsiz çalışanın, verimsizliğin büyüklüğünün ve nispeten işe yarar bir çerçeve ile verimsizliği ortadan kaldırmak için etki eden çalışan değerlerinin bulunmasına yardımcı olmaktadır.

Kümeleme analizinin kendisi belirli bir algoritma değil, çözülmesi gereken genel görevdir. Bir kümelenmeyi neyin oluşturduğu ve bunların nasıl etkili bir şekilde bulunabileceği konusundaki anlayışların önemli ölçüde farklı olan çeşitli algoritmalar ile başarılabilir. Kümelenmelerin popüler kavramları arasında küme üyeleri arasındaki küçük mesafeler, veri alanının yoğun alanları, aralıkları veya belirli istatistiksel dağılımlar yer alır. Kümeleme bu nedenle çok amaçlı bir optimizasyon problemi olarak formüle edilebilir. Uygun kümeleme algoritması ve parametre ayarları (kullanılacak mesafe işlevi, yoğunluk eşiği veya beklenen kümelerin sayısı gibi parametreler dahil), tek tek veri kümesine ve sonuçların kullanım amacına bağlıdır. Küme analizi, otomatik bir görev değil, deneme ve başarısızlığı içeren yinelemeli bir bilgi keşif süreci veya etkileşimli çok amaçlı optimizasyon sürecidir. Sonuçta istenen özellikleri elde edinceye kadar, veri ön işleme ve model parametrelerini değiştirmek genellikle gereklidir.

Bu makalenin amacı, insan kaynakları birimi için bir değerlendirme metodolojisi sağlamaktır. Çalışan profilleri, eğitim sürecinin performansını iyileştirmek için değerlendirme sürecinde çok önemli bir rol oynamaktadır. Bu çalışma, çalışanların özelliklerini profillerine göre temsil eden belirli kategorilerde kümelenmeye bulmaya odaklanmaktadır. Çalışma sonucunda en değerli veri nesneleri proje numarası ve üretim zamanıdır.

Anahtar Kelimeler: K-Means, Çalışan Sınıflandırılması, Kümeleme

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Overview

Most of the organizations or companies have a formal performance evaluation system in which employee job performance is graded on a regular basis, usually once or twice a year. A good performance evaluation system can prominently benefit an organization. It helps employee behavior toward organizational aims by permitting employees know what is expected for them, and it yields information for making employment decisions, such as those regarding pay raises, promotion or releases. Developing and implementing an effective system is not an easy task. An Employee can improve their performance by way of monitoring the progression of their performance. Machine learning algorithms i.e. clustering algorithm and decision tree of data mining technique can be used to find out the key characteristics of future prediction of an organization. Clustering is a method to group data into categories with identical characteristics in which the similarity of intra-class is maximized or minimized.

The main purpose of this study is to enable employees to distinguish between homogeneous groups according to their characteristics and abilities using clustering. This study uses cluster analysis to divide employees into groups according to their performance. Based on the performance results of the employee, it is possible to decide whether to require further training, talent enhancement or more qualification. These practices also help administrative staff improve the quality of their organization.

1.2 Literature Review

In the past, many scientists made researches on the topic of human resources performance evaluation and classification through clustering. This study proposes an approach that can roughly cluster a data set with fuzzy linguistic entries as a prior data arrangement for performance evaluation of R&D employees. The extension principles of fuzzy linguistic numbers are used to modify the K-means method for handling the linguistic data set. The absolute difference of fuzzy linguistic variables is defined as fuzzy distance. Based on this definition, the K-means approach can be modified slightly for clustering purposes. The performance of employees engaged in designing and

R&D-oriented jobs is possibly related to some qualitative attributes and the evaluation of such attributes for each employee has a tendency toward semantic scales. In the proposed approach, the supervisor can evaluate the performance of each employee directly using a semantic scale. The modified K-means approach can roughly cluster their performance into different classes in advance of applying some other sophisticated processes (Hong Tau Lee , Sheu Hua Chen , Jie Min Lin,1996)

Using an innovative research methodology known as Kohonen's Self-Organizing Maps, is exploring the link between competitive advantage between human resources management and perceived organizational performance in the European Union private and public sectors. The results of the study demonstrated the utility of an innovative technique when applied to the research conducted to date through traditional methodologies and led to questions about the universal applicability of the widely accepted relationship between better human resource management and better business performance (Stavrou, Charalambous and Spiliotis,2007).

The success or failure of an organization depends on the employee performance. Hybrid procedure based on clustering of data mining method may be used by the authority to predict the employees' performance for the next year. Data clustering method can be applied for evaluating the employee's performance as well in decision making process. The inefficient employee, magnitude of inefficiency and aids to eliminate inefficiencies with a relatively easy to employ framework (Sarker, Shamim, Zama and Rahman,2018).

2. ABOUT THE DATA

2.1 Data Description

The given dataset covers the employee records in a huge transactional database. The study is based on data from employee of a company in Istanbul, Turkey.

The transaction database consists of the following information:

- Employee ID – Unique basket identification;
- Production Time – Working hours;
- Seniority – The time spent in company;
- Age – Employee age;
- Project Number – The number of project per month;
- Quality ratio – Quality ratio of employee's job
- Customer Points - Job score

This study contains in total 4244 employee data.

3. PROJECT DEFINITION

3.1 Problem Statement

The performance evaluation system is used to determine the organization's skilled and best performing employees. It is implemented in the organization to increase their salaries and other benefits. However, most employees are not satisfied with the performance evaluation and therefore are leaving their job. This study created the work of employees and the production of values.

3.2 Project Scope

The scope of this project includes:

- to develop employee classification model.
- the analysis will be based on Clustering.
- to understand the clustering within the classification.

4. MEDHODOLOGY

4.1 Clustering

Data Clustering is an unsupervised and arithmetic data analysis procedure. Cluster analysis is used to separate a large set of data into subsets called clusters. While the data points in the same group have similar features, the data points in the different groups must have different characteristics. Each cluster is a collection of data objects that are similar with one another data object within the same cluster but are dissimilar to objects place other clusters. It is used to classify the same data into a homogeneous group. It's also used to operate on a large data-set to discover hidden pattern and relationship which helps to make decision quickly and efficiently. Data clustering algorithm is an effective method that is being used to evaluate the employee performance. (Sarker, Shamim, Zama and Rahman,2018).

4.2 K Means Algorithm

K-means Clustering (3) is a type of unsupervised learning, which is used in unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

K-Means clustering intends to partition n objects into k (number of clusters) clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function. Given a set of data or documents (x_1,x_2,…, x_i), where each data point is an M-dimensional real vector, the objective of the algorithm is to partition n documents into k clusters ($k \leq n$) with minimizing an objective function, which may be expressed in Formula 1.

$$j = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^j - c_j||^2$$

Formula 1

$$|\left|x_i^j - c_j\right|\,|^2$$

Formula 2

Where J is the object function, k number of cluster, n number of cases, and Formula 2 a chosen distance measure between a data point xi and the cluster center c_j, The algorithm and flow-chart of K-means clustering is explained step by step and the flowchart is shown in Figure 1.

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values

Step 2: Initialize the first K clusters

- Take first k instances or
- Take random sampling of k elements

Step 3: Calculate the arithmetic means of each cluster formed in the dataset.

Step 4: K-means assigns each record in the dataset to only one of the initial clusters.

Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all the clusters in the dataset.

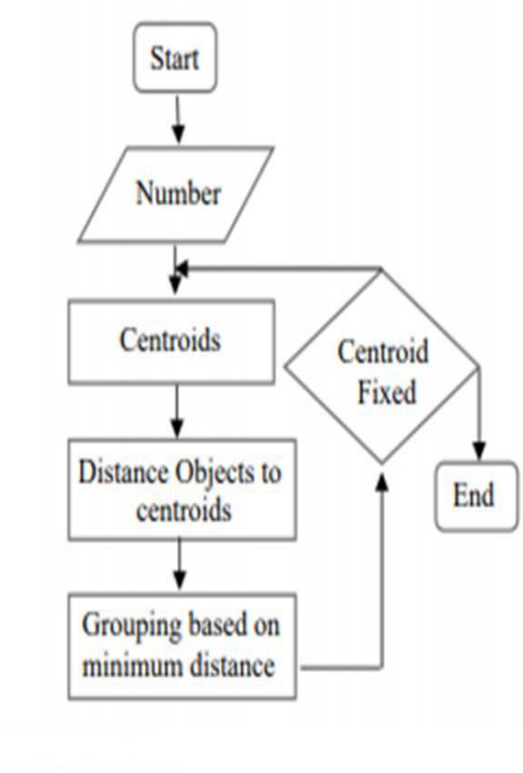

Figure 1. Flowchart of K means Clustering

4.3 Hierarchical Algorithm

In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Generally, there are two types strategies for hierarchical clustering:

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram.

The standard algorithm for hierarchical agglomerative clustering (HAC) has a time complexity of $Q(n^3)$ and requires $Q(n^2)$ memory, which makes it too slow for even medium data sets. However, for some special cases, optimal efficient agglomerative methods (of complexity $Q(n^2)$ are known: SLINK for single-linkage and CLINK for complete-linkage clustering. With a heap the runtime of the general case can be reduced to $Q(n^2 \log n)$ at the cost of further increasing the memory requirements. In many programming languages, the memory overheads of this approach are too large to make it practically usable (Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome,2009).

Except for the special case of single-linkage, none of the algorithms (except exhaustive search in $Q(2^n)$ can be guaranteed to find the optimum solution.

Divisive clustering with an exhaustive search is $Q(2^n)$, but it is common to use faster heuristics to choose splits, such as k-means.

5. RESULT

Results obtained from K-Means Algorithm and Hierarchical Clustering Algorithms are presented in detail.

5.1 K-Means Algorithm

The Min Max Scaler method converts properties by scaling each property to a specific range. Each property is individually adjusted on the data set, that is, within the specified range from zero to one. This process is distributing the variable values according to the range 0-1. Scaling was performed in the data set between 0 and 1, followed by the conversion process.

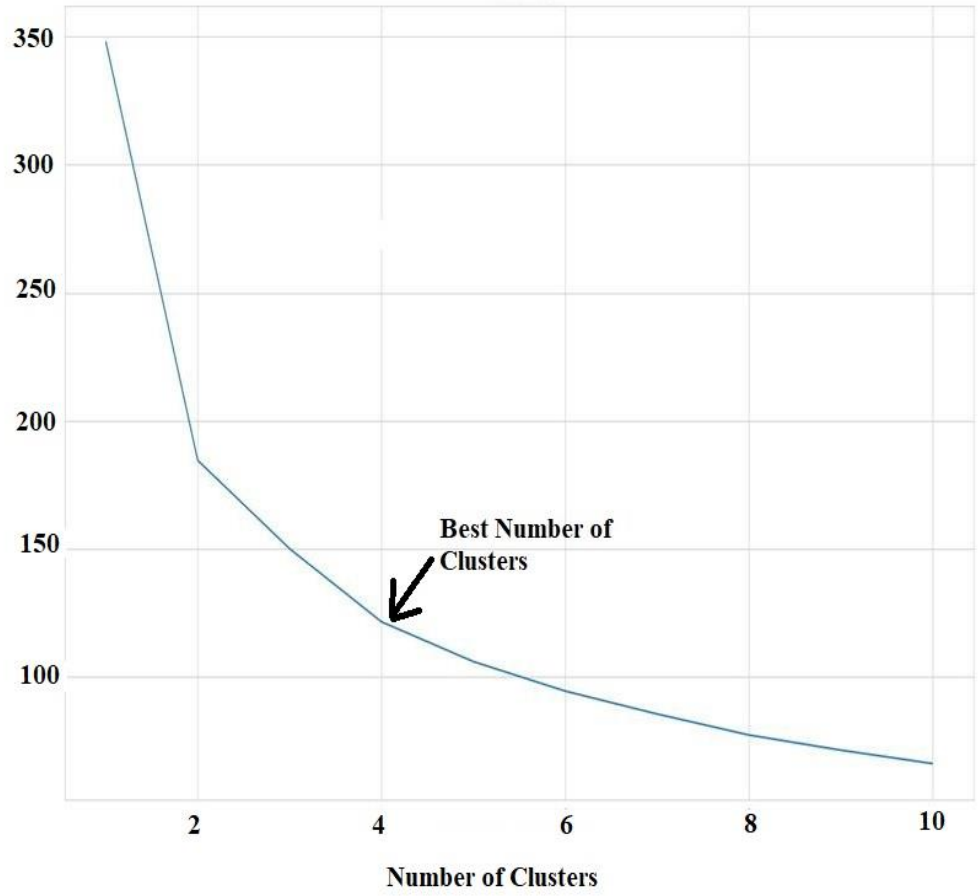

Figure 2. Within Clusters Sum of Square (WCSS)

Figure 2 shows the number of clusters and the relationship between WCSS. The line graph of the sum of the square errors (SSE) for each number of clusters value is displayed. Good model WCSS is lower. The objective is to achieve a small SSE value, but the SSE tends to fall to 0 as it increases k. Up to cluster number 4, WCSS falls very quickly. 4 is the optimal number of clusters for this data set. Because more clusters do not reduce WCSS, it is also reducing the interpretability of the model.

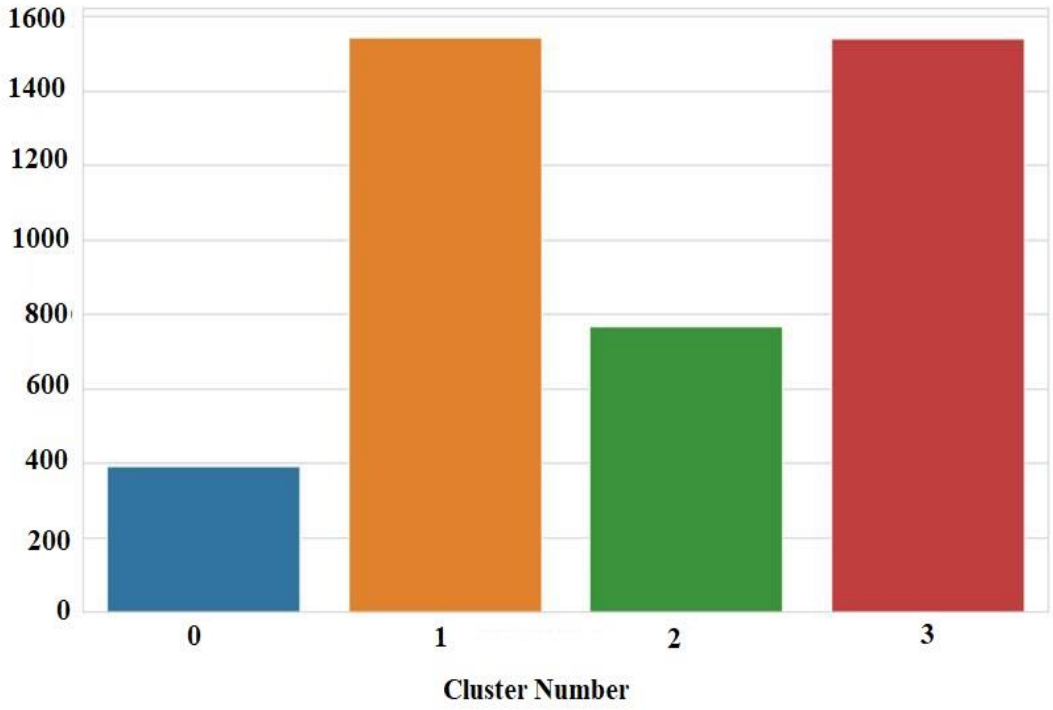

Figure 3. Data Distribution Graph of Clusters

It’s known that a cluster is the largest and a cluster has the least number of employees. The data in Figure 3 represents mostly Cluster 1 and Cluster 3.

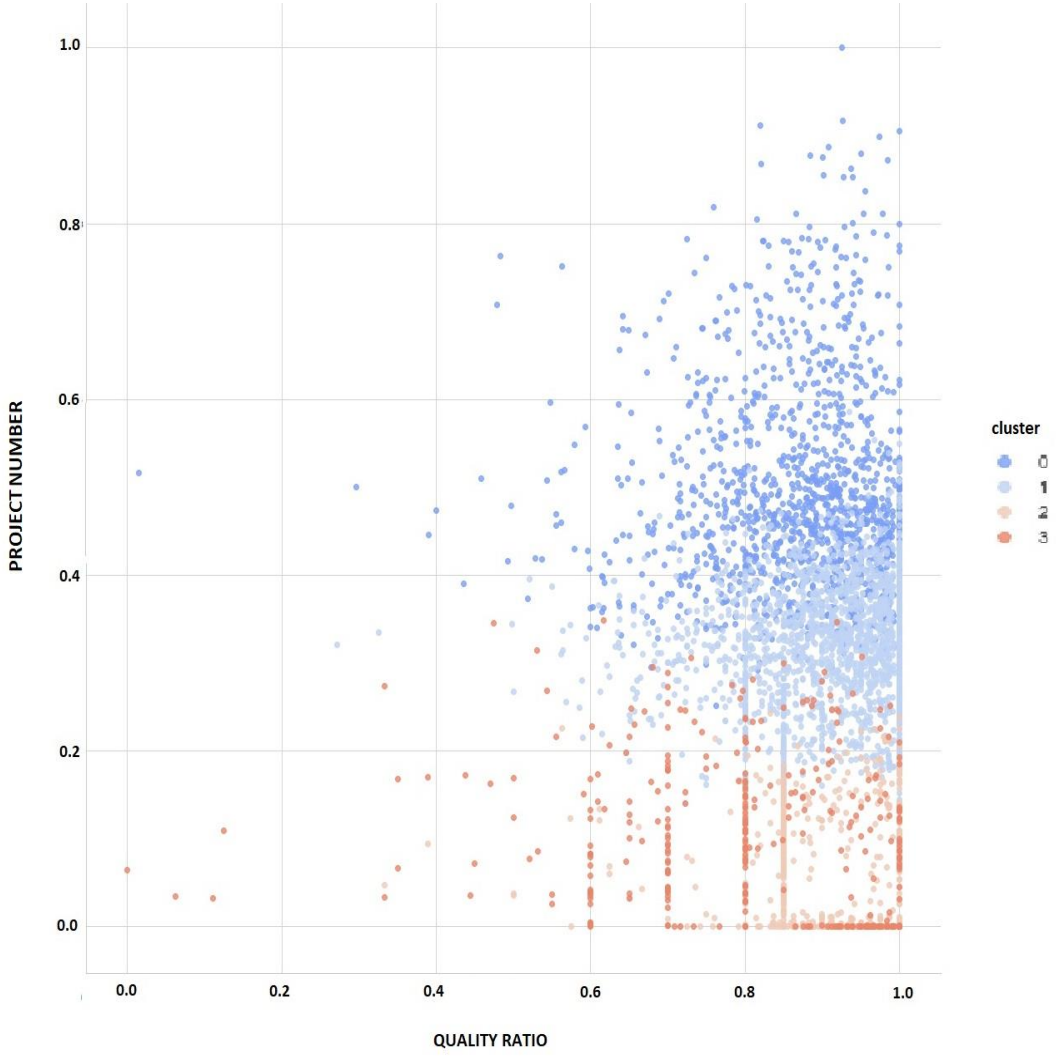

Figure 4. Project Number-Quality Ratio

Figure 4 shows that distribution of employee's Project Number and Quality Ratio. Cluster 2 and Cluster 3 are located under other clusters related to the project number. This means that they have fewer projects than the other clusters.

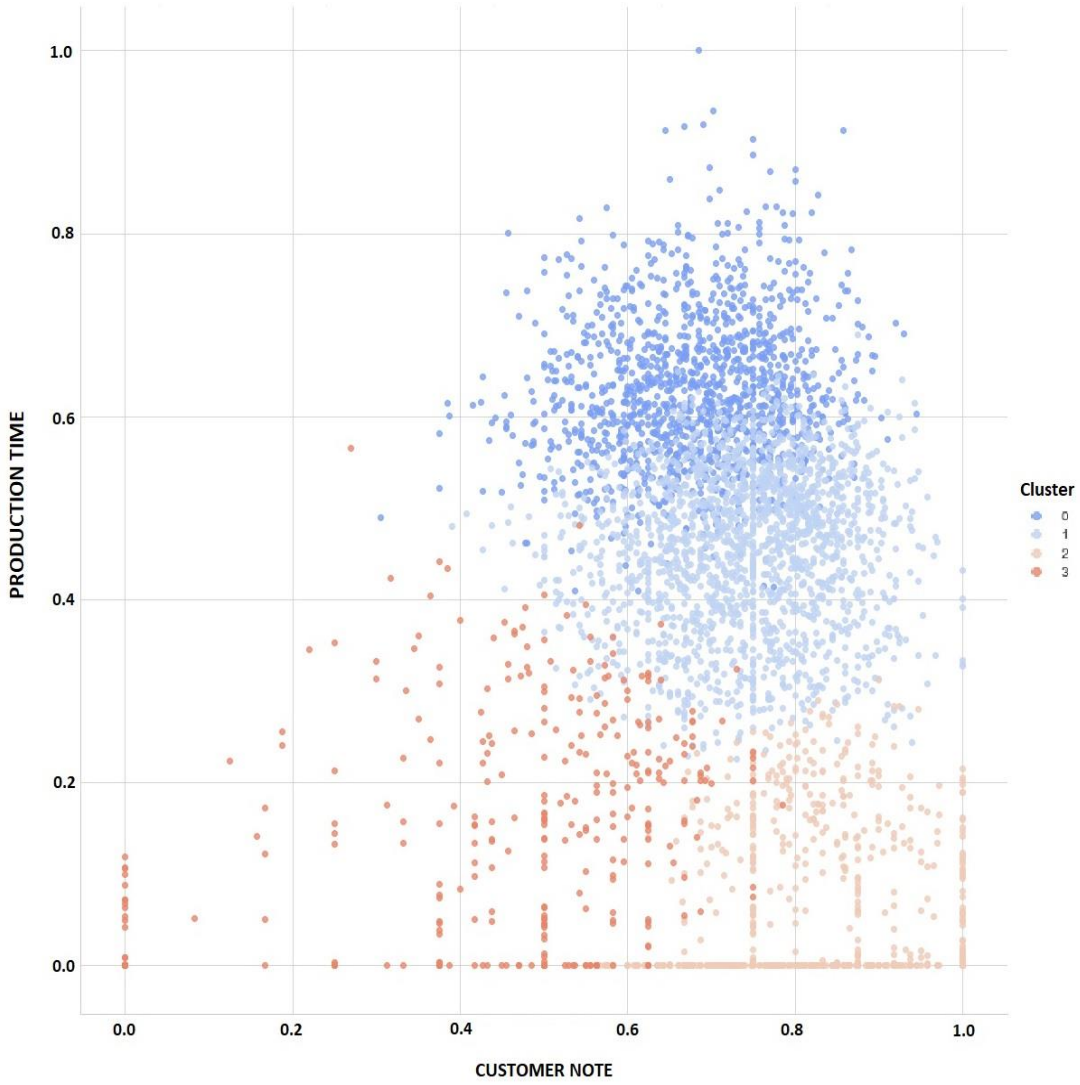

Figure 5. Production Time – Customer Note

Cluster 0 is located upper other clusters to the production time (Figure 6). It means that Cluster 0 has employees who higher production time and customer note. In Figure 6, some employees of Cluster 3 appear to have lower customer ratings than employees in other clusters.

5.2 Hierarchical Clustering Algorithms

Hierarchical clustering algorithms fall into 2 categories: top-down or bottom-up. Bottom-up algorithms treat each data point as a single cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged

into a single cluster that contains all data points. Bottom-up hierarchical clustering is therefore called hierarchical agglomerative clustering or HAC. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample (Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome,2009).

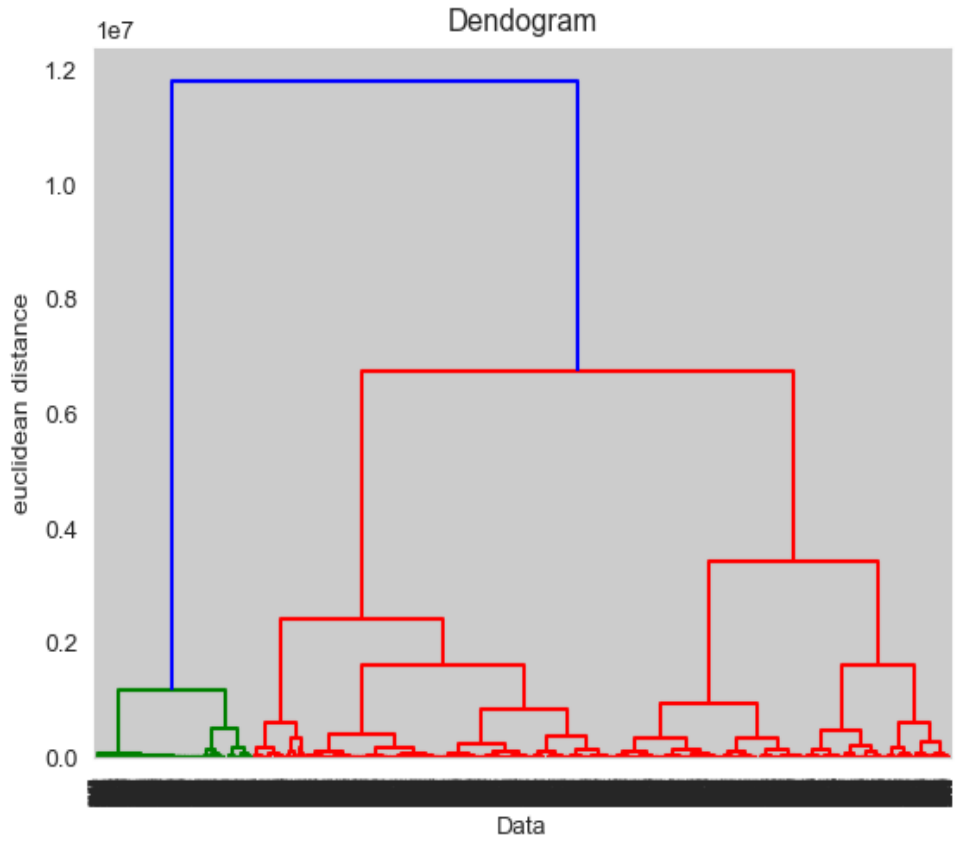

Figure 6. Euclidean Distance

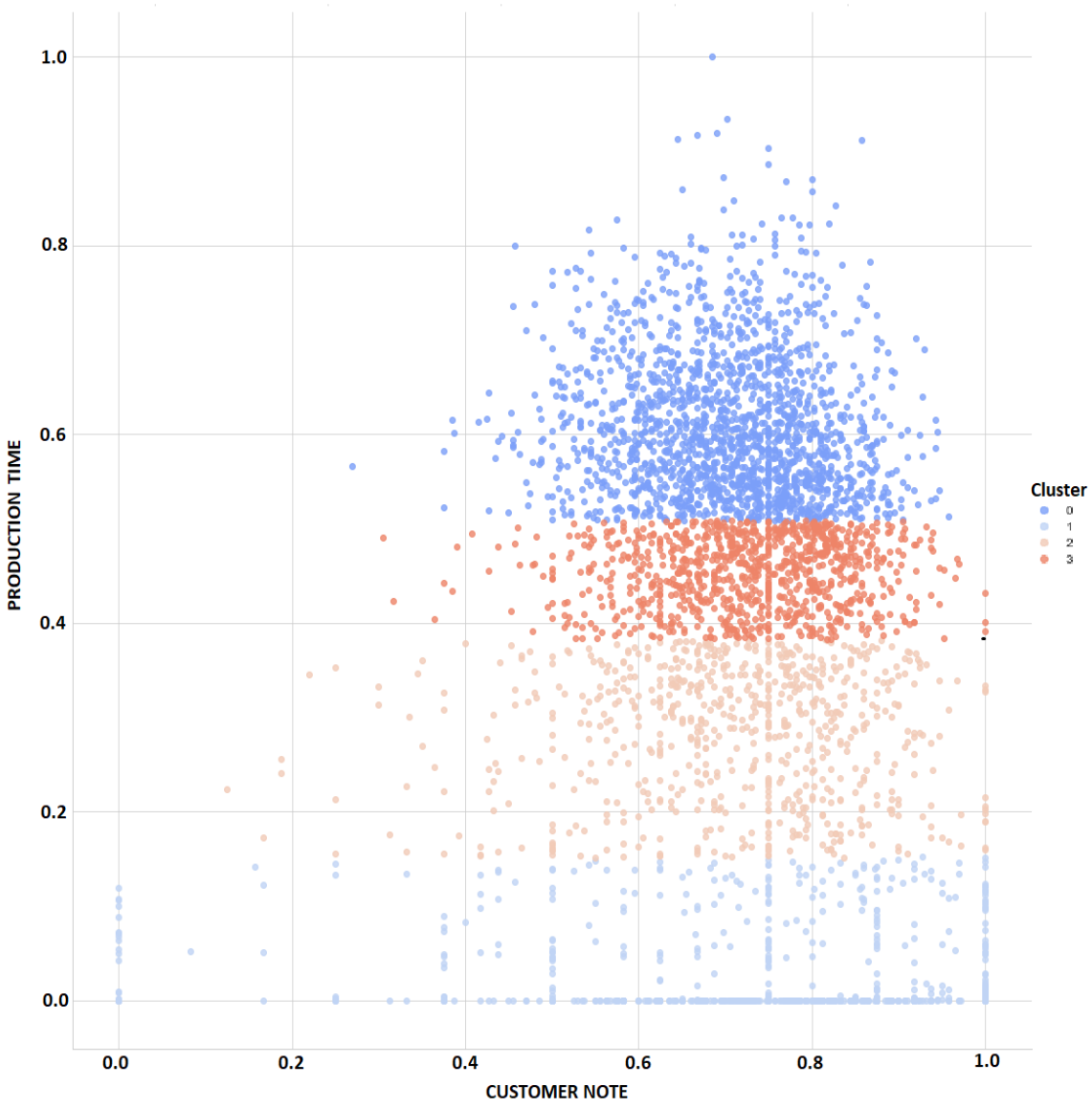

Figure 7. Production Time – Customer Note

Cluster 1 is located under other clusters related to the production time (Figure 8). So, Cluster 1 has employees who lower production time and customer note. Cluster 0 is located above the other clusters related to the production time (Figure 8). It means that Cluster 0 has employees who higher production time and customer note.

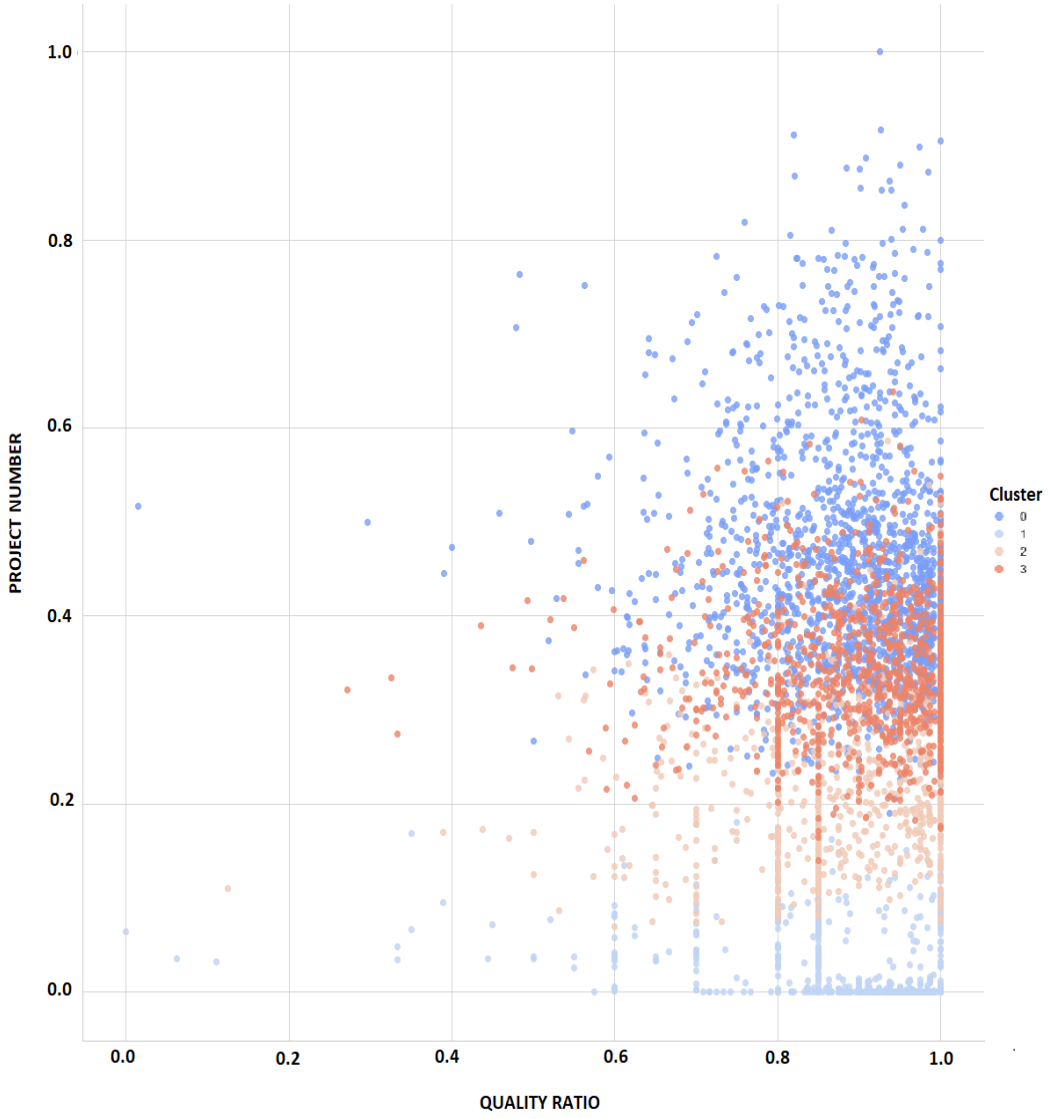

Figure 8. Project Number – Quality Ratio

Cluster 1 is located under other clusters related to the project number (Figure 10). So, Cluster 1 has employees who lower project number. Although, other data objects are very close to each other so there is not saw any separated clusters.

6. SOCIAL AND ETHICAL ASPECTS

Understanding employees' behaviors, managing and deciding employees according to their needs has become a tremendous problem. Technological innovations shave

upholstery in the faster processing of queries and response times of the second. Data mining tools have become the most powerful tool to analyze data in large quantities and to break ground in making the right decisions. The purpose of this project report is to analyze the data, thus to evaluate the behavior of the employees and to make decisions by the management according to the situation of the employees. Experimental analysis used clustering method to prove its value according to traditional methodologies.

7. CONTRIBUTION

The K means algorithm provides better clustering K means when compared with the Hierarchical algorithms. Cluster 0 has the low number of projects, the low customer score and the low production value, so training can be provided such as communication, productive work, etc. it is necessary to work to increase the work for these employees. Cluster 1 work hard, but they need to update their customer notes. The target in this group can be determined and rewarded. Cluster 2 has a low number of projects, high customer score and low production value. For this group work motivations can be increased.

8. REFERENCES

1- Hameed, A., & Waheed, A. International journal of business and social science, 2(13), 2011.

2- Azar, A., Sebt, M. V., Ahmadi, P., & Rajaeian, A.,"A Model for Personnel Selection with A Data Mining Approach: A Case Study in A Commercial Bank: Original Research". SA Journal of Human Resource Management, 11(1), 2013. p. 1-10.

3- Shovon, M., Islam, H., & Haque, M.,"An Approach of Improving Students Academic Performance by using k Means Clustering Algorithm and Decision Tree". arXiv preprint arXiv:1211.6340, 2012.

4- Çalışkan, S.K. ve Soğukpınar, İ. KxKNN: K-means ve K en yakın komşu yöntemleri ile ağlarda nüfuz tespiti. 2. Ağ ve Bilgi Güvenliği Sempozyumu, Girne, 2008. p. 120-124.

5- Stavrou, E. T., Charalambous, C. and Spiliotis, S. "Human resource management and performance: A neural network analysis," European Journal of Operational Research, vol. 181, 2006. p. 453-67.

6- Hong Tau Lee ,Sheu Hua Chen ,Jie Min Lin, **"**K-means method for rough classification of R&D employees' performance evaluation", 1996.

7- Kumar SA, Vijayalakshmi MN. Mining of student academic evaluation records in higher education. In: Recent Advances in Computing and Software Systems (RACSS), 2012 International Conference on IEEE; 2012. p. 67-70.

8- Geng X, Luo L. Multilabel ranking with inconsistent rankers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 3742-7.

9- Bouhmala N. How good is the Euclidean distance metric for the clustering problem. In: Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on IEEE; 2016. p. 312-5.

10- Esteves RM, Hacker T, Rong C. Competitive k-means, a new accurate and distributed k-means algorithm for large datasets. In: Cloud Computing Technology and Science (Cloud Com), 2013 IEEE 5th International Conference on IEEE. Vol. 1; 2013. p. 17-24.

11- Nazeer, K. A., & Sebastian, M. P., "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm". In Proceedings of the World Congress on Engineering 2009.

12- Oyelade, O. J., Oladipupo, O. O.,& Obagbuwa, I. C.,"Application of k Means Clustering algorithm for prediction of Students Academic Performance". arXiv preprint arXiv:1002.2425, 2010.

13- Kumar KM, Reddy AR. A fast K-means clustering using prototypes for initial cluster center selection. In: Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on IEEE; 2015. p. 1-4.

14- Poteraş CM, Mocanu ML. Evaluation of an optimized K-means algorithm based on real data. In: Computer Science and Information Systems (Fed CSIS), 2016 Federated Conference on IEEE; 2016. p. 831-5.
15- Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu, " A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", The Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, 1996.
16- Adriano Moreira, Maribel Y. Santos and Sofia Carneiro, "Density-based clustering algorithms – DBSCAN and SNN", 2005.
17- Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. p. 321-352.
18- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome "Hierarchical clustering". The Elements of Statistical Learning (2nd ed.). Springer. New York, 2009. p. 520–52.