ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter06852020.pdf
https://doi.org/10.30534/ijeter/2020/06852020

Decision Support Model for Employee Recruitment Using Data Mining Classification

Clarissa Elfira Amos Pah[1], Ditdit Nugeraha Utama [2]
Computer Science Department, BINUS Graduate Program - Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480. Email: [1]clarissa.pah@binus.ac.id, [2]ditdit.utama@binus.edu

ABSTRACT

This article explains a decision support model (DSM) construction for employee recruitment in a particular IT consultant company using the conception of data mining classification. The created model addresses the company's need in recruiting employees objectively by utilizing historical data to find patterns of potential employees for the company. There are four data mining classification algorithms compared in this case, namely C4.5 Decision Tree, Naive Bayes, Support Vector Machine, and Random Forest. The algorithm with the highest accuracy, specificity, and sensitivity values are operated to produce rules for the DSM. The results are C4.5 Decision Tree has the highest accuracy of 88.24%, specificity of 88.10% and sensitivity of 100%;thus this algorithm is selected to process the value of eight predictor parameter to propose the most valuable employee candidate.

Key words: Data Mining, Decision Support Model, Employee Recruitment

1. INTRODUCTION

High potential employees are important assets to boost business growth; conversely, the less potential employees can only be detrimental to the business. The right employee selection can predict business success in the future [1]. An employee selection process is a treatment to find a qualified employee. If the process is running right, the right employees will come as the results [2]. The selection process depends on some assessments that can distinguish between the quality of one candidate to another and is able to envisage the performance of candidates in the future. The assessment begins by looking at the attributes' values that are inherent in the candidates and compliance with the criteria determined by the company [1].

As a company grows, more and more criteria are determined to assess employees, starting with criteria per division, criteria per department, criteria per employee position and other criteria according to company regulations. In addition, the large number of applicants also makes the recruitment process take a long time because they have to compare one applicant with other applicants in order to get the applicant who best matches the specified criteria. Seeing this condition, some researchers started to create some decision support models (DSM) that make some companies easy to make a decision with match employees with the job criteria offered by the companies. According to [3], a DSM allows the results of decisions that are objective and can be justified logically and scientifically because this model is built on logical and correct reasoning. To produce this reasoning requires accurate calculations and the involvement of various decision parameters. So that, in this works, the DSM is intended to do profile matching. According to [4], profile matching is a decision making mechanism that assumes that there are ideal predictor attribute values that can be owned by prospective employees. In matching the profile, a company can determine what kind of prospective employee who are match with the company criteria and who are not in accordance with company criteria.[5]has built a DSM for marketer selection and found that a DSM can help a company to find best prospective marketer by putting aside feeling and presumption.

With this DSM, the assessment of prospective employees becomes more objective and consistent ;as numerous assessment factors are carried out automatically by a system, and this can help decision-makers determine the right prospective employees to be hired [7]. According to [8], DSM help users to process and analyze large volume of varies information quickly and comprehensive. In addition, with the large number of prospective employees, DSM can provide an assessment of each prospective employee and produce an order of employees that best meets company criteria [9].

2. PERSONNEL SELECTION IN IT CONSULTANT COMPANY

PT. Phincon is an IT consultant company that has been established in Indonesia since 2008. The company experiences an average increase in the number of employees by 48% annually and also an increase in average employee turnover which reaches 11% annually. Surely the number of applicants that are evaluated to be hired will be greater in number than those that are literally recruited.

The process of recruiting employees at the company also went through several stages, including screening, psychological testing, interviews with the human resources department (HRD), user interviews with division head (DH), and job offers. The whole stages can take a minimum of two weeks, and a maximum of approximately two months. The main problem that occurs is the amount of time wasted focusing on prospective employees who are not necessarily potential. All employees who apply are chosen by HRD staff in the screening phase to be tested in psychological tests and interviewed. After being considered as qualified, the step will be continued to an interview with the DH. At this stage, there is often a mismatch between HRD staff choice and the choice of DH who should be considered more because it is directly related to the work that the prospective employee will face. Finally, all efforts from screening to HRD interviews were in vain. To minimize this incident, DSM will be placed at the screening stage where profile matching will be carried out between the attributes of prospective employees with the criteria desired by the company (DH or decision-maker).

However, the desired criteria have never been formulated currently, so the company only depends on the decision of interviews with the DH. Therefore, in this study, researchers will apply data mining classification to classify high-potential and less-potential employees based on historical data of employees who have been accepted to work. Every employee who has worked at the company has a key performance indicator (KPI) average that will be used as a criterion. If the employee KPIaverage is A, then employees will be categorized as high-potential employees, while employees who have an average KPI B, C, D, and E will be categorized as less-potential employees. The data mining classification algorithm will formulate the pattern of what attribute values that high-potential employees have. This pattern will be used to formulate the DSM validations at the screening stage.

3. DATA MINING CLASSIFICATION FOR PERSONNEL SELECTION

In 2008, [2] used data mining classification to design a DSM for personnel selection in a high-tech industry. They compared 4 classification algorithms namely CHAID, CART, ID3, and C4.5 and finally decided to use the CHAID decision tree algorithm in their DSS because most of their data were categorical. In 2013, [10] also used data mining classification for personnel selection in a commercial bank. They compared the accuracy of 4 data mining classification algorithms namely QUEST, CHAID, C5.0 (extension of C4.5), and CART. As a result, they found that C5.0 had the best accuracy of 80.43% and used this algorithm to determine the pattern of the attributes of potential employees.

In the same year 2013, [11] compared a number of data mining classification algorithms to predict the performance of teachers in a school. The involved data mining classification algorithms including Naive Bayes (NB), ID3, CART, and LAD. As a result, NB had the highest accuracy of 80.35% in predicting the performance of potential teachers. In addition, [12] also compared the C4.5 and NB algorithms to predict the potential of students towards a chosen major and found that the accuracy for the C4.5 algorithm was higher than the NB one. C4.5 produced an accuracy of 93.31 while NB produced 82.64. [13] applied data mining classification to predict the possibility of employee turnover. They compared Random Forest (RF), Back Propagation (BP), C4.5, Logistic algorithms and as a result, RF has the highest accuracy of 92.65%. In 2019, [14] conducted an exposition of several data mining algorithms for predict a customer churn. The goal was to know what factors that owned by customers can influence the customers churn selections. This scenario is similar with the personnel selection scenario but differentiated by business needs.

4. RESEARCH METHODOLOGY

Figure 1 shows the case study flow for employee selection. The flow consists of data collection, data preparation, data mining process, data mining evaluation to select the best data mining algorithm by finding the highest accuracy, specificity, and sensitivity of several algorithms to produce some classification rules.

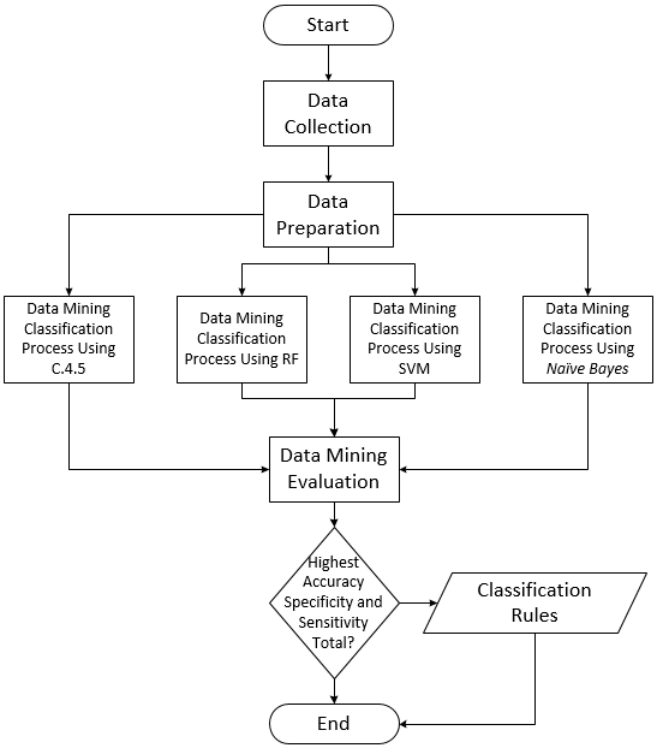

Figure 1: Case Study Flow for Employee Selection Using Data Mining Classification

In data collection, all hired employee data are collected including personal data, educational data, experience data, recruitment source, and performance result average. In data preparation, all attributes and attribute values that have been defined are processed to produce the final dataset that will be used in the data mining process. The data preparation done includes attribute selection, data cleaning, data transformation, predictor attributes definition, and target

attribute definition. The target attribute is also called the label attribute where the value of this attribute will be used as a label that defines each record in the dataset. In this case, the target attribute KPI average that classifies each employee into the "Recommended" label or the "Not Recommended" label. Meanwhile, the predictor attributes are obtained from all attributes other than the label attribute, such as Age, Marital Status, Degree and so on which will be used to predict the value of the target attribute. During this data preparation process, experts and representatives from the talent management division (TMD) remain involved to avoid mistakes.

Then, the dataset produced is used in the data mining process. There are four algorithms that are used to be compared, namely C4.5, NB, RF, and Support Vector Machine (SVM). In this case study, authors choose to use C4.5 rather than C5.0 because C4.5 is still widely used until this day with fairly high accuracy, even this case study also shows that C4.5 is still superior to the NB and RF classification algorithms which also have good accuracy records. SVM is also considered in this study as a trial.

C4.5 algorithm has several steps, they are preparing training data, selecting attributes as roots, creating branches for each value, and repeating the process for each branch until all cases in the branch have the same class [15]. The attribute with the highest gain is selected as the root. Gain is obtained by finding the entropy value as shown in equation 1; where *n* is the number of values contained in the target attribute, *pi* is the ratio between the number of samples in class *i*, and the sum of all samples in the data set. After getting entropy, the next is to get the gain value as shown in the equation 2; where A is a particular attribute, i is a possible value for attribute A, | Si | is the number of samples for values i and | S | is the sum of all data samples.

$$\boldsymbol{Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi} \quad \textbf{(1)}$$

$$\boldsymbol{Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\,(S_i)} \quad \textbf{(2)}$$

The highest gain value will be selected as the root. After getting the root attribute, the number of categories in the root attribute determines the many branches of the root. After that, recalculate the gain to determine the branch attribute until all cases in the branch have the same class [16].

The NB classification algorithm refers to Bayesian theorem calculations by looking for probabilities as shown in equation 3; where *X* is a sample or data record, while *H* is a hypothesis. *P(H | X)* is the probability that X is true for hypothesis *H*,*P(H)* is the prior probability for hypothesis *X* while *P(X)* is the prior probability for sample *X*.

$$\boldsymbol{P\,(H|X) = \frac{P(X|H)P(H)}{P(X)}} \quad \textbf{(3)}$$

Furthermore, there is a SVM classification algorithm that classifies classes by looking for hyperplanes with the maximum distance between classes (margins). To find the hyperplane with the best margins, what needs to be done is to look for support vectors or the outermost objects closest to the hyperplane. After defining support vectors, optimization is performed until finding hyperplanes with maximum distance [17].

RF algorithm is an example of the application of decision tree-based ensemble learning. Some decision tree models are built randomly from training data samples and then voting to determine decisions [17].

Each algorithm is evaluated to determine the most suitable in this case study that will produce the classification rules. Each algorithm will be trained using X-Fold Cross-Validation to produce a confusion matrix as shown in Table 1. This matrix can be used to measure the accuracy, sensitivity, and specificity. The algorithm with the highest accuracy, sensitivity, and specificity will be selected as the best classification model. The accuracy, specificity, and sensitivity formulas are shown in equations 4, 5, and 6 as mentioned by [17].

Table 1: Confusion Matrix

		Actual Value	
		True	**False**
Prediction Value	**True**	TP (True Positive)	FP (False Positive)
	False	FN (False Negative)	TN (True Negative)

$$\boldsymbol{Accuracy: \frac{TP+TN}{TP+TN+FP+FN}} \quad \textbf{(4)}$$

$$\boldsymbol{Sensitivity = \frac{TP}{P}} \quad \textbf{(5)}$$

$$\boldsymbol{Sensitivity = \frac{TN}{N}} \quad \textbf{(6)}$$

5. RESULT AND DISCUSSION

This section explains the result obtained in this study including final dataset that produced from data preparation, data mining classification algorithms comparison result and generated rules from the selected data mining classification algorithm.

5.1 Data Preparation

In this research, only employees with Consultant positions were processed. There were 170 employees with 145 active employees and 25 resigned employees. There were 21 maintained employee attributes, namely Employee ID, Employee Name, Position, Role, Class, Join Date, Length of Service, Gender, Marital Status, Religion, Date of Birth, Join Age, Degree, University, University Accreditation, Employee

Status, Resignation Flag, Resignation Date, KPI Average, Recruitment Sources, and Work Experience. But only 8 attributes selected by the experts, they are Join Age, Working Experience, Degree, University Accreditation, Recruitment Source, Marital Status, Sex as predictor attributes and KIP Average as target attribute. In previous study, these attributes have also been used as considerations for employee recruitment as shown Table 2. So, comparing expert choice with the previous studies attributes will produce the mapping as shown in Table 3.

Table 2: Predictor Attributes References from Previous Case Study

No	Predictor Attribute	References
1	Join Age	[2], [10], [18], [9], [19]
2	Sex	[2], [10]
3	Marital Status	[2], [10]
4	Degree	[2], [10], [20], [21], [22], [9], [19], [23]
5	University	[2], [24], [10], [20]
6	University Majority	[2], [10]
7	Work Experience	[2], [10], [25], [18], [19], [23]
8	Recruitment Source	[2]

Table 3: Comparison Attribute Based on Expert Choice and Previous Study

Predictor Attribute	Experts Choice	Previous Study
Join Age	V	V
Working Experience	V	V
Degree	V	V
University Accreditation	V	-
Recruitment Source	V	V
Marital Status	V	V
Sex	V	V
University	-	V
University Majority	-	V

5.2 Data Mining Classification Process

This research was using the Rapidminer tool to process the data mining classification algorithm. Each algorithm was processed using some operators as shown in Table 4.Every process had a Retrieve Data and a Set Role operators to import a final dataset after data preparation, then, to set each attribute role whether as a predictor attribute or a target attribute.

Furthermore, one of the data mining algorithms was applied namely Decision Tree, SVM, RF, or NB. In this study, C4.5 used the Decision Tree operator with information gain as the splitter. After that, the selected data mining algorithm was evaluated by a Cross-Validation operator. This operator applied X-Fold Cross-Validation to produce a confusion matrix for accuracy, sensitivity and specificity calculation. Table 5 shows the confusion matrix result for each data mining classification algorithms to be compared.

Based on Table 5, C4.5 Decision Tree has the highest accuracy value of 88.24%, followed by RF, NB, and SVM. The highest sensitivity value is produced by NB and C4.5 Decision Tree at 100% and the highest specificity value is generated by C4.5 Decision Tree at 88.10%. Thus, C4.5 produces the best algorithm that will be used to determine classification rules.

Table 4: Data Mining Operator Used in Rapidminer

		Data Mining Algorithm			
		C4.5	SVM	RF	NB
Operator	**Retrieve Data**	V	V	V	V
	Set Role	V	V	V	V
	Decision Tree	V	-	-	-
	Cross-Validation	V	V	V	V
	Nominal To Numerical	-	V	-	-
	SVM	-	V	-	-
	Random Forest	-	-	V	-
	Naive Bayes	-	-	-	V

Figure 2 shows the produced C4.5 decision tree that involving University Accreditation, Join Age, Marital Status, Sex and Working Experience attributes. Where University Accreditation has the highest gain value so it is placed as the root and each value from University Accreditation is made a branch for the next node. The determination of the next node is also obtained from the highest gain of the attributes under the branch, so that the node Join Age, Marital Status and Working Experience. There are 9 rules as the result that can be used to construct a decision support model as shown in Table 6.

The Yes Percentages will be used as the rule weights to determine the prospective employee's potential level. The more the prospective employee matched the rule, the more he will be considered to be hired.

6. CONCLUSION AND FUTURE WORKS

This paper has explained the detail of the data mining process to classify the prospective employee whether will be hired or not based on experimental or historical data. Researcher found that C.4.5 decision tree algorithm is the most accurate algorithm to be used to produce classification rules. The data mining classification was indeed a very good way to avoid the subjectivity in decision making because it use the historical data, but in this case study, the collected data tends to be small and even though with high accuracy, there are imbalance total of label class in the dataset (total of "Yes" label and "No" label as target). So in the future case study, we can combine with the other method to deal with the imbalance label class.

In this study, the scope of the recruitment process is only limited to the Screening stage, so that the data used were limited to employee personal data, education data and working experience data. In the future, the data collection can be expanded to the results of interviews and other test values such as the results of psychological tests, IQ, EQ, Personality test results, English proficiency tests and other related tests to optimize the data mining classification rules.

Table 5: Accuracy, Sensitivity and Specificity Comparison of Data Mining Classification

No	Data Mining Classification	*Confusion Matrix*						Sensitivity	Specificity	Accuracy
		TP	TN	FP	FN	P	N			
1	C4.5 Decision Tree	2	148	0	20	2	168	100%	88.10%	88.24%
2	SVM	0	148	0	22	0	170	0%	87.06%	87.06%
3	RF	2	147	1	20	3	167	67%	88.02%	87.65%
4	NB	1	148	0	21	1	169	100%	87.57%	87.64%

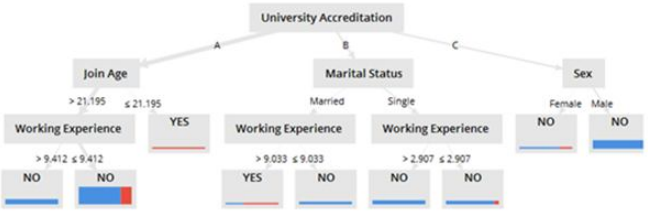

Figure 2: C4.5 Decision Tree Result Model

Table 6: Decision Support Model Rules for Employee Recruitment

Rule#	Rule	Decision Count	Yes Percentages
1	IF University Accreditation is "A" AND Join Age is ">21.195" AND Working Experience is ">9.412" THEN Decision is "NO"	NO=19, YES=0	0.00
2	IF University Accreditation is "A" AND Join Age is ">21.195" AND Working Experience is "<=9.412" THEN Decision is "NO"	NO=58, YES=15	0.21
3	IF University Accreditation is "A" AND Join Age is "<=21.195" THEN Decision is "YES"	NO=0, YES=3	1.00
4	IF University Accreditation is "B" AND Marital Status is "Married" AND Working Experience is ">9.033" THEN Decision is "YES"	NO=1, YES=2	0.67
5	IF University Accreditation is "B" AND Marital Status is "Married" AND Working Experience is "<=9.033" THEN Decision is "No"	NO=8, YES=0	0.00
6	IF University Accreditation is "B" AND Marital Status is "Single" AND Working Experience is ">2.907" THEN Decision is "No"	NO=14, YES=0	0.00
7	IF University Accreditation is "B" AND Marital Status is "Single" AND Working Experience is "<=2.907" THEN Decision is "No"	NO=11, YES=1	0.08
8	IF University Accreditation is "C" AND Sex is "Female" THEN Decision is "No"	NO=3, YES=1	0.25
9	IF University Accreditation is "C" AND Sex is "Female" THEN Decision is "No"	NO=34, YES=0	0.00

ACKNOWLEDGEMENT

We would like to thank Bina Nusantara University who has supported and sponsored our studies and works, particularly Bina Nusantara Graduate Program, Master of Computer Science.

REFERENCES

1. R. M. Guion, S. Highhouse and D. Doverspike, **Essentials of Personel Assesment and Selection**, New York: Routledge, 2016, pp. 3-15.
2. C. F. Chien and L. F. Chen, **Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry**., *Expert Systems with Applications*, vol. 34, no. 1, pp. 280-290, 2008. https://doi.org/10.1016/j.eswa.2006.09.003
3. D. N. Utama, **Sistem Penunjang Keputusan: Filosofi, Teori dan Implementasi**, Yogyakarta: Penerbit Garudhawaca, 2017.
4. C. Mursa, D. Utama and Z. Fananie, **Implementasi Analisis Gap untuk Sistem Pendukung Keputusan (SPK) Kenaikan Jabatan**, *Studi Informatika: Jurnal Sistem Informasi*, vol. 4, no. 1, pp. 1-17, 2011.
5. S. Oktafiani and D. N. Utama, **Generic Model of Fuzzy Profile Matching for Determining the Best Marketer**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1, pp. 859-869, 2020. https://doi.org/10.30534/ijatcse/2020/123912020
6. M. D. T. P. Nasution, Y. Rossanty, A. D. GS, S. Sahat, R. Rosmawati, N. Kurniasih, A. S. Ahmar, E. Susanto, Y. Novitasari, S. Suhardi, I. A. Kadir and R. Rahim,

Decision Support Rating System with Analytical Hierarchy Process Method, *International Journal of Engineering & Technology*, pp. 105-108, 2018. https://doi.org/10.14419/ijet.v7i2.3.12629

7. M. A. Ahmad, I. Tvoroshenko, J. H. Baker and V. Lyashenko, **Modeling the Structure of Intellectual Means of Decision-Making Using a System-Oriented NFO Approach**, *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 11, pp. 460-465, 2019. https://doi.org/10.30534/ijeter/2019/107112019
8. S. Salmon and B. Harpad, **Penerapan Metode Analytical Hierarcy Process (AHP) Pada Pemilihan Staf Laboratorium Komputer STMIK Widya Cipta Dharma Samarinda**," *Sebatik STMIK Wicida*, pp.22-29,2018. https://doi.org/10.33299/jpkop.22.1.1322
9. A. Azar, M. Sebt, P. Ahmadi and A. Rajaeian, **A model for personnel selection with a data mining approach: A case study in a commercial bank**, *SA Journal of Human Resource Management*, vol. 11, no. 1, pp. 1-10, 2013.
10. A. K. Pal and S. Pal, **Evaluation of Teacher's Performance: A Data Mining Approach**, *International Journal of Computer Science and Mobile Computing*, vol. 2, no. 12, pp. 359-369, 2013.
11. L. Swastina, **Penerapan Algoritma C4.5 Untuk Penentuan Jurusan Mahasiswa**, *Gema Aktualita*, vol. 2, no. 2, pp. 93-98, 2013.
12. X. Gao, J. Wen and C. Zhang, **An Improved Random Forest Algorithm forPredicting Employee Turnover**, *Mathematical Problems in Engineering*, pp. 1-12, 2019.
13. P. Hooda and P. Mittal, **An Exposition of Data Mining Techniques for Customer Churn in Telecom Sector**, *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 11, pp. 506-511, 2019. https://doi.org/10.30534/ijeter/2019/177112019

14 K. and E. T. Luthfi, **Algoritma Data Mining**, Andi Publisher, 2009, pp. 13-21.

15. R. T. Vulandari, **Data Mining Teori dan Aplikasi Rapidminer**, Penerbit Gava Media, 2017, pp.13-33.
16. Suyanto, **Data Mining Untuk Klasifikasi dan Klasterisasi Data**, Bandung: Informatika, 2019, pp.140-290.
17. B. Supriaty, R. Malani and O. D. Nurhayati, **Design of Information System for Acceptance Selection of Prospective Employees Online Using Tahani Fuzzy Logic Method and Simple Additive Weighting (SAW)**, *International Journal of Computing and Informatics (IJCANDI)*, vol. 1, no. 1, 2016.
18. W.-S. Tai and C.-C. Hsu, **A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method**, in *9th Joint International Conference on Information Sciences (JCIS-06)*, 2006.
19. V. Lytvyn , V. Vysotska, P. Pukach, I. Bobyk and B. Pakholok, **A Method for Constructing Recruitment Rules Based on The Analysis of a Specialist's Competences**, *Eastern-European Journal of Enterprise Technologies*, pp. 4-16, 2016.
20. M. H. Mammadova and Z. G. Jabrayilova, **Decision-Making Support in Human Resource Management Based on Multi-Objective Optimization**, *TWMS Journal of Applied and Engineering Mathematics*, vol. 9, no. 1, pp. 55-72, 2018.
21. M. Mammadova and Z. Jabrayilova, **Application of Fuzzy Optimization Method in Decision-Making for Personnel Selection**, *Intelligent Control and Automation*, pp. 190-204, 2014. https://doi.org/10.4236/ica.2014.54021
22. R. C. Gustilo and C. C. Escolar-Jimenez, **An Analytic Hierarchy Process Approach in the Shortlisting of Job Candidates in Recruitment**, *International Journal of Emerging Trends in Engineering Research*, vol. 7, no. 9, pp. 333-339, 2019. https://doi.org/10.30534/ijeter/2019/17792019
23. A. Kelemenis and D. Askounis, **A New TOPSIS-based multi-criteria approach to personel selection**,*Expert Systems with Applications*, p. 4999–5008, 2010. https://doi.org/10.1016/j.eswa.2009.12.013
24. B. Karatop, C. Kubat and O. Uygun, **Talent management in manufacturing system using fuzzy logic approach**,*Computers & Industrial Engineering*, pp. 1-10, 2014. https://doi.org/10.1016/j.cie.2014.09.015