Iranian Journal of Management Studies (IJMS) 2022, 15(2): 207-226
RESEARCH PAPER

Providing a Hybrid Clustering Method as an Auxiliary System in Automatic Labeling to Divide Employee Into Different Levels of Productivity and Their Retention

Seyed Alireza Mousavian Anaraki, Abdorrahman Haeri*, Fateme Moslehi

School of Industrial Engineering, Iran University of Science & Technology, Tehran, Iran

(Received: April 13, 2020; Revised: June 29, 2021; Accepted: August 15, 2021)

Abstract
Identifying productive employees and analyzing their turnover by data mining tools without human intervention is an attractive research field in human resource management. This study develops an innovative auxiliary system for automatic labeling of numerical data by providing a hybrid clustering algorithm of K-means and partition around medoids (PAM) methods to identify organizational productive employees and to divide them into different productivity levels. The model is evaluated by calculating the differences between actual and labeled values (93% labeling accuracy) and an innovative criterion for image processing of the final clusters using the singular value decomposition (SVD) algorithm. Ultimately, the results of the algorithm determine four labels of middle and good productive employees who leave the organization and excellent and weak productive employees who stay in the organization; according to each cluster, policies are adopted for their retaining, productivity improvement, and replacement.

Keywords: productive employees, employee turnover, hybrid clustering, auto labeling, image processing

1. Introduction

Productivity is a measure of the success of a system in collecting and using resources to achieve goals (Okoye & Ezejiofor, 2013) from a managerial perspective that encompasses both the concepts of effectiveness and efficiency (Tajeddini, 2015). It is, therefore, related to the concepts of efficiency, effectiveness, profitability, quality, innovation, quality of work-life, and culture, as well as to a combination of the above (Pritchard, 1992). Efficiency and effectiveness – as definitions of productivity – make it comprehensive and this is confirmed by many scholars, some of whom believe that limiting productivity to efficiency or effectiveness causes ambiguity and does not provide accurate information to the organization (Ilgen & Klein, 1988). Productivity could be defined as the effectiveness of using the agents of production to produce goods and services (Shaker Ardakani et al., 2016). Effectiveness is defined as the level of output (Berman & Berman, 1998) and considered as one of the essential goals of an organization, and every organization tries to raise such output (Rahmati et al., 2014). Efficiency is the optimal conversion of inputs into outputs (Salem, 2003). Organizational productivity is the most proper use of human resources to move to corporate goals and objectives with minimal time and cost, which measures the organization’s performance. It is also an indication of the efficiency and competitiveness of departments. Measuring organizational productivity has many benefits (Faghihi & Mousavi kashi, 2010),

* **Corresponding Author, Email:** ahaeri@iust.ac.ir

including (a) strategic and operational planning, (b) allocating organizational budgets and resources, (c) identifying opportunities for improvement (especially quality improvement) and helping managerial decisions (Clarke, 1991), (d) comparison of organizational performance with internal and external standards, and (e) providing a basis for determining salaries and bonuses (Prichard, 1990). In addition, productivity affects the foresight and organizational transformation in all areas of human resource management (HRM) (Friedman, 2006), such as selection and hiring, performance management, employee retention, training and development, talent management, and competency management (Hajiheydari et al., 2017). Human resources are the most crucial resources that managers can use to achieve organizational goals, and employees are the most critical factor in improving productivity (Abtahi & Kazemi, 2001). Thus, paying attention to human resources is one of the essential strategies for increasing the efficiency and productivity of an organization (Bertschek et al., 2006). Employee turnover and retention are among the most important arenas of human resources, and productivity must be considered for these fields. This is because the long-term success or failure of the organization depends on the retention of human resources, which in turn increase the productivity of the business (Azar et al., 2010; Storey, 2016; Vasantham & Swarnalatha, 2015). In addition, it is difficult to find a suitable replacement for skilled employees. It takes time, money, and a lot of effort to bring new employees to the levels of expertise and productivity of the previous employees (Saradhi & Palshikar, 2011). There are also various forms of employee turnover, namely voluntary (employee resignation) or involuntary (induced by the company) (Ribes et al., 2017), functional or inefficient, and avoidable or inevitable (Chalkiti & Sigala, 2010). Leaving an organization without highly productive staff is one of the adverse effects of an inefficient loss, and this separation is effective when it comes to low-productivity employees (Ilmakunnas et al., 2005). Functional turnover means dismissing employees with poor performance and retaining those with excellent performance in their jobs. Inefficient turnover also means that good-performing employees leave their work, and poor-performing ones stay in their jobs (Sexton et al., 2005). As the human resources sector has essential employee information, the data mining process can be used in HRM domains. Data mining is a systematic process for extracting valuable knowledge across a large volume of data to solve business problems (Provost & Fawcett, 2013). It uses advanced tools for discovering sensitive data patterns (Roiger, 2017). Accurate predictions with data mining tools enable companies to take action for employee retention or succession planning (Ajit, 2016; Chanodkar et al., 2020). As one of the most important data mining methods (Reddy et al., 2013), clustering is among the most basic ways of data analysis with complete applications (Simić et al., 2018) used in many research fields, including pattern recognition, machine learning, statistical learning, data mining, text mining, bioinformatics, and numerous application domains (Cao & Liang, 2011). Clustering is an unsupervised learning method for organizing unlabeled data that exist in similar groups known as clusters (Jain et al., 1999). Clustering algorithms are classified into two groups, namely hierarchical and partitional clustering (Datta & Datta, 2003). Partitional clustering algorithms (such as K-means and partition around medoids (PAM)) are the most popular methods (Celebi, 2014). They have aspects such as simple implementation and low computational complexity (Anaraki et al., 2021). In the meantime, the topic of automatic labeling has been neglected in numerical data, but it has been chiefly used in text mining and image processing (Kusumaningrum, 2017).

The present study presents a hybrid clustering approach for auto-labeling as an assistance system for dividing employees into different levels of productivity and surveying the type of their turnover to provide appropriate policies for their retention, trying to answer the following questions:

1. What are the effective variables (of each human resource dataset) representing effectiveness or efficiency in determining employee productivity of an organization?
2. To what extent has the proposed clustering approach been effective (in terms of quantitative and qualitative criteria) in automatic labeling to identify and classify productive employees at different levels and investigate their departure?

2. Literature Review

Related papers are reviewed in the hybrid clustering, auto labeling, and human resources and retention management subsections.

2.1. Hybrid Clustering

From providing a hybrid multivariate observations clustering method (a combination of K-means elements and single linkage clustering techniques) to discover high-density clusters in 1982 (Wong, 1982) to present three-stage clustering method for solving general clustering problems – including nonconvex clusters in 2019 (Amiri et al., 2019) – many researchers have used various hybrid clustering methods in different domains with diverse goals. K-means and K-medoids algorithms have been the most popular partition-based clustering algorithm for decades. Their combinations have been used in many studies. The K-MM algorithm combined these two in image clustering (Drias, Cherif et al., 2016) and web information foraging (Drias, Kechid et al., 2016) with efficient and effective results. Some studies have also proposed a new hybrid clustering method using the PAM and self-organizing map (SOM) approaches. Firstly, the dataset is clustered using the SOM algorithm, and then the clustering results are used in the PAM algorithm (Zhang et al., 2007). Other studies include combining three partition-based algorithms, namely PAM, clustering around large applications (CLARA), and clustering large applications based on randomized search (CLARANS) with k-medoid distance-based distance tracking, to improve the outlier detection and elimination process (Murugavel & Punithavalli, 2011). In addition, they combined a new genetic algorithm with K-means for determining the number of clusters automatically (Rahman & Islam, 2014). Hybrid clustering methods are used in various fields, including fuzzy modeling (Wong & Chen, 1999), medical disease diagnosis, and operating room planning (Santoso et al., 2017; Simić et al., 2018) strategic planning related to determining the best logistics for distribution centers (Simić et al., 2017), wireless sensor network data collection (Jung et al., 2009), ecosystem mapping (Tchuenté et al., 2011), resource assessment and agricultural development (He et al., 2010), accessing huge data improvement (Ebadati & Tabrizi, 2016), and many other topics. In the human resources field and related topics, including areas of employee retention and their training, studies have been also conducted using hybrid clustering techniques to analyze the turnover rate of technology professionals by performing artificial neural networks and SOM (Fan et al., 2012), and have evaluated staff profiles to predict their training requirements using a hybrid clustering and the optimization algorithm (Esmaeilzadeh et al., 2016). Moreover, they have used a hybrid data mining model with the K-medoids and C4.5 technique in classifying small businesses (SMEs) with business prospects, goals, development, and empowerment programs (Tosida et al., 2019).

2.2. Auto Labeling

Many automated labeling types of research have been done in the areas of text mining, web mining, and image processing in order to provide a supervised approach to automatically label the subject clusters of readers' opinions into online news (Aker et al., 2016) and multinomial topic models in text mining (Mei et al., 2007). They have also analyzed government agencies' textual information (Treeratpituk & Callan, 2006), public access to a corpus annotated with cyber-security entities (Bridges et al., 2013), and the acoustic-phonetic properties of speech (Tanaka et al., 1986). In addition, a semi-surveillance approach to improve the sequencing method in internet explorer through a class of algorithms with self-learned features (Qi et al., 2009), providing an integrated clustering framework for learning representation of images and cluster centers jointly based on a fully secure automated encoder (Li et al., 2018), conducting shape clustering to facilitate the automatic labeling of objects in a set of images (Yankov & Keogh, 2006), introducing a new and intuitive method of automatic labeling for single stroke primitives (Zhen et al., 2012), and providing semi-supervised fuzzy clustering approach for automatic labeling (de Abreu Lopes & de Arruda Camargo, 2012) are the critical issues addressed in web mining and image processing. With regard to numerical data labeling, studies have been carried out on the categorical data labeling, including a method for labeling categorical data using rough relative entropy to measure information uncertainty and applying cluster purity for distance detection (Reddy et al., 2013), an algorithm for labeling categorical data and analyzing its time complexity (Cao & Liang, 2011), and a mechanism called maximal resemblance labeling data for categorical data (Chen et al., 2005). Besides, it should be noted that only one study has dealt with the automatic labeling of numerical data using the chi-square analysis of human health resources (Kusumaningrum, 2017).

2.3. Human Resources and Retention Management

According to recent reviews about employee retention (Al-Emadi et al., 2015; Das & Baruah, 2013; Ramlall, 2004; Singh, 2019), the biggest challenge for companies is not only managing human resources but also retaining them, which is affected by employee satisfaction (Das & Baruah, 2013) and motivation (Ramlall, 2004). Having employees is an obligation to continue the company to achieve a competitive advantage (Walker, 2001; Zineldin, 2000). An organization's ability to maintain its employees depends entirely on managing them (Kaliprasad, 2006). In addition to management, several factors such as compensation (Kumar & Arora, 2012; Moncarz et al., 2009), reward (Alhmoud & Rjoub, 2019; Silbert, 2005), promotion (Eyster, 2008), participation in decision making (Khalid & Nawab, 2018; Noah, 2008), work-life balance (Hyman & Summers, 2004), work environment (Kundu & Lata, 2017; Ramlall, 2003), training and development (Diah et al., 2020; Handy, 2008), leadership (Fang et al., 2009; Rao et al., 2018), job-security (Rosenblatt & Ruvio, 1996), and economic, psychological, affiliation, and self-actualization (Kurdi & Alshurideh, 2020) affect employee retention. Patgar and Kumar (2015) identified the main factors of retention management strategies in companies, the most important of which was participation in management. Ribes et al. (2017), emphasizing that employee retention needs an in-depth turnover analysis, proposed a method to employee retention with machine learning techniques, and designed retention policies. Giri et al. (2019) used structural equation modeling (SEM) to analyze the impact of influential factors on employee retention. Elsafty and Ragheb (2020) investigated the role of HRM towards employee retention during the covid-19 pandemic in the medical supplies sector.

According to this review of the related literature, no research has been done using the K-means and PAM hybrid clustering method to determine the levels of employee productivity and turnover to retain them. The innovation of this research is in the automatic labeling of numerical data in the studied dataset related to HRM. The results of the algorithm are due to the high consistency with the initial interpretations of data exploration. Another innovation of this study is the evaluation of the model results, which provides a simple criterion for examining the differences. Moreover, the article processes the resulting image of clusters using the singular value decomposition (SVD) algorithm, leading to complete confidence in the model and its results.

3. Materials and Methods

3.1. CRISP-DM Methodology

The cross-industry standard process for data mining (CRISP-DM) is often known as the most common and popular process model and influential technique in data mining literature (Schröer et al., 2021). It is about 2 decades old and according to many surveys and user opinions, it is still the standard for developing data mining and knowledge discovery projects (Martínez-Plumed et al., 2019; Piatetsky, 2014). According to this approach, the life cycle of a data-mining project consists of six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al., 2000).

First step – business understanding: This introduces the business environment. Here, a problem is defined according to the needs and goals of the organization to be addressed by data mining techniques.

Second step – data understanding: Here, the first step is to collect data, after which the dataset is initially processed to complete data recognition and obtain initial insights from the data.

Third step – data preparation: This is the longest and most crucial step in the data mining process. Creating a table, selecting attributes and records according to the intended purpose, and converting and cleaning data for use in the model are among the tasks performed in this phase.

Fourth step – modeling: Various modeling techniques are selected and used for data analysis and knowledge extraction.

Fifth step – evaluation: In this step, the model results must be evaluated to ensure that they are in line with the goals defined for the project.

Sixth step – deployment: The focus of this step is on applying the knowledge gained in business processes to solve the business needs (Moslehi et al., 2018, 2019). These steps are illustrated in Figure 1 (Chapman et al., 2000).

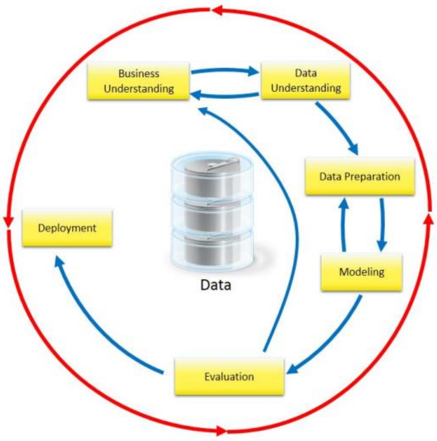

Figure 1. Crisp Data Mining Methodology

3.2. Clustering

The clustering technique is an unsupervised and process-based method that divides a heterogeneous dataset into homogeneous clusters (Ngai et al., 2009). K-means clustering and K-medoids are common and well-known clustering methods (Kaufman & Rousseeuw, 1990; MacQueen, 1967). Among the many algorithms available for K-medoids, the partition around medoids (PAM) method is one of the best ways for dealing with small data (Kaufman & Rousseeuw, 1990).

3.3. SVD

The SVD method can convert matrix (A) to three USV^T matrices, where U and V are two orthogonal matrices and S is a diagonal matrix (Hoecker & Kartvelishvili, 1995). It has different applications, including image processing (Cao, 2006).

4. Results

The steps of CRISP-DM used to direct this study are as follows.

4.1. Business Understanding

Identifying productive employees and their turnover is one of the most pressing problems in the HRM sector of an organization, which is affected by various variables. For solving this problem, employees should be divided into multiple levels of productivity. Then the type of their turnover is identified as functional or inefficient for performing good policies tailored to each level. The purpose of data mining is automatic labeling on staff by considering features, such as indicators of effectiveness and efficiency, as determinants of productivity, which is suggested as an innovative method and evaluated by new criteria. The steps for understanding business issues are summarized in Table 1.

Table 1. Steps of Business Understanding

Steps of business understanding		Description
Application	The problem	Productive employees' turnover
	Causes of problem	Valid variables in determining productivity
	Improvement	Reducing inefficient turnover and retain productive employees
Aim of business		Identifying and dividing productive employees into various productivity levels and determining the type of turnover for performing tailored policies.
Purpose of data mining		Automatic labeling of the staff by considering features such as indicators of effectiveness and efficiency as determinants of productivity.

4.2. Data Understanding

The present dataset was collected from www.kaggle.com[1], which presents the information of 14999 employees in a big company during 5 years based on the HRM unit report. The variables of this dataset and the characteristics of each variable are described in tables 2 and 3.

1. www.kaggle.com/jacksonchou/hr-analytics

Table 2 . Variables Definition

Variable name	Data type	Range (description)
Satisfaction level	Numeric	0 to 1
Last evaluation (yearly)	Numeric	0 to 1 (Performance evaluation)
Number project	Integer	2 to 7
Average monthly hours	Integer	96 to 310
Time spend company (in years)	Integer	2 to 10
Work accident (within the past 2 years)	Integer	0: Didn't do 1: Do
Left	Integer	0: didn't leave 1: left
Promotion-last-5 years	Integer	0: Didn't have 1: had
Department	Categorical	Accounting, HR, IT, Management, Marketing, Product-mng, Rand-D, Sales, Support, Technical
Salary	Categorical	Low, Medium, High

Reference: www.kaggle.com

Table 3. Characteristics of Variables

Variable name	Min.	1 st Qu.	Median	Mean	3 rd Qu.	Max.
Satisfaction level	0.09	0.44	0.64	0.61	0.82	1
Last evaluation	0.36	0.56	0.72	0.72	0.87	1
Number project	2	3	4	3.8	5	7
Average monthly hours	96	156	200	201	245	310
Time spend company	2	3	3	3.5	4	10
Work accident	0: 12830, 1: 2169					
Left	0: 11428, 1: 3571					
Promotion-last-5 years	0: 14680, 1: 319					
Department	Sales: 4140, Technical: 2720, Support: 2229, IT: 1227, Other: 4683					
Salary	High: 1237, Low: 7316, Medium: 6446					

4.3. Data Preparation

The data preprocessor consists of four steps, respectively:

1. Data cleaning in which missing values are cleaned or filled with appropriate values. Data were no cleaned here because there were no missing values and contradictions in the dataset.
2. Data reduction: features and duplicate rows are deleted or sampled. Given the dimension of the present dataset (14999*10), the duplicate rows were deleted and six features (satisfaction level, time spend company, work accident, promotion-last-5 years, department, salary) were eliminated based on the modeling aims. Finally, the dataset dimension was reduced to 11991*4.
3. Data integration: data from different places and sources are integrated, which did not apply to this study.
4. Data conversion: in this step, the type of features *last evaluation*, *average monthly hours*, *number project*, and *left* are changed into the category type, as explained below.

4.4. Data Exploration

Data exploration is one of the most critical steps that should be taken to fit the purpose of the research. Although there are many definitions of productivity, employees' productivity comes from the performance, average monthly hours, and the number of projects according to the present dataset features. As such, the performance and the number of projects are effectiveness (doing the right thing), and average monthly hours is efficiency (doing things right) because effectiveness is the kind of goal, and the purpose of the manager is to improve employee performance and maximize the number of projects. However, efficiency is a kind of resource (especially time and cost resources), where the more time employees spend the time, the more the efficiency decreases (Answer research question 1). Figures 2, 3, and 4 show the relationships of dataset features as effectiveness and efficiency variables with various levels of employees' productivity. Four levels of productivity as weak, medium, good, and excellent are shown in figures 2 and 3. In addition, Figure 4 shows the cubic space where the z-axis represents the average monthly hours, and the x and y axes represent the number of projects and the last evaluation, respectively. Blue circles are employees who left the organization, and red circles are those who stayed therein. As it is well known, the interpretation of figures is complex with increasing variables and their communications. Hence, it is necessary to present a method that can automate identifying and labeling, as explained in the modeling section.

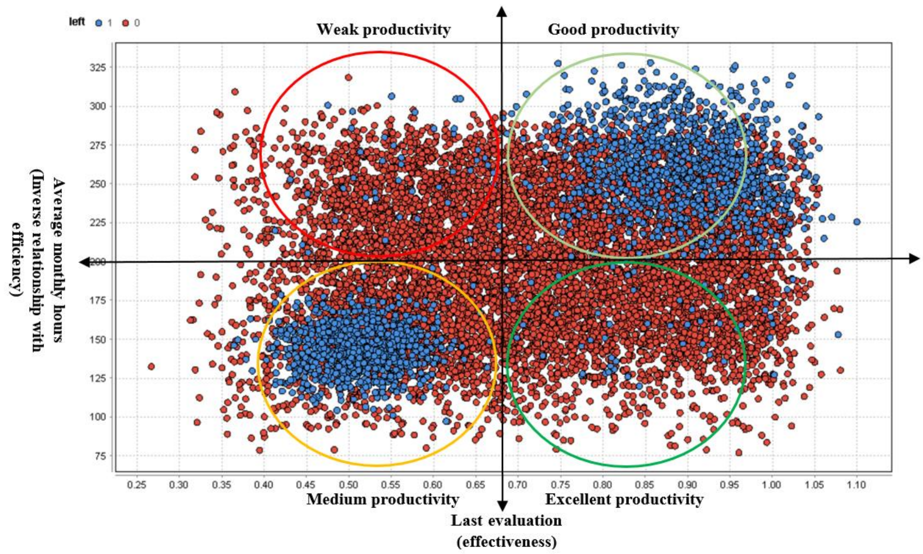

Figure 2. Investigating Employees' Turnover and Four Levels of Productivity Considering Average Monthly Hours as Efficiency and Last Evaluation as Effectiveness

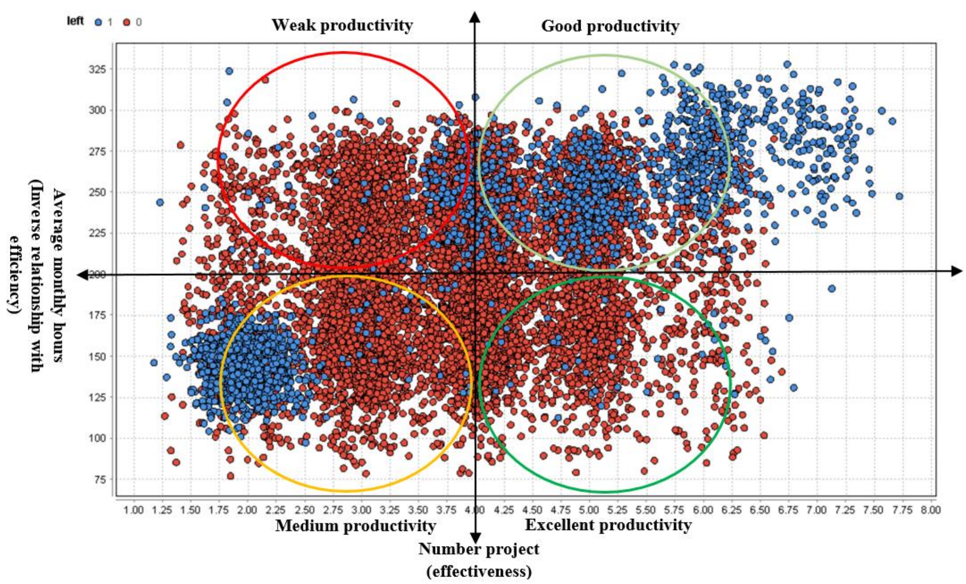

Figure 3. Investigating Employees' Turnover and Four Levels of Productivity Considering Average Monthly Hours as Efficiency and Number Project as Effectiveness

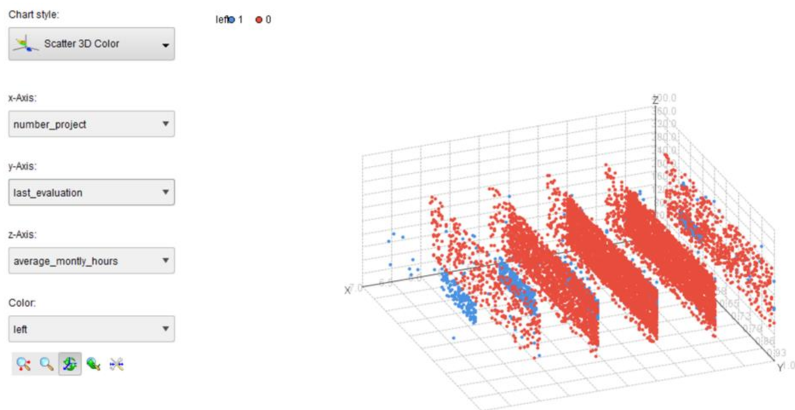

Figure 4. Investigating Employees' Turnover Considering the Communication of the Last Evaluation, Average Monthly Hours, and Number Project

4.5. Modeling

A hybrid clustering algorithm was used in the modeling discussion to separately cluster and then categorize the variables of the last evaluation and average monthly hours by K-means clustering (tables 4 and 5). Then, the two variables, along with those of project numbers and left, were implemented in the PAM clustering algorithm in the normalized state with Manhattan distance. The whole process was executed in the 'for' loop and the final model results would generate automatic labeling for the data set after repeated iterations and reaching the stopping condition, as shown in Figure 5 (recognition with model). The flowchart of the suggested method is plotted in Figure 6. The stopping conditions are related to how the results of the K-means clustering are located in each cluster center of the PAM algorithm.

Table 4. Clustering of Last Evaluation

Number of each cluster	Last evaluation	Labels
1	0.9179640	Excellent
2	0.6108240	Medium
3	0.4869621	Weak
4	0.7617411	Good

Table 5. Clustering of Average Monthly Hours

Number of each cluster	Average monthly hours	Labels
1	199.9763	Medium
2	145.8609	Low
3	255.9609	Top

```
Medoids:
        ID last_evaluation number_project average_montly_hours left
[1,]     1               3              1                    2    2
[2,]     2               1              4                    3    2
[3,] 11977               2              3                    1    1
[4,]  2049               3              3                    3    1
Clustering vector:
   [1] 1 2 2 2 1 1 2 2 2 1 1 2 2 1 1 1 1 2 1 2 2 1 2 1 1 2 2 1 1 1 2 1 1 1 2 2 1 1 2 1 2 1 1 2 2 1 1 1 1 1 2 1 1 1 2 1 2 2 1 2 2
  [62] 2 2 1 2 2 1 2 2 1 1 2 1 2 2 2 2 1 1 1 1 1 2 2 2 2 2 1 1 2 1 2 1 1 2 1 2 2 2 2 1 1 1 2 1 2 2 1 1 1 1 2 1 2 2 1 2 2 1 1 2 1
 [123] 2 2 1 2 1 1 2 1 2 1 2 1 1 2 2 1 2 2 2 1 2 1 2 1 2 2 2 1 2 1 1 1 1 2 2 2 1 1 2 2 1 2 2 2 1 2 1 1 2 1 2 2 1 1 1 1 1 2 1 2 2
 [184] 2 1 1 2 1 1 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1 2 2 1 1 1 1 2 2 1 1 1 1 2 1 2 2 1 1 2 1 2 1 1 2 2 1 2 2 1 2 2 1 1 1 2 2 1 1 1
 [245] 2 1 1 1 2 2 2 2 2 2 1 1 2 1 2 2 1 2 1 2 1 2 2 1 1 2 1 2 2 2 2 2 2 2 2 1 2 1 2 2 1 1 2 2 2 1 1 2 1 2 2 1 1 1 1 2 1 2 2 1 2
 [306] 1 1 2 1 2 2 2 1 1 2 2 1 1 2 1 1 2 1 2 1 1 1 2 2 1 1 2 2 1 2 1 2 1 1 2 2 1 2 1 1 2 2 2 2 1 2 1 2 1 2 1 1 1 2 2 2 2 1 2 1 2
 [367] 2 1 2 2 1 1 1 1 2 2 1 1 2 2 1 1 1 1 1 2 2 2 1 1 2 1 2 2 1 1 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 1 2 1 2 2 1 2 2 2 2 1 1 1 1 2 1
 [428] 1 1 2 1 1 2 2 1 2 1 2 1 1 1 2 2 1 2 1 2 1 2 2 2 2 1 2 1 2 2 1 1 1 1 1 2 2 1 1 2 2 1 2 1 2 1 2 1 2 2 1 2 1 2 2 1 1 2 1 2 2
 [489] 2 2 1 2 2 2 2 2 2 1 1 2 2 2 1 1 2 2 1 1 1 1 1 2 2 2 2 2 2 1 1 2 2 2 1 2 1 1 2 2 1 1 1 2 1 2 1 2 1 2 2 1 2 2 2 1 2 1 1 1 1
 [550] 2 2 2 2 1 2 1 1 1 2 2 1 1 2 1 1 2 2 1 1 1 2 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 1 2 1 1 1 2 1 2 2 2 2 1 2 2 2 2 2 2 2 1 2 1 2 2
 [611] 1 1 2 2 2 1 2 1 1 1 2 1 1 1 2 1 1 2 1 2 2 2 1 2 1 1 1 1 1 1 2 2 1 2 1 2 1 1 1 1 2 2 1 2 2 2 1 1 2 1 1 2 2 2 2 2 2 2 1 2 2
 [672] 1 1 1 1 2 2 1 2 2 1 2 1 1 2 2 2 2 2 1 1 2 2 1 2 2 2 1 1 1 1 1 1 1 1 2 2 2 1 2 1 2 2 1 1 1 2 2 2 1 2 1 1 1 1 1 1 1 2 1 2 1
 [733] 1 2 2 2 1 2 2 2 2 1 2 1 2 1 1 1 2 1 1 1 1 2 1 2 1 1 1 2 2 1 1 2 1 1 1 2 1 2 2 1 1 1 1 2 2 2 2 1 2 2 1 1 2 2 1 1 2 2 2 2 2
 [794] 2 1 1 2 2 1 2 1 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 1 1 2 2 1 1 1 2 2 2 1 1 2 2 1 1 2 2 1 2 2 1 1 2 2 2 1 1 2 2 2 2 1 1 2 1 2
 [855] 1 2 2 2 2 2 1 1 1 2 2 1 2 1 1 2 2 2 2 2 2 2 1 2 1 2 2 1 1 2 1 1 2 2 1 2 2 2 2 2 1 1 2 2 2 1 1 1 1 2 2 2 2 2 2 1 2 1 2 2 2
 [916] 1 2 2 1 1 1 1 1 2 2 1 2 1 2 2 1 2 2 1 2 2 2 1 2 1 2 2 2 2 1 1 2 2 2 1 2 1 1 2 1 1 1 2 1 2 1 1 1 2 2 2 1 1 1 2 2 1 2 1 2 2
 [977] 1 2 1 1 2 1 2 1 1 2 2 1 2 2 1 2 1 1 1 2 1 1 2 2
 [ reached getOption("max.print") -- omitted 10991 entries ]
Objective function:
   build     swap
1.593873 1.524541

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"  "clusinfo"   "silinfo"    "diss"       "call"
[10] "data"
```

Figure 5. The Final Result of the Clustering Detected by the Model and the Auto-Labeling Perfectly Match

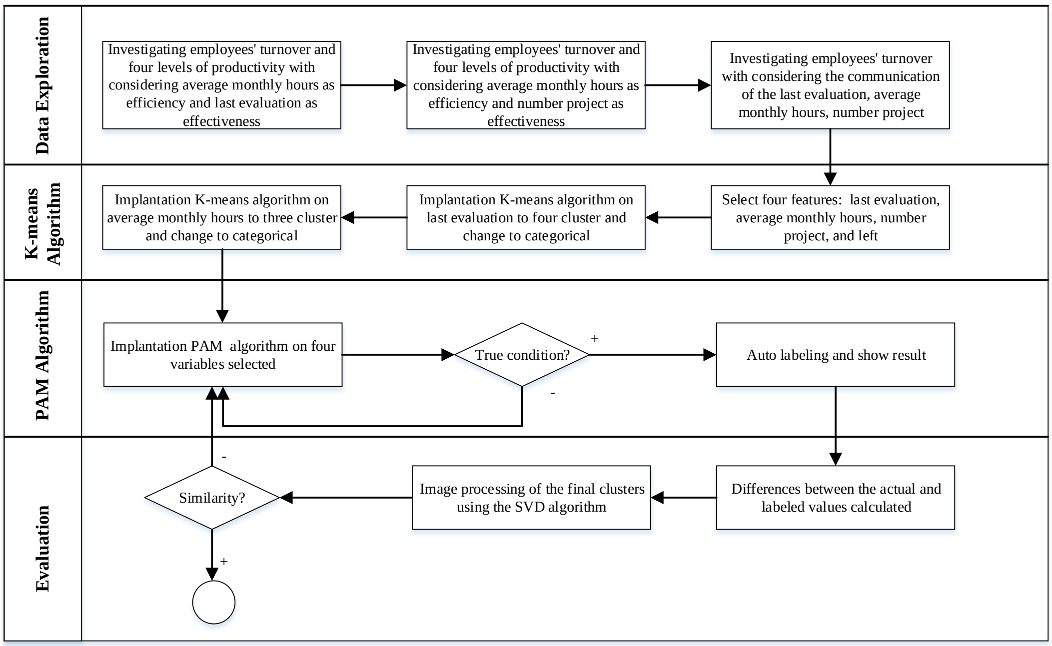

Figure 6. The Flowchart of the Method Suggestion

4.6. Evaluation and Analysis

This section first interprets the results of the algorithm implementation, which demonstrates a secure link between the data exploration and modeling sectors, and then evaluates the model with two criteria.

Table 6 shows the results of automatic labeling in identifying productive employees and their type of turnover, and presents a more straightforward and understandable interpretation of the results in Figure 5.

Table 6. Results Interpretation

Cluster number	Cluster size	Last evaluation		Number project		Effectiveness	Average monthly hours		Efficiency	Left		Productivity levels	Turnovers type
1	901	3	Weak	1	Weak	Low	2	Low	High	2	Leave	Medium	Inefficient
2	1090	1	Excellent	4	Excellent	High	3	High	Low	2	Leave	Good	Inefficient
3	4453	2	Medium	3	Medium	High	1	Medium	High	1	Stay	Excellent	Functional
4	5547	3	Weak	3	Weak	Low	3	High	Low	1	Stay	Weak	Inefficient

Based on Table 6, the clusters are interpreted as follows:

Cluster 1: Employees with the weakest performance, the lowest number of projects, and the lowest average monthly working time. The first two variables behave as low effectiveness and the last variable behaves as high efficiency, both of which represent medium productivity as identified by the model correctly based on the segmentation in the data exploration (figures 2, 3, and 4). This cluster is named medium productivity employees who left the organization, which is an inefficient turnover.

Cluster 2: Employees with the best performance, the highest number of projects, and the utmost average monthly working time. The first two variables behave as high effectiveness and the last variable behaves as low efficiency, both of which represent good productivity as identified by the model correctly based on the segmentation in the data exploration (figures 2, 3, and 4). This cluster is named good productivity employees who left the organization, which is an inefficient turnover.

Cluster 3: Employees whose number of projects and performance are above average, with an average monthly working time. The first two variables behave as high effectiveness and the last variable behaves as high efficiency, both of which represent excellent productivity as identified by the model correctly based on the segmentation in the data exploration (figures 2, 3, and 4). This cluster is named outstanding productivity employees who stay in the organization, which is a functional turnover.

Cluster 4: Employees whose performance is weak. The number of projects and their average monthly work time are above average. The first two variables behave as low effectiveness and the last variable behaves as low efficiency both of which represent poor productivity as the model correctly identified based on the segmentation in the data exploration (figures 2, 3, and 4). This cluster is named reduced productivity employees who stay in the organization, which is an inefficient turnover.

Two criteria are used to evaluate the model (and answer research question 2):

First criterion: The labeled clusters aim to examine the employees who left the organization. Clusters 1 and 2 show good and medium productive employees who left the organization. Accordingly, a criterion is proposed to verify the cluster accuracy based on the differences between the actual and labeled values (cluster size). Tables 7 and 8 show the number of left variables for good and medium productivities by applying filters to the primary data set. Table 9 shows the total values of the left variable. In tables 7 and 8, the leaving amounts cover 93% of the values in Table 9. The difference between the actual and labeled values for clusters 1 and 2 is obtained based on the leaving numbers in tables 7 and 8. In addition, the size of clusters in Table 6 is calculated for the twice run of the model in Table 10. The results of both running times show auto labeling, but the second time is better than the first one.

Table 7. Initial Data Filter Based on Good Productivity

Limitations	Number of left=1 (leaving)	Number of left=0 (staying)
Average monthly hours>=200 Number project>=4 Last evaluation>=0.7	1015	1699

Table 8. Initial Data Filter Based on Medium Productivity

Limitations	Number of left=1 (leaving)	Number of left=0 (staying)
Average monthly hours<200 Number project<4 Last evaluation<0.7	838	1204

Table 9. The Total Values of the Left Variable in the Initial Data Set

Variable name	Number of 1 (leaving)	Number of 0 (staying)
Left	1991	10000

Table 10. First Evaluation Criterion (Difference of Labeled Values With Real Values)

Model implementation	The result of subtraction	
	Cluster 1	Cluster 2
1	124	262
2	114	252

Second criterion: The second criterion for evaluating the model is the image processing of clusters resulting from clustering using the algorithm SVD as an innovative criterion.

In Figure 7, two plots inside the left scenario show the results of the proposed method as automatic labeling, and the plot inside the right scenario shows that labeling is not automatic.

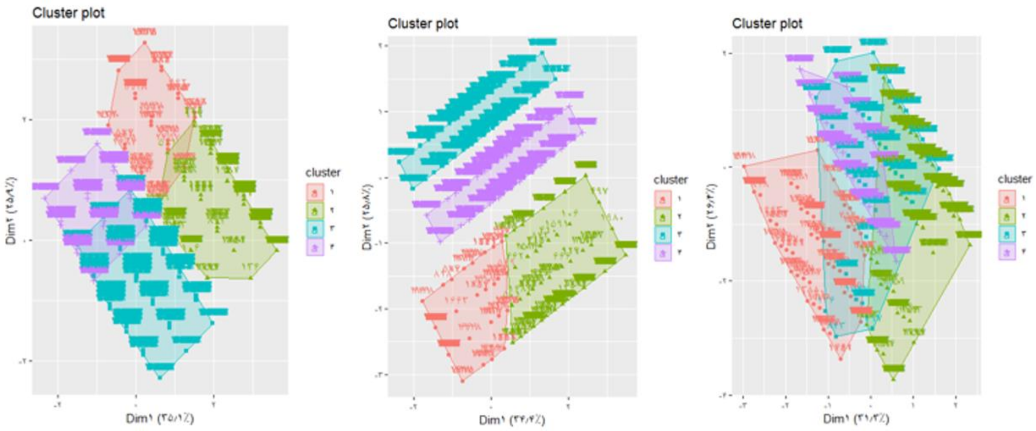

Figure 7. Comparison of Cluster Plots in Automatic Labeling Mode and Other Than This

According to the SVD algorithm, each shape is transformed into a matrix – and is called ‘A’ for example. These matrices decompose into three USV^T matrices for each image. The singular values as outputs from the SVD algorithm are sorted in descending order. Hence, the first columns of U (left singular vectors) and V (right singular vectors) give the main structure of the image, which is the difference of the first two-element of matrices in Figure 8 to confirm this.

```
>>> D
Array ([1.13065155e+05, 5.83896910e+03, 3.81860955e+03, 3.34088747e+03,
        3.10145231e+03, 2.82908745e+03, 2.63195235e+03, ………………...,
        1.12606859e-11, 1.12606859e-11, 1.12606859e-11, 1.12606859e-11,
        6.57435422e-12])
```

Figure 8. The Outputs From the SVD Algorithm That Sorted in Descending Order

Figures 9, 10, and 11 are the signals from the first columns of U (first row of matrix A) and the first column of V (matrix column A) as the main SVD output components from cluster plots of Figure 7. The signals and oscillations in the U indicate the difference of cluster plots that have been automatically labeled with other cluster plots.

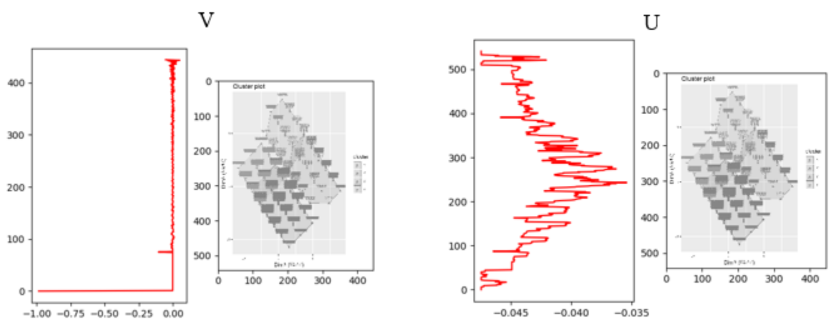

Figure 9. The Signals From the Cluster Plot Indicate Automatic Labeling

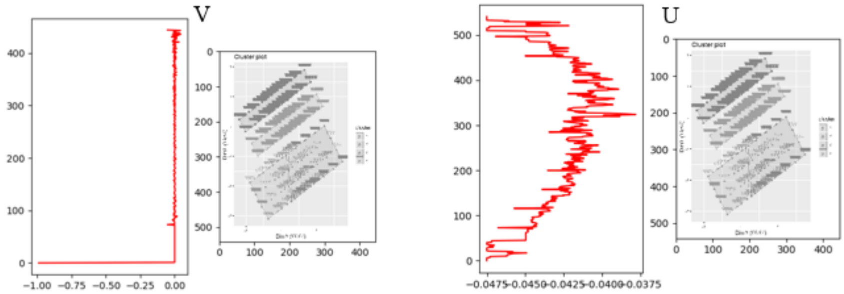

Figure 10. The Signals From the Cluster Plot Indicate Automatic Labeling

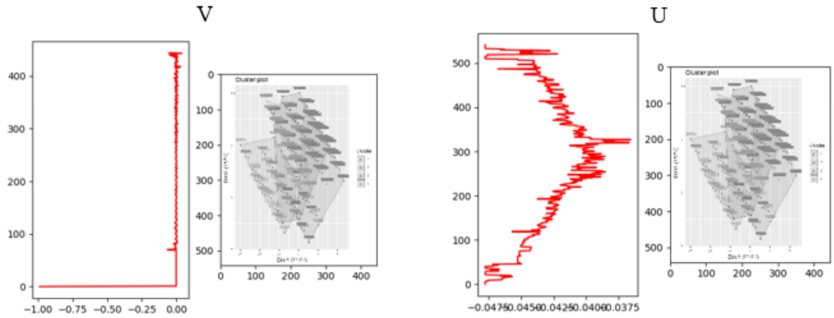

Figure 11. The Signals From the Cluster Plot do not Indicate Automatic Labeling

Figures 9 and 10 are similarly based on the fluctuations (ascent and drop) of the U signals, while Figure 11 is not, and this indicates the correct performance of the proposed criterion.

It is also possible to compare multiple vectors of the shapes for better comparison.

5. Discussion

The main findings of study in Table 6 verify theories related to human resources and productivity studies for practice. Theories related to various forms of employee turnover (functional or inefficient) (Chalkiti & Sigala, 2010) and their relationship with productivity and the concepts of effectiveness and efficiency (Ilmakunnas et al., 2005; Sexton et al., 2005) were discussed in Introduction. Distinctive columns of Table 6 illustrate the relationship between model results and previous theories so that the analysis of the performance and number of projects determine levels of effectiveness, and average of monthly hours shows the levels of efficiency. The levels of productivity are the result of effectiveness and efficiency (the answer to research question 1). Finally, the type of employee turnover (i.e., functional and inefficient) is concluded from analyzing productivity levels and left variables.

The practical implications of these results are (a) quantitative performance evaluation that reduces human involvement and mental judgments in qualitative performance assessment, (b) variable payments such as bonuses to increase employees motivation, (c) identifying employees with functional turnover who have stayed in the company (cluster 3), which leads to (d) their career development, (e) succession management, and (f) the selection of senior managers among them.

Table 11 compares labeling accuracy and time of the proposed method with previous studies in human resources fields. This table shows that the proposed method performed better than previous studies methods in hybrid clustering and auto labeling (thus providing the answer to research question 2).

Table 11. Comparison of the Proposed Method With Previous Studies Based on Labeling Accuracy and Time

Method		Labeling accuracy	Time
Proposed method		93 %	65.58s
Hybrid clustering			
Fan et al., 2012	K-means	63.5 %	-
	BPN	87.2 %	-
	SOM+BPN	92 %	-
Tosida et al., 2019	k-medoids–C45	71.8 %	-
	k-medoids–C5	62.95 %	-
Esmaeilzadeh et al., 2016	ICA	-	1282s
	FICA	-	1129s
	PSO	-	1071s
	GA	-	1148s
Auto labeling			
Kusumaningrum, 2017		89.14 %	-

According to tables 6 and 11, the proposed approach – with its high accuracy of automatic labeling compared to previous methods – has led to the identification of functional and inefficient employee turnover at different levels of productivity. In addition, the results of the two tables verify the aims of management and data mining in Table 1.

6. Conclusion and Suggestions

The workforce is a determining factor in organizational productivity. Identifying and categorizing productive employees at different levels of productivity, then analyzing their exit from the company, and their type of turnover are essential issues in human resources that can be examined well by data mining tools, which are mentioned in this paper. In this study, various stages of data-based decision-making in HRM are described according to the CRISP-DM. Accordingly, features related to the concepts of efficiency and effectiveness were identified in the data exploration section, and their relationship explored the concept of productivity. The modeling section presented the automatic labeling of numerical data by a hybrid of K-means and PAM clustering algorithms. The difference between the actual and labeled values was calculated to evaluate the model; In addition, image processing was used on plot clusters to evaluate this method. The clusters labeled by the model were automatically named as four clusters, and policies were suggested for each cluster for retention and improved productivity.

Cluster 1 indicates medium productive employees who left the organization and had an inefficient turnover. Cluster 2 denoted good productive employees who left the organization and had an inefficient turnover. Cluster 3 represents excellent productive employees who stay in the organization and have a functional turnover. Cluster 4 is indicative of poor productive employees who remain in the organization and have an inefficient turnover. Productivity improvement and retention policies were applied for clusters 1, 2, and 3, and replacement policies were used for cluster 4, along with the consideration of other variables. These policies improve inefficient turnover and convert it into a functional one. The improvement of succession and bonus management, reduction of human involvement, and mental judgments are practical implications of these changes.

Automatic labeling of numerical data, detecting it by the model, and minimizing human interference can revolutionize clustering. The steps taken in this research in the HRM field on a given dataset can hopefully be used for bright days of this topic. In future work, a comprehensive algorithm can be of interest that can be executed on valuable and relevant datasets.

It should be noted that access to more diverse data from a wide range of organizations that might lead to obtaining profound findings is one of the limitations of the present study. In addition, the organized records of why employees have left, work experience, and demographic characteristics such as education, age, gender, etc., could be helpful and lead to a more comprehensive analysis.

References

Abtahi, S., & Kazemi, B. (2001). *Productivity* (3rd ed.). Institute For Trade Studies and Research (In Persian).

Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *Algorithms*, *4*(5), 22-26.

Aker, A., Paramita, M. L., Kurtić, E., Funk, A., Barker, E., Hepple, M., & Gaizauskas, R. (2016). Automatic label generation for news comment clusters. In Isard, A., Rieser, V., & Gkatzia, D. (Eds.), *Proceedings of 9th International Natural Language Generation Conference (INLG 2016)*, (pp. 61-69). Edinburgh, UK: Association for Computational Linguistics.

Al-Emadi, A. A. Q., Schwabenland, C., & Wei, Q. (2015). The vital role of employee retention in human resource management: A literature review. *IUP Journal of Organizational Behavior*, *14*(3), 7-32.

Alhmoud, A., & Rjoub, H. (2019). Total rewards and employee retention in a Middle Eastern context. *SAGE Open*, *9*(2), 2158244019840118.

Amiri, S., Clarke, B. S., Clarke, J. L., & Koepke, H. (2019). A general hybrid clustering technique. *Journal of Computational and Graphical Statistics*, *28*(3), 540-551.

Anaraki, S. A. M., Haeri, A., & Moslehi, F. (2021). A hybrid reciprocal model of PCA and K-means with an innovative approach of considering sub-datasets for the improvement of K-means initialization and step-by-step labeling to create clusters with high interpretability. *Pattern Analysis and Applications*, 24(3), 1387-1402.

Azar, A., Ahmadi, P., & Sabt, M. V. (2010). Model design for personnel selection with data mining approach (Case study: A Commerce Bank of Iran). *Journal of Information Technology Management*, *2*(4), 3-22. (In Persian).

Berman, E. M., & Berman, E. M. (1998). *Productivity in public and nonprofit organizations: Strategies and techniques*. Sage.

Bertschek, I., Fryges, H., & Kaiser, U. (2006). B2B or Not to Be: Does B2B e-commerce increase labour productivity? *International Journal of the Economics of Business*, *13*(3), 387-405.

Bridges, R. A., Jones, C. L., Iannacone, M. D., Testa, K. M., & Goodall, J. R. (2013). Automatic labeling for entity extraction in cyber security. *ArXiv*, abs/1308.4941.

Cao, F., & Liang, J. (2011). A data labeling method for clustering categorical data. *Expert Systems With Applications*, *38*(3), 2381-2385.

Cao, L. (2006). Singular value decomposition applied to digital image processing. *Division of Computing Studies, Arizona State University Polytechnic Campus, Mesa, Arizona State University Polytechnic Campus*, 1-15.

Celebi, M. E. (2014). *Partitional clustering algorithms*. Springer.

Chalkiti, K., & Sigala, M. (2010). Staff turnover in the Greek tourism industry: A comparison between insular and peninsular regions. *International Journal of Contemporary Hospitality Management*, *22*(3), 335-359.

Chanodkar, A., Changle, R., & Mamtani, D. (2020). Prediction of employee turnover in organizations using machine learning algorithms. *Prestige International Journal of Management and Research*, *12*(1/2), 222-226.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc.

Chen, H. L., Chuang, K. T., & Chen, M. S. (2005, November). Labeling unclustered categorical data into clusters based on the important attribute values. In Han, J., Wah, B. W., Raghavan V., Wu, X., & Rastogi, R. (Eds.), *Fifth IEEE international conference on data mining (ICDM'05)* (pp. 1-8). Houston, TX, USA: IEEE.

Clarke, R. L. (1991). The measurement of physical distribution productivity: South Carolina, a case in point. *Transportation Journal*, 31(1), 14-21.

Das, B. L., & Baruah, M. (2013). Employee retention: A review of literature. *Journal of Business and Management*, *14*(2), 8-16.

Datta, S., & Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, *19*(4), 459-466.

de Abreu Lopes, P., & de Arruda Camargo, H. (2012, August). Automatic labeling by means of semi-supervised fuzzy clustering as a boosting mechanism in the generation of fuzzy rules. In Zhang, C., Joshi, J., Bertino, E., Thuraisingham, B. (Eds.), *2012 IEEE 13th International conference on information reuse & integration (IRI)* (pp. 279-286). Las Vegas, NV, USA: IEEE.

Diah, A. M., Hasiara, R. L. O., & Irwan, M. (2020). Employee retention of pharmaceutical firms in Indonesia: Taking investment in employee development and social and economic exchange as predictors. *Systematic Reviews in Pharmacy, 11*(1), 564-572.

Drias, H., Cherif, N. F., & Kechid, A. (2016). K-MM: A hybrid clustering algorithm based on k-means and k-medoids. In Pillay N., Engelbrecht A., Abraham A., du Plessis M., Snášel V., & Muda A. (Eds.), *Advances in nature and biologically inspired computing* (pp. 37-48). Cham: Springer.

Drias, H., Kechid, A., & Fodil-Cherif, N. (2016). A hybrid clustering algorithm and web information foraging. *International Journal of Hybrid Intelligent Systems*, *13*(3-4), 137-149.

Ebadati, E. O. M., & Tabrizi, M. M. (2016). A hybrid clustering technique to improve big data accessibility based on machine learning approaches. In Satapathy S., Mandal J., Udgata S., & Bhateja V. (Eds.), *Information Systems Design and Intelligent Applications* (pp. 413-423). New Delhi: Springer.

Elsafty, A. S., & Ragheb, M. (2020). The role of human resource management towards employees retention during Covid-19 pandemic in medical supplies sector - Egypt. *Business and Management Studies, 6*(2), 50-59.

Esmaeilzadeh, M., Abdollahi, B., Ganjali, A., & Hasanpoor, A. (2016). Evaluation of employee profiles using a hybrid clustering and optimization model. *International Journal of Intelligent Computing and Cybernetics*, 9(3), 218-236.

Eyster, L. (2008). *Current strategies to employ and retain older workers*. Urban Institute.

Faghihi, A., & Mousavi kashi, Z. (2010). A model for the measurement of productivity in the public sector of Iran. *Journal of Public Administration*, 2(4), 107-126. (In Persian).

Fan, C. Y., Fan, P. S., Chan, T. Y., & Chang, S. H. (2012). Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. *Expert Systems with Applications, 39*(10), 8844-8851.

Fang, C.-H., Chang, S.-T., & Chen, G.-L. (2009). applying structural equation model to study of the relationship model among leadership style, satisfaction, organization commitment and performance in hospital industry [paper presentation]. *The 2009 international conference on e-business and information system security*, Wuhan, China: IEEE.

Friedman, T. L. (2006). *The world is flat: The globalized world in the twenty-first century*. Penguin London.

Giri, A., Gangopadhyay, S., Majumder, J., & Paul, P. (2019). model development for employee retention in indian construction industry using structural equation modeling (SEM). *International Journal of Civil Engineering and Technology*, 10(04), 196-204.

Hajiheydari, N., Khabiri, S. H., & Talafi Daryani, M. (2017). A framework for data mining approach applications in human resource management. *Iranian Journal of Management Sciences, 12*(47), 21-50. (In Persian).

Handy, L. A. W. (2008). *The importance of the work environment variables on the transfer of training*: North Carolina State University.

He, Z., Zhang, R., & Wu, D. (2010). Evaluation of resource-saving and environment-friendly agriculture development status based on hybrid clustering. In Zhang, J., Xu, L., Zhang, X., & Yi, P. (Eds.), *ICLEM 2010: Logistics For Sustained Economic Development: Infrastructure, Information, Integration* (pp. 1962-1971). Chengdu, China: ASCE.

Hoecker, A., & Kartvelishvili, V. (1995). SVD approach to data unfolding. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 372(3), 469-481.

Hyman, J., & Summers, J. (2004). Lacking balance? Work-life employment practices in the modern economy. *Personnel Review*, 33(4), 418-429.

Ilgen, D. R., & Klein, H. J. (1988). Individual motivation and performance: Cognitive influences on effort and choice. In J. P. Campbell, R. J. Campbell, et al. (Eds.), *Productivity in Organizations:*

New Perspectives from Industrial and Organizational Psychology, (pp. 143-176), London: Jossey-Bass.

Ilmakunnas, P., Maliranta, M., & Vainiomäki, J. (2005). Worker turnover and productivity growth. *Applied Economics Letters*, *12*(7), 395-398.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, *31*(3), 264-323.

Jung, W. S., Lim, K. W., Ko, Y. B., & Park, S. J. (2009, February). A hybrid approach for clustering-based data aggregation in wireless sensor networks. In Takahashi, Y., Berntzen, L., & Smedberg, Å. (Eds.), *The third international conference on digital society* (pp. 112-117). Cancun, Mexico: IEEE.

Kaliprasad, M. (2006). The human factor I: Attracting, retaining, and motivating capable people. *Cost Engineering*, *48*(6), 20-26.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Khalid, K., & Nawab, S. (2018). Employee participation and employee retention in view of compensation. *SAGE Open*, *8*(4), 2158244018810067.

Kumar, R., & Arora, R. (2012). Determinants of talent retention in BPO industry. *Indian Journal of Industrial Relations*, 48(2), 259-273.

Kundu, S. C., & Lata, K. (2017). Effects of supportive work environment on employee retention: Mediating role of organizational engagement. *International Journal of Organizational Analysis*, 25(4), 703-722.

Kurdi, B., & Alshurideh, M. (2020). Employee retention and organizational performance: Evidence from banking industry. *Management Science Letters*, *10*(16), 3981-3990.

Kusumaningrum, R. (2017, January). An automatic labeling of K-means clusters based on chi-square value. *Journal of Physics: Conference Series*, *801*, 012071.

Li, F., Qiao, H., & Zhang, B. (2018). Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognition*, *83*, 161-173.

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M., & Neyman, J. (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (vol. 1, no. 14, pp. 281-297), California: University of California Press.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., … & Flach, P. A. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories [paper presentation]. *IEEE Transactions on Knowledge and Data Engineering*, 33(8), 3048-3061.

Mei, Q., Shen, X., & Zhai, C. (2007, August). Automatic labeling of multinomial topic models. In Berkhin, P., Caruana, R., Wu, X. (Eds.), *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 490-499). New York, NY, USA: Association for Computing Machinery.

Moncarz, E., Zhao, J., & Kay, C. (2009). An exploratory study of US lodging properties' organizational practices on employee turnover and retention. *International Journal of Contemporary Hospitality Management*, 21(4), 437-458.

Moslehi, F., Haeri, A., & Gholamian, M. R. (2019). Investigation of effective factors in expanding electronic payment in Iran using data mining techniques. *Journal of Industrial and Systems Engineering*, *12*(2), 61-94.

Moslehi, F., Haeri, A. R., & Moini, A. R. (2018). Analyzing and investigating the use of electronic payment tools in Iran using data mining techniques. *Journal of AI and Data Mining*, *6*(2), 417-437.

Murugavel, P., & Punithavalli, M. (2011). Improved hybrid clustering and distance-based technique for outlier removal. *International Journal on Computer Science and Engineering (IJCSE)*, *3*(1), 333-339.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592-2602.

Noah, Y. (2008). A study of worker participation in management decision making within selected establishments in Lagos, Nigeria. *Journal of Social Sciences, 17*(1), 31-39.

Okoye, P., & Ezejiofor, R. A. (2013). The effect of human resources development on organizational productivity. *International Journal of Academic Research in Business and Social Sciences, 3*(10), 250-268.

Patgar, S., & Kumar, V. (2015). A study on the factors affecting employee retention in a textile industry. *International Journal of Recent Research in Civil and Mechanical Engineering*, *1*(2), 1-5.

Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDD News.* https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

Prichard, R. (1990). *Measuring and improving organizational productivity*. Praeger Publishers.

Pritchard, R. D. (1992). Organizational productivity. In Dunnette, M. D., & Hough, L. M. (Eds.), *Handbook of industrial and organizational psychology* (vol. 3, pp. 443-471). Palo Alto, CA: Consulting Psychologists Press.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

Qi, Y., Kuksa, P., Collobert, R., Sadamasa, K., Kavukcuoglu, K., & Weston, J. (2009, December). Semi-supervised sequence labeling with self-learned features. In Wang, W., Kargupta, H., Ranka, S., Yu, P. S., & Wu, X. (Eds.), *The ninth IEEE international conference on data mining* (pp. 428-437). Miami Beach, FL, USA: IEEE.

Rahman, M. A., & Islam, M. Z. (2014). A hybrid clustering technique combining a novel genetic algorithm with K-Means. *Knowledge-Based Systems*, *71,* 345-365.

Rahmati, M. H., Hosseini Fard, S. M., & Alimadadi, A. (2014). Investigating effectiveness of in-service training in the public sector. *Iranian Journal of Management Studies*, *7*(2), 305-327.

Ramlall, S. (2003). Managing employee retention as a strategy for increasing organizational competitiveness: Organizational Application. *Applied HRM Research, 8*(2), 63-72.

Ramlall, S. (2004). A review of employee motivation theories and their implications for employee retention within organizations. *Journal of American Academy of Business*, *5*(1/2), 52-63.

Rao, D. K. S., Sahyaja, C., Akhil, P., & Narasimha, N. L. (2018). Role of leadership on employee retention–A study on corporate hospitals. *International Journal of Mechanical Engineering and Technology*, *9*(2), 161-172.

Reddy, H. V., Agrawal, P., & Raju, S. V. (2013, August). Data labeling method based on cluster purity using relative rough entropy for categorical data clustering. In Manjunath, S., Kumar, V., & Nagasundara, K. B. (Eds.), *The 2013 international conference on advances in computing, communications and informatics (ICACCI)* (pp. 500-506). Mysore, India: IEEE.

Ribes, E., Touahri, K., & Perthame, B. (2017). Employee turnover prediction and retention policies design: A case study. *arXiv e-prints,* https://ui.adsabs.harvard.edu/abs/2017arXiv170701377R.

Roiger, R. J. (2017). *Data mining: A tutorial-based primer*. CRC press.

Rosenblatt, Z., & Ruvio, A. (1996). A test of a multidimensional model of job insecurity: The case of Israeli teachers. *Journal of Organizational Behavior*, *17*(S1), 587-605.

Salem, H. (2003). *Organizational performance management and measurement.* United Nations Economic and Social Council.

Santoso, L. W., Sinawan, A. A., Wijaya, A. R., Sudiarso, A., Masruroh, N. A., & Herliansyah, M. K. (2017, November). Operating room scheduling using hybrid clustering priority rule and genetic algorithm. In Sutopo, W., Jauhari, W. A., Nor, F. M., & Kurniawan, D. (Eds.), *AIP conference proceedings* (Vol. 1902, No. 1, pp. 20-32). Miri, Malaysia: AIP Publishing LLC.

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999-2006.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526-534.

Sexton, R. S., McMurtrey, S., Michalopoulos, J. O., & Smith, A. M. (2005). Employee turnover: A neural network solution. *Computers & Operations Research*, *32*(10), 2635-2651.

Shaker Ardakani, M., Abzari, M., Shaemi, A., & Fathi, S. (2016). Diversity management and human resources productivity: Mediating effects of perceived organizational attractiveness, organizational justice and social identity in Isfahan's steel industry. *Iranian Journal of Management Studies, 9*(2), 407-432.

Silbert, L. (2005). *The effect of tangible rewards on perceived organizational support* (unpublished Master's thesis). University of Waterloo.

Simić, D., Ilin, V., Svirčević, V., & Simić, S. (2017). A hybrid clustering and ranking method for best positioned logistics distribution centre in Balkan Peninsula. *Logic Journal of the IGPL, 25*(6), 991-1005.

Simić, S., Banković, Z., Simić, D., & Simić, S. D. (2018, June). A hybrid clustering approach for diagnosing medical diseases. In de Cos Juez F. et al. (Eds.), *International conference on hybrid artificial intelligence systems* (pp. 741-752). Cham: Springer.

Singh, D. (2019). A literature review on employee retention with focus on recent trends. *International Journal of Scientific Research in Science and Technology, 6*(1), 425-431.

Storey, D. J. (2016). *Understanding the small business sector*. Routledge.

Tajeddini, K. (2015). Exploring the antecedents of effectiveness and efficiency. *International Journal of Hospitality Management*, *49*, 125-135.

Tanaka, K., Hayamizu, S., & Ohta, K. (1986, April). A demiphoneme network representation of speech and automatic labeling techniques for speech data base construction. In FUJISAKI, H. et al. (Eds.), *ICASSP'86. IEEE international conference on acoustics, speech, and signal processing* (Vol. 11, pp. 309-312). Tokyo, Japan: IEEE.

Tchuenté, A. T. K., De Jong, S. M., Roujean, J. L., Favier, C., & Mering, C. (2011). Ecosystem mapping at the African continent scale using a hybrid clustering approach based on 1-km resolution multi-annual data from SPOT/VEGETATION. *Remote Sensing of Environment*, *115*(2), 452-464.

Tosida, E. T., Andria, F., Wahyudin, I., Widianto, R., Ganda, M., & Lathif, R. R. (2019, July). A hybrid data mining model for Indonesian telematics SMEs empowerment. In Rindengan, A. J., Montolalu, C. E. J. C., Nainggolan, N., Tumilaar, R., & Bakhtiar, T. (Eds.), *IOP conference series: Materials science and engineering* (vol. 567, No. 1, pp. 1-12). Manado, Indonesia: IOP Publishing.

Treeratpituk, P., & Callan, J. (2006, May). Automatically labeling hierarchical clusters. In Fortes, J. A. B., & Macintosh, A. (Eds.), *Proceedings of the 2006 international conference on digital government research* (pp. 167-176). San Diego, California, USA: Digital Government Society of North America.

Vasantham, T., & Swarnalatha, C. (2015). Importance of employee retention. *International Journal of Research in Finance and Marketing*, *5*(8), 7-9.

Walker, J. (2001). Perspectives of human resource planning. *Journal of Management, 24*(1), 6-10.

Wong, C. C., & Chen, C. C. (1999). A hybrid clustering and gradient descent approach for fuzzy modeling. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 29*(6), 686-693.

Wong, M. A. (1982). A hybrid clustering method for identifying high-density clusters. *Journal of the American Statistical Association*, *77*(4), 841-847.

Yankov, D., & Keogh, E. (2006, December). Manifold clustering of shapes. In Clifton, C. W., Zhong, N., Liu, J., Wah, B. W., & Wu, X. (Eds.), *Sixth international conference on data mining (ICDM'06)* (pp. 1167-1171). Hong Kong, China: IEEE.

Zhang, Z., Wang, S. Z., & Zhang, Y. (2007). New clustering method based on hybrid of SOM and PAM. *Jisuanji Yingyong/ Journal of Computer Applications*, *27*(6), 1400-1402.

Zhen, J., Blagojevic, R., & Plimmer, B. (2012, June). Automated labeling of ink stroke data. In Kara, L. B., & Singh, K. (Eds.), *Proceedings of the international symposium on sketch-based interfaces and modeling* (pp. 67-75). Goslar, Germany: The Eurographics Association.

Zineldin, M. (2000). *TRM: total relationship management*. Studentlitteratur.