Journal of Theoretical and Applied Information Technology
15th September 2022. Vol.100. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TOWARDS A NEW CLASSIFICATION METHOD OF TEAM PERFORMANCE USING CLUSTERING AND FEATURES REDUCTION TECHNIQUES

[1] ZBAKH MOURAD, [2]AKNIN NOURA, [3] CHRAYAH MOHAMED, [4] ELKADIRI KAMAL EDDINE

[1]FS, Abdelmalek Essaadi University, Tetouan, Morocco

[2]FS, Abdelmalek Essaadi University, Tetouan, Morocco

[3]ENSATE, Abdelmalek Essaadi University, Tetouan, Morocco

[4]FS, Abdelmalek Essaadi University, Tetouan, Morocco

E-mail: [1]mourad.zbakh@etu.uae.ac.ma, [2]noura.aknin@uae.ac.ma, [3]chrayah@gmail.com, [4]kelkadiri@uae.ac.ma

ABSTRACT

The human factor is becoming more and more decisive in making the company more efficient and more competitive, improving the performance of their teams represents a challenge for the Human Resources department, which is also experiencing a profound digital transformation of data and its management [1].

Traditional Human Resources tools are no longer effective at managing skills, On the one hand, the performance evaluation of these resources objectively becomes more and more one of the very complex tasks of a manager, especially with the mass of current data presented by new non-traditional sources, on the other hand, the departure of key skills is a phenomenon that is not predictable by current HRIS tools, the financial cost is high as well as the technical loss of knowledge and know-how, and flexibility that this presents.

In this article, we will propose a new approach to the classification of teams according to several performance indicators. This method is based on the K-means algorithm to classify the members of a team, assessed against performance indicators linked to some soft and technical skills. The result of this work represents a decision support model for managers to develop a team adapted to the overall mission, to adapt its management style to each cluster, and to prepare future hires to compensate for the skill gaps of the team in place.

Keywords: *Human Resources Management, Key Performance Indicators, K-Means, Cluster Algorithm.*

1. INTRODUCTION

Developing and sustaining high-performance work teams has always been a primary concern of companies in their quest to be a leader in their fields of activity, to increase their market share, and to increase profit margins.

Indeed, how companies mobilize their teams is a key element in performance improvement. According to Barney's article published in 1991 [2], Resource management theory first links the development of certain specific skills to team performance, the study argues that organizations should develop skills that are Valuable, Rare, Inimitable, and Non-substitutable called "VRIN" criteria to be more competitive by doing things differently. According to the same study, companies must put considerable effort into identifying, understanding, and classifying essential skills.

For decades, companies have constantly sought new ways to use IT technologies to develop the skills of their teams. The human resources department is one of the most relevant application sectors for Big Data and machine learning, which represents an opportunity to transform the way companies manage their teams. Companies must spend a large part of their budget dedicated to Human Resources technologies to achieve their goals [3]. Improving performance is one of the challenges of the application on human resource

management that we will develop through this work.

Nevertheless, many empirical studies on factors affecting team performance have been published, the effectiveness of training is an example of improving the skills of individuals (Arthur, Bennett, Edens, & Bell, 2003 [4] of the performance at work (Aguinis & Kraiger, 2009) [5], and to increase the overall performance of organizations ([AASanchez, MIBAragon & RSValle, 2003) [6].

But these studies focused on a single factor or a maximum of three factors. Studies analyzing the relationships between individuals' technical skills, their teamwork skills, and their levels of achieving individual goals on the one hand and their influence on team performance, on the other hand, are absent.

This study aims to propose an original approach to monitor and develop performance through appropriate management, we propose a model that allows evaluating performance base on several factors that affect individual performance by classifying the team members.

To this end, we used the K-mean algorithm for classifying a data set of 23 team members assessed against eleven variables.

The results of the present study will allow managers to adopt an appropriate management style to each cluster of profiles, to plan targeted training, and to prepare for the future by managing critical skills.

2. BIG DATA, MACHINE LEARNING AND HUMAN RESOURCES MANAGEMENT

2.1. Big Data

With the quantitative explosion of digital data, "Big data" is emerging as a set of processes and techniques that allow an organization to create, manipulate and manage data and extract new knowledge to create economic value.

In the literature, the concept of Big Data is defined through the theory of 5V (Volume, Variety, Velocity, Veracity, and Value).

Firstly, Volume is the size of data streams constantly arriving at an exponential size ranging from petabyte to exabyte, the global volume of digital data keeps increasing year after year. These flows vary between data internal to the company (Customer relationship management "CRM", Internal information system "IS", etc.), external data (social media, emails, mobile devices, etc.), structured data (documents, images, etc.), or unstructured (tweets, GPS data, sensors, etc.) and difficult to handle using conventional computer tools.

The third characteristic is the speed which corresponds to the speed of production of this data, as for the fourth characteristic, it is the veracity of the data, finally, the last characteristic is the added value of these data and their applications [7].

The real richness of a Big Data project is to cross heterogeneous data in real-time and to imagine possible combinations and correlations.

2.2. Machine Learning

The "Big Data" strategy has emerged as one of the major issues related to the development of new technologies within the organization. It is considered to be the engine of innovation, customer satisfaction, and the achievement of greater profit margins (Barton & Court, 2012) [8]. It allows better productivity when decisions are made based on analyzes and data cross-referencing. A study by McAfee and Brynjolfsson (2012) [9] has shown that companies that have adopted advanced techniques in data analysis achieve higher rates of productivity and profitability than their competitors. It also allows better management of information in terms of use and classification of information by priority. According to Kaufman (1973) [10]. Managers do not need to acquire more information to make better decisions, but the rather better organization and better use of the information at their disposal.

Traditional analytical tools are not performing well enough to fully exploit the value of big data. The volume of data is too large for comprehensive analyzes, and the correlations and relationships between these data are too large for analysts to test all hypotheses to derive value from the data.

Basic analytical methods are used by business intelligence and reporting tools for reporting amounts, for doing accounts, and for performing SQL queries. Online analytical processing is a systematic extension of these basic analytical tools that require human

intervention to specify what needs to be calculated.

2.3. Human Resource Management

According to the ISO standards which govern human resources management, the latter is defined as the set of practices implemented to identify, administer and develop the human resources involved in the activity of a given company, its purpose is to optimize the contribution of people to the financial success of the organization. Among these practices, we can cite recruitment, communication, payroll management, and skills management.

Human resources management has a large amount of data on employees. However, for a long time, this data was only used for descriptive reporting purposes, as in the social report for example (annual document in which companies must publish indicators on the population of the company).

More recently, a new trend has emerged called "HR analytics" (Marler and Boudreau, 2016) [11]. It is presented as a more sophisticated way of mobilizing data, notably by using more complex statistical methods, but above all by aiming at a different objective. It is no longer a question of providing purely descriptive reporting, but of mobilizing data to better understand a phenomenon, to improve decision-making.

In terms of HR, very little research works on big data, in particular, its application to improve team performance [12]. Thus, the development of a decision support model to develop both individual and collective results is an urgent and important issue. For this, we have chosen to use clustering algorithms, in particular, K-means to build our model.

3. RELATED WORKS

Human resources analysis has become a powerful tool that makes it possible to move from reporting to the extraction of new knowledge, it is a process that collects and analyzes Human Resources data from traditional sources or by using another source now available to improve overall performance by improving that of teams [13].

Optimal performance management, through the objective assessment of that of individuals, is one of the main issues in the application of big data to HRM, this task is done subjectively by managers. In short, current SI tools do not make it possible to measure this performance or to know the factors that influence it.

In terms of human resources, but the correlation of several factors is absent, yet much research is focused on the impact of one or two factors on the overall performance, training is one of these factors, the article published in 2019 by researchers Josh-ua S. Bendickson and Timothy D. Chandler [14] was the study we were inspired to conduct this research.

The researchers referred to data from 2003 to 2011, including 30 Major League Baseball organizations (along with their affiliates), which were analyzed using regression models to examine the impact of baseball programs. Human Capital Development (HCDP) on financial performance through operational performance. The results support the hypothesis that better human capital development programs lead to operational performance, which in turn leads to increased revenue and sales.

To test the theoretical model, a regression model including control variables is used to determine if the development ranking has a significant impact on team wins. Next, mediation models are used to determine the impact of development rankings on earnings and average attendance for team wins.

This study proves that teams win more matches two years after the increase in development ranking (the number of development ranking decreases), this model represents an inspiration for how a skills ranking (which varies from one observe to another) can be used to increase overall performance.

But in an industrial context, this task is even more complex and complicated, with traditional HR tools many questions remain unanswered, namely: how to measure the contribution of an individual to collective

success? why does an employee seek to leave a given company? that he is the right candidate to recruit?

In this section, we will try to summarize the work that has attempted to establish models that make performance predictions based on data collected from individuals.

Iwamoto et al. [15] have proposed an approach based on a multiple regression model which assesses individual performance based on the purely financial results of employees which influence a performance indicator of the organization, this represents a first step towards the objective assessment of performance. individual performance but our goal is to establish the link with other factors as well, years later Abdullah et al. [16] established a model that demonstrates the importance of knowledge versus skills through a case study in Malaysia. The analytical hierarchy process is used to integrate the multifaceted preferences of the five criteria of human capital to determine the importance of the four identified indicators,

Chen and Chen [17] applied a data mining algorithm, based on decision trees and association rules, to employee characteristics and performance, which represents a starting point for our research which aims to First, establish a resource classification model to objectively assess their performance and predict initial risks.

4. TOWARDS A NEW METHOD OF CLASSIFICATION OF TEAM PERFORMANCE

4.1. Research Design

Among the standard research models that frame the research work in time, effectiveness, and efficiency, we have decided to adopt the (CRISP-DM) model. This model described in (figure 1), is one of the most used in the industrial context.

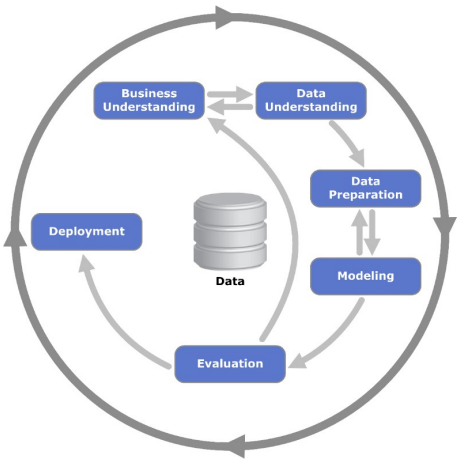

Figure 1: CRISP-DM process model

4.2. Business Understanding, Data Understanding And Preparation

Considering the factors that influence performance according to the literature review, we have chosen a database published by IBM. It describes a team of 1470 people, assessed by IBM scientists against 34 factors, they collected information on age, income, seniority, and some other HR data concerning their profiles, the variables are summarized in a data set is made of 1470 rows and 35 columns.

We use for the rest of our study the R software to process the HR dataset and to build our model, the data was transferred to a data frame of 1470 rows and 35 columns. The types of data were then reviewed and modified.

Data cleaning process is performed on the data set to remove any missing values. Then a correlation analysis is performed on the data sets,

After that we split the data into a training sample and a test sample using a random split technique. The distribution used is 80% for learning, 20% as testing data.

4.3. Modeling

Aims to propose an original approach to monitor and develop performance through appropriate management by identifying and grouping similar, which then allows us to discover the performance factors that characterize a large work team described by several factors.

we keep for the rest of the study seven variables that we deemed useful based on the literature review. These variables are summarized in the table 2.

TABLE 2. Active variables

Name	*Description*
AGE	Age
JOB LEVEL	Level of job
MONTHLY INCOME	Monthly salary
NUM COMPANIES WORKED	Number of companies worked at
TOTAL WORKING YEARS	Total years worked
YEARS AT COMPANY	Total number of years at the company
YEARS IN CURRENT ROLE	Years in current role

the objective of our study is to build a model that makes it possible to identify and group similar employees according to several issues of performance factors from HR data. For this, we will refer to dimensionality reduction algorithms in particular principal component analysis (PCA) to both facilitate visualization and compress and synthesize our training and test database.

Principal Component Analysis is one of the dimensionality reduction algorithms, it consists of transforming interdependent variables (called "correlated" in statistics) into new variables uncorrelated from each other. These new variables are called “principal components”, or principal axes [18].

In the next part of our model we opted for clustering, in our study context, we decided to apply separate K-means and hierarchical clustering algorithms on the compressed training set.

In fact, Kmeans and Hierarchical clusturing are the most suitable clustering techniques in terms of ease, speed, and ability to be implemented at scale. Then we tested both models on a test set. Finally, we evaluated the two results to validate the most efficient model.

The model construction steps are summarized in (Figure 2).

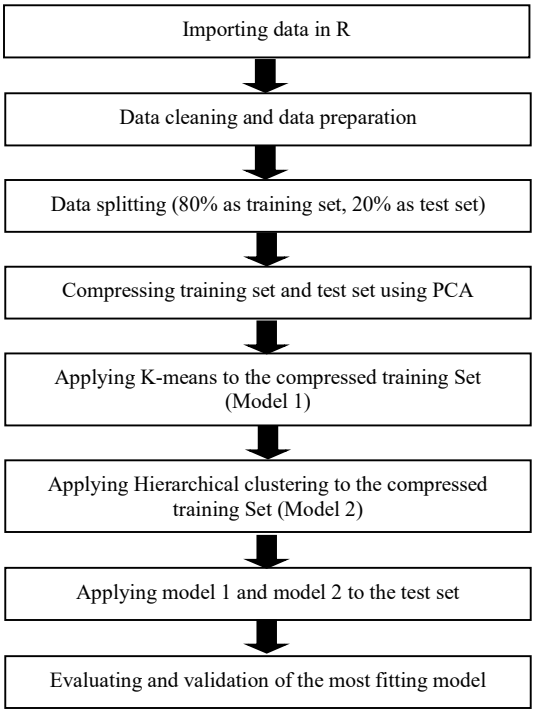

Figure 2: Model building steps

4.4. Results

4.4.1. Graph of individuals using PCA

The graph below (figure 3) shows the distribution of individuals from each department with concerning the first two factors of the Principal Component Analysis of the training set.

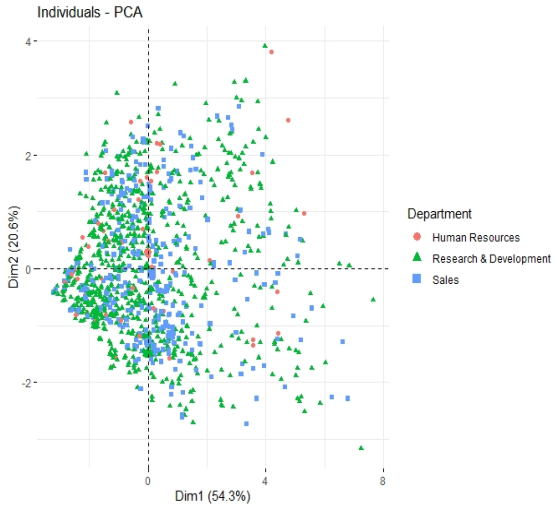

Figure 3: Employees distribution using PCA

The distribution of variables summarized in the table below (table 3) it possible to consider the factor 1 (PC1) as experience factor (53.4% of the variance), and factor 2 (PC2) as stability factor (20.6%) (Table 3).

Factor 1, the experience factor, is the most important in terms of variance. It includes the variables that assess the age, the level of the position, the monthly income, the total number of years of experience, and the length of service in the company. As for factor 2, stability factor, it comprises the variables which assess the number of antecedent companies and the length of service in the company.

TABLE 3. The distribution of variables

Variables	*Factor 1 (expertise)*	*Factor 2 (stability)*	*Factor 3*
AGE	0,6650504		
JOB LEVEL	0,896769		
MONTHLY INCOME	0,8808675		
NUM COMPANIES WORKED		0,7451965	0,5383931
TOTAL WORKING YEARS	0,915279		
YEARS AT COMPANY	0,7347542	-0,5407864	

- Values greater than 0.50 have been retained as significant, values less than 0.50 have been deleted for the clarity of the table.
- A high level of Factor 1 can be explained by a high starting risk, and subsequent-ly, on a population, we can consider that the stability within the team is risky.
- A high level of Factor 2 can be explained by a risk of having a critical skill or having an overqualified skill.

4.4.2. Kmeans results (model 1)

To apply the K-means algorithm on our database we used the "factoextra" package on the R software.

To define the number of clusters k to generate we use such that the total within-cluster sum of square (WSS) is as small as possible (Figure 4), in the rest of the study we decided to take k = 5.

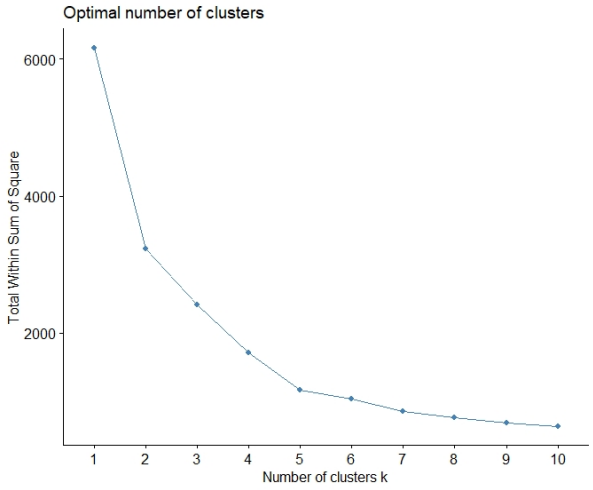

Figure 4: Optimal number of cluster identification using Elbow method

The result generated for k = 5 is presented in the following table (Table 4)

Journal of Theoretical and Applied Information Technology
15th September 2022. Vol.100. No 17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4. Results

Cluster number	*Size of the Cluster*	*WCSS*	*Average PC1 per cluster*	*Average PC2 per cluster*
1	303	260,81	0.4527564	1.1897820
2	111	173,94	-2.7204782	1.6069972
3	368	243,85	1.8368283	-0.3454493
4	84	193,47	-4.3964170	-1.0816469
5	310	287,66	-0.4576319	-1.0351514

with:

BSS/TSS = 81.2 % (1)

The result of the classification of individuals by our model is represented by the graph below (Figure 5).

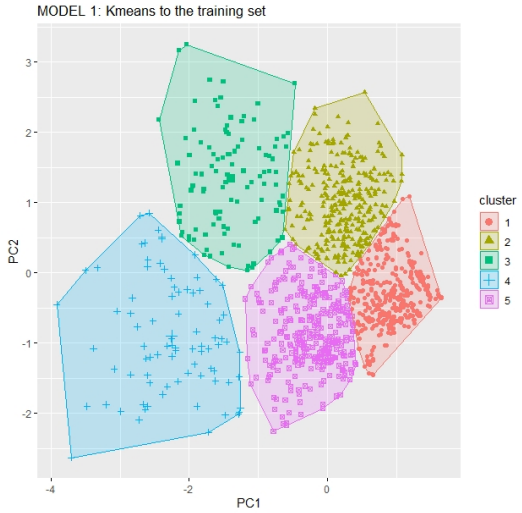

Figure 5: Cluster graph using Kmeans

4.4.3. Hierarchical clustering results (model 2)

To apply the K-means algorithm on our database we used the "factoextra" package on the R software.

To define the number of clusters k to generate we use dendrogram method in (Figure 6), in the rest of the study we decided to take k = 5.

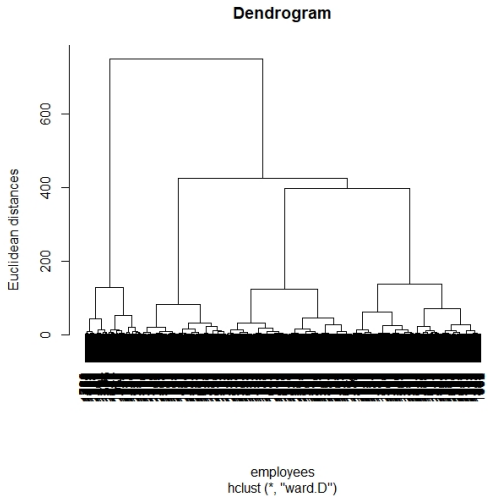

Figure 6: Dendrogram

The result of the classification of individuals by model 2 is represented by the graph below (Figure 7).

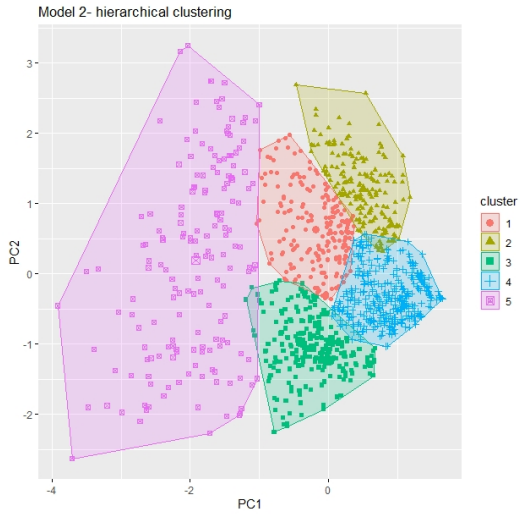

Figure 7: Cluster graph using Hierarchical clustering

5. EVALUATION AND DISCUSSION

According to the results previously obtained, we can categorize the kmeans clusters as summarized in the following table (Table 5).

Table.5 Kmeans clusters interpretation

Cluster	*Interpretation*
1	Stable expert
2	Unstable senior
3	Unstable junior
4	Stable Junior
5	Stable senior

From an HR point of view, the "cluster 4" corresponds to a stable population generally at the start of their career and in the process of learning, it is the cluster that requires the most support in training and coaching. Their missions must be specific on positions that are not critical for the company.

"Cluster 3" also corresponds to a junior population but with a high risk of departure. The investment in its elements may not be profitable for the company.

"Cluster 2" and "cluster 5" represent a senior population with more autonomy than "cluster 4" and "cluster 3", the instability of "cluster 2" represents a risk of the departure of some key talents, and subsequently a risk for the company and the continuity of the activity if they occupy key positions.

To deal with this problem, the manager must be very close to this population to understand individually their motivations which allows making a talent retention plan, or if necessary, to prepare alternatives to anticipate any departure in the short term. Ideally, key positions should be reserved for clusters that are relatively more stable.

In the end, the "cluster 1" represents stable experts which is the locomotive of the company, a manager must put more efforts to treat them carefully by very meticulous management of talents. The sudden departure of one of these skills can be very costly.

According to the HC clustering results previously obtained, we can categorize the HC clusters as summarized in the following table (Table 6).

Table.6 HC clusters interpretation

Cluster	*Interpretation*
1	Unstable senior
2	Unstable Expert
3	Stable senior
4	Stable Expert
5	Junior

From an HR point of view, the first cluster 5 does not take into account the stability of the employee, it is just considered as a young population generally at the start of their career or in the process of learning on positions that do not require an expertise and subsequently their departures should not impact the company in a considerable way.

"Clusters 1" and "Cluster 3" represent a senior population who are more autonomous compared to "cluster 5", the instability of "cluster 1" (unstable senior) represents a risk of departing some key skills like the same results of the same category on the first model using kmeans, and subsequently the same measures must be taken to control the risk of departure of employees who hold key positions or who have key skills in this cluster.

In this model, the Expert category is devised this time on two clusters compared to the first model using kmeans, this category takes into account the stability criterion, "Cluster 2" and "Cluster 4" represent an expert population who are with more autonomy compared to "cluster 1" and "cluster 3" and which lead the company's activity, the instability of "cluster 2" (unstable expert) represents a serious risk to the continuity of the activity.

Taking into account the above, we have decided to adopt the first model using kmeans, from an HR point of view, on the one hand, this model allows us to rationalize the expenses and the supervision of the junior population, paying for years training on unstable elements can be very costly to the company, the HC model does not take into account the stability criterion for this category, which means that this cluster contains a larger population which makes management for the manager is both expensive and difficult.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

On the other hand, the expert population must imperatively be stable, it is an existential question for any company, the sudden departure of an expert can be fatal taking into account the scarcity on the market of several critical skills, the risk of departure towards a competitor, the recruitment time, its costs of its latter.

6. CONCLUSION

Using classical methods of classification based on average scores, grouping team members into different categories based on their performance is a complex and complicated task. The proposed method makes it possible to obtain a global view of the state of employee performance and simultaneously discover the details of their performance from time to time.

In comparison with HC clustering model, the K-means clustering model represents a suitable method for monitoring the development of team performance. The result of its application is a decision aid for the manager to monitor the performance of these employees regularly, the manager can also use it to plan targeted actions to boost both individual and collective performance.

the novelty and the contribution of our article to the research theme is to have a tool that makes it possible to classify employees according to the two factors level of Experience (PC1) and stability (PC2) which impact performance, which opens up several opportunities to improve several human resource practices challenges, namely:

• Make decisions during recruitments in the event of hesitation between candidates by simulating their performance

• Classify a recruit to get an idea of his level of performance upon integration.

• Master the team structure, for a balanced team by launching recruitment at the right time according to the distribution of performance.

• Prioritize elements whose performance is low to launch targeted training, coaching, or the possibility of sponsorship by elements whose performance is higher.

• Decide in the event of departure, the departure of a performing element is more critical than that of a less performing element and subsequently the results of our work will be at the service of talent retention.

• Create a turnover of critical elements whose departure can impact the overall performance.

• Adapt the payroll strategy in particular the annual increases to the performance of individuals.

Despite the satisfactory results of our research, the scarcity of open-source HR data was one of the obstacles we encountered, implementing our model on real data from a large manufacturing company with more employee's number will help to have a better result.

as a work perspective, we recommend using another algorithm of dimensionality reduction to seek even lower data loss rate.

ACKNOWLEDGMENT

I would like to express my deep gratitude to FS TETOUAN and ENSATE professors for their able to guidance and support in completing this project; indeed, their contribution by stimulating suggestions and encouragement helped me to finalize this work.

REFERENCES:

[1] Yousra Karim, Abdelghani Cherkaoui, "Assessment and improvement of human and organizational factors in an auto-parts manufacturing plant using the fuzzy analytical network process combined with the fuzzy comprehensive evaluation method", *Journal of Theoretical and Applied Information Technology (JATIT)*, February 15th, Vol. 100, No. 03, 2022.

[2] Barney, J. B, "Firm resources and sustained competitive advantage", *Journal of Management*, 1991, 17(1), pp. 99–120.

[3] Evaristus Didik Madyatmadja, Lydia Liliana, Johanes Fernandes Andry,Hendy Tannady, "risk analysis of human resource information Systems using cobit 5", *Journal of Theoretical and Applied Information Technology*, Vol.98, No 21, 2020.

[4] Arthur, Bennett, Edens, Bell, "Effectiveness of training in organizations: a meta-analysis of design and evaluation features", *J Appl Psychol*. 2003 Apr ;88(2):234-45.

[5] Aguinis & Kraiger, "Benefits of Training and Development for Individuals and Teams, Organizations,and Society", *Annu. Rev. Psychol,* 2009, pp. 451–74.

[6] Antonio Aragón Sánchez, María Isabel Barba Aragón & Raquel Sanz Valle, "Effect of training on business results", *The International Journal of Human Resource Management,* 2009*,* 14(6), pp. 956-980.

[7] Sagiroglu, S. and Sinanc, D, "Big Data: A Review, Collaboration Technologies and Systems", *International Conference on Digital Object Identifier*, 2013*,* pp.42-47.

[8] D.Barton, D.Court, "Marketing Advanced Analytics Work for You", *Harvard business review*, Oct;90(10), 2012, pp.78-83.

[9] McAfee, Brynjolfsson, "Big Data: The Management Revolution", *Harvard business review*, 2012, 90(10):60-6, 68, 128.

[10] Kaufmann, W, "Without guilt and justice: From decidophobia to autonomy", *NewYork: P.H. Wyden, 1973.*

[11] Marler, Boudreau, "An evidence-based review of HR Analytics", *The International Journal of Human Resource Management*, November 2016.

[12] Jeffrey B Arthur, "Effects of Human Resource Systems on Manufacturing Performance and Turnover", *The Academy of Management Journal,* 37(3), 1994, pp. 670-687

[13] S.N. Mishra, D.R. Lama, Y. Pal, "Human resource predictive analytics (HRPA) for HR management in organizations", *Int J Sci Technol*. Res. 5 (5), 2016, pp. 33–35.

[14] Joshua S. Bendickson, Timothy D, "Operational performance: The mediator between human capital developmental programs and financial performance". *Journal of Business Research*, 2019, Volume 94.

[15] H. Iwamoto, M. Takahashi, "A quantitative approach to human capital management", *Proc.-Soc. Behav. Sci*, 172, 2015, pp. 112–119.

[16] L. Abdullah, S. Jaafar, I. Taib, "ranking of human capital indicators using analytic hierarchy process", *Proc.-Soc. Behav. Sci,* 107, 2013, pp. 22–28.

[17] C.F. Chen, L.F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry, Expert Syst", Appl. 34 (1), 2008, pp. 280–290.

[18] Fahim A. M., Salem A. M., Torkey F. A. , Ramadan M. A., "An efficient enhanced k-means clustering algorithm", *Journal of Zhejiang University Science A*., 2006, pp. 1626–1633.