Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 9, No. 2, April 2021, pp.799-807

Hybrid of K-means clustering and naive Bayes classifier for predicting performance of an employee

Zainab Mahmood Fadhil
Department of Computer Engineering, University of Technology, Iraq

ABSTRACT

Predicting the performance of an employee in the future is a requirement for companies to succeed. The employee is the organization's main component, the failure or organization's success based on the performance of an employee, this has become an important interest in almost all types of companies for decision-makers and managers in the implementation of plans to find highly skilled employees correctly. Management thus becomes involved in the success of these employees. Particularly to guarantee that the right employee at the right time is assigned to the convenient job. The forecasting of analytics is a modern human resource trend. In the field of predictive analytics, data mining plays a useful role. To obtain a highly precise model, the proposed framework incorporates the K-Means clustering approach and the Naïve Bayes (NB) classification for better results in processing performance data of employees, implemented in WEKA, which enables personnel professionals and decision-makers to predict and optimize their employees' performance. The data were taken from the previous works, this was used as a test case to illustrate how the incorporates of K-Media and Naïve Bayes algorithms increases the exactness of employee performance predicting, compared with the K-Means and Naïve Bayes methods, the proposed framework increases the accuracy of predicting the performance of an employee.

Keywords: Performance of an Employee, NB, K-Means

Corresponding Author:

Zainab Mahmood Fadhil
Department of Computer Engineering
University of Technology, Iraq
Baghdad, Iraq
E-mail: 120094@uotechnology.edu.iq

1. Introduction

Human resources became one of the key concerns of managers in virtually every sector, consist of governmental organizations, private businesses, and educational institutions [1] Corporate companies want to schedule the correct selection of employees. After recruiting staff, management was concerned about their performance in order to retain the successful performers of their employees [2] by management construct assessment systems. Data Mining (DM) is knowledge discovery and an information area, which is young and promising [3].
Information can be extracted and accessed through DM techniques to convert database work from storage and recovery to the learning and extraction of knowledge [4]. DM has many tasks including clustering and classification [5] [6] [7]. The techniques for classification are supervised learning methods, which classify data items into limited class labels. It is one of the most important methods for DM for building classification models from the input data set [8] [9]. Models are widely used in classification methods to forecast future data patterns [10]. One of the best-known algorithms of classification is NB [11]. NB classification is another method used to predict a target class [12]. It relies on probabilities when calculating it, but also offers a specific method for designing different learning algorithms that do not use likelihoods directly [13]. The findings of this classification are therefore more precise, productive, and sensible to recent data inserted into the dataset [14].
A good business is a corporation that can distinguish the performance of their employee. Not all employees should be handled the same as others because every employee performs variously [15]. DM can assist the organization by using its algorithms. i.e. clustering algorithm and NB of DM method can be used to discover

the key proof features of future prediction of an organization [16]. Clustering is the most commonly used procedure for future prediction to group data into groups with the same proof features by which intra-class similarity is maximized or minimized [17].

2. Problem statement

The performance evaluation process of an employee is one of the difficult processes, as these processes seek to feed every employee into his / her performance, as well as to make promotion decisions and raise salaries, they also recognize aspects for the workplace that need to be improved and altered. Human resources in most parts of other corporations' public sectors use conventional evaluation methods that do not enable them to achieve the perfect evaluation of their results. Therefore, many previous works use supervised classification algorithms in DM to create a prediction performance model for their employees. Supervised classification challenges include a category-dependent variable dataset and several independent variables which are useful (or not!) in class predicting. Unsupervised learning takes a dataset without labels and tries to find a latent structure in the data such as clustering. This paper demonstrates how to boost the efficiency of the classifier by using k-means to identify latent "clusters" in the data set.

3. Literature survey

In 2012 this paper, DM algorithms were used to create a grading model to predict employee efficiency. The Decision Tree was the key method for DM used in the development of the model for the classification to generate rules, which collected from questionnaire to 130 employees in several IT companies. Several experiments were carried out to validate the created model by using10-Fold Cross-Validation and Hold-out (60%), for evaluating three algorithms such as the accuracy of ID3 equal to 50% and 43.7%, C4.5 (J4.8) equal to 60.5% and 56.2%, Naïve Bayes that equal to 65.8% and 68.7% [18]. In the 2016 this paper the performance data of employees were gathered from the database of the Human Resource Department at the Kenya School of Government. The process of classification was implemented with three various algorithms of DM such as ID3, C4.5, and NB to distinguish the best and most suitable classification algorithm. The data gathered was to 5 years and consist of 206 assessment reports defined in 14 criteria of the employee's performance. These data have been divided into 2 datasets. For training and testing, the first 110-base dataset was used, while the other 96-base data set was used to validate the model, the comparison of accuracy for different classification methods was given as follows, ID3 64.5%, NB 80.33% C4.5 (J4.8) 82.60%, 92.60% [19]. In 2019 this paper offered the performance prediction of an employee in a company using the classification of NB method, which employed to build the model of prediction. The data used consist of details about 310 employees. There are 28 parameters in the dataset. The result presented that NB successfully graded correctly instances to 95.48% accuracy [20]. In 2019 this paper focuses on the potential to create a predictive employee performance model using classification techniques to the actual data, which was gathered from the Egyptian Civil Aviation Ministry during a questionnaire for 145 employees. The classification process takes place after the preparation and preprocessing of the data. The employee's performance prediction model was developed using the three classification algorithms, SVM, DT, and NB, to obtain the most adequate DM algorithm that may affect and predict the performance of an employee, from the results of experiments, it has been shown that best result for accuracy of C4.5 (J48), Naïve Bayes and SVM were equal to 79.31 %82.07 %86.90 % respectively [21].

4. Materials and methods

4.1. K-Means clustering

K-mean clustering is a commonly used type of clustering. This algorithm is the most common method for the clustering of science and industry [22]. K-means are primarily intended to describe k centroids, one per cluster [23]. Due to various position reasons, these centroids should be positioned ingeniously. The safest option is, therefore, to keep them as far apart as possible [24].

Each point in a given data set is the next step and associated with the next center. If there is no need for an early group the first move is taken. In this case, it is appropriate to update the cluster center from the first step. Following these new centroids a new bond between the same datasets and the nearest new centroid must be achieved, Figure1 shows the general description of K-Means [24] [25].

We have created a loop, which means the k centroids are gradually changing their position until no more changes are made. In other words, centroids do not shift anymore. This algorithm moves objects between clusters until the sum cannot be decreased further. The consequence is a collection of clusters that are as compact and independent as possible as indicates in Figure1 [26]. In this case, the purpose of this algorithm is to minimize an objective function of squared error [27]. The clustering method or cluster analysis is mainly used in applications such as market research, prediction, image processing, pattern recognition, and data analysis. There are many advantages and disadvantages of the K-means Algorithm [28] [29], some of them shown in Table1.

Table 1. Advantages and Disadvantages of K-means Clustering

Advantages	Disadvatages
Easy to implement.	Difficult to predict the number of clusters (K-Value).
With a large number of variables, K-Means may be computaionally faster than hierarchical clustering (if K is small).	Initial seeds have a strong impact on the final results.
k-Means may produce higher clusters than hierarchical clustering.	The order of the data has an impact on the final results.

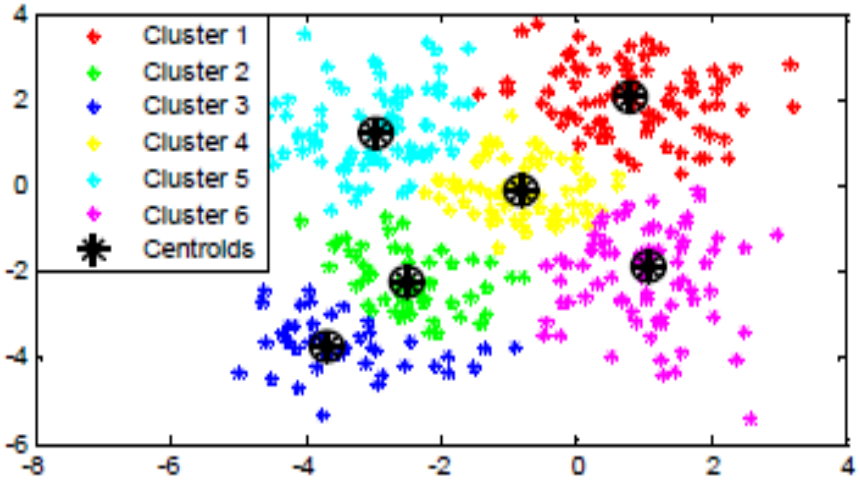

Figure 1. K-Means Clustering [31]

4.2. Naïve bayes classifier

The NB is a classification system that is a statistical classification depend on the Bayes theorem and the maximum posterior hypothesis. This classification is so popular and straightforward that it is simple to implement. [30]. NB can predict the probability of class membership of tuple data that will enter a certain class, according to probability calculations. This method is often used to solve problems in the field of machine learning because it is known to have a high degree of accuracy with simple calculations [32]. NB depends on a strong and fairly simple construction presumption of independence. Figure2 shows the NB classifier that is used to classify classifies all three clusters into more specific categories [33] [34]. There are many advantages and disadvantages of Naïve Bayes algorithms some of them shown in Table2 [35] [36].

Table 2. Advantages and Disadvantages of Naïve Bayes Classifier

Advantages	Disadvatages
When the independent assumption holds then this classifier gives outstanding accuracy.	If the independent assumption does not hold then performance is very low
Easy to implement as only the probability is to be calculated.	Smoothing turns out to be a over-head and a must to do step when probability of a feature turns out to be zero in a class.
It works well with high dimensions such as text classification.	Vanishing value is also a problem due to product of many small probability(eg. 0.05^3).

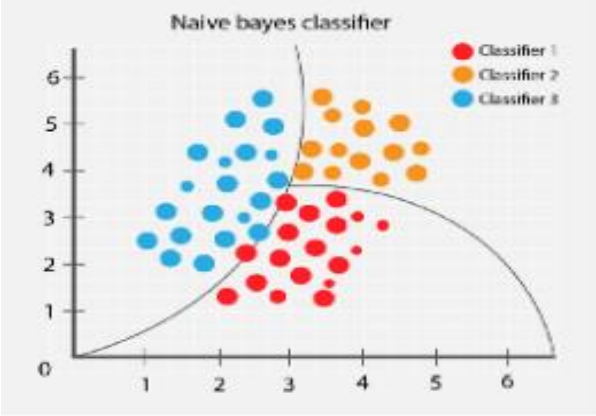

Figure 2. Naïve Bayes Classifier

4.3. Performance measures

There are some parameters based on which we evaluated the performance of the classifiers such as Accuracy (ACC), Precision, and Recall. The ACC of a classifier in a certain test range is the percentage of the test settings identified correctly by the classifier. The confusion matrix used to calculate these measures, which contain four useful parameters as, True Positive (TP): a positive example classified as positive, False Negative (FN): a positive example misclassified as negative, True Negative (TN): a negative example classified as negative, and False Positive (FP): a negative example misclassified as positive. The ACC can be calculated as in formula1. Precision is defined as what fraction of the recommended items the user consumed as in the following formula2. Recall defined as what, out of all the items that the user consumed, was recommended as in the following formula3. Provide sufficient detail to allow the work to be reproduced. Methods already published should be indicated by a reference: only relevant modifications should be described [37] [38].

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{\text{TP}}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

4.4. Methodology

We have combined the K-means technique and the NB classifier to enhance the accuracy of the classification model. The combination of this classification with the K-means clustering technique has shown promising improvements over previous methods, NB has become one of the most effective learning algorithms. Figure3 flowchart for the methodology of the hybrid model is presented. Which contains two algorithms, the first algorithm is K-Means clustering and the second algorithm is NB classifier. K-Means shows several steps to cluster cleaning dataset that is free of noise data or invalid data of employees, used in the analysis of cluster.

This data comes from previous works that are mentioned in literature reviews. The clustering and classification models have been established following the data preparation. To assess the efficacy of previous research, a group of attributes has been chosen. The attributes consist of personal information, Education information, and Professional information. These data were used to predict employee performance. In the first phase, the clustering process using the K-Means algorithm is performed first to determine training data. Employee's performance data collection has many classes, such as Excellent, Very Good, Good, Average, and Bad. In the second phase, data testing will be carried out using the NB algorithm. The centroid is the initial step in the clustering of K-mean results. This centroid uses the Euclidean distance equation to measure the distance between the data and the centroid, which is shown in formula4 [39] [40].

$$d(x,y) = \sqrt{(x_1 - y_n)^2 + \cdots + (x_n - y_n)^2} \quad (4)$$

Then describe the class cluster for every row of data. The average data for each variable and each cluster is determined after all data have been clustered. If the average data for each variable is not equal to the centroid value, then you must repeat the distance calculation until the average data is equal with the centroid value for each variable. The NB classifier with the K-mean clustering algorithm is used in this hybrid process. While NB delivers good classification results, it still needs to be improved in predicting employee performance, this hybrid approach works in a mixture of classification and clustering approaches. NB classifier can handle very high dimensional data and provides fast calculation training. The clustering algorithm in the hybrid approach requires the layout of the original dataset and adds to the training dataset the number of clusters, penalty factor parameters, and the kernel of NB. This framework aims at describing the number of clusters k. The irrelevant and unnecessary features are deleted to accelerate the calculation time. K-means partitions the data into k clusters and preserves the principal dataset distributions. In the last step, NB is applied.

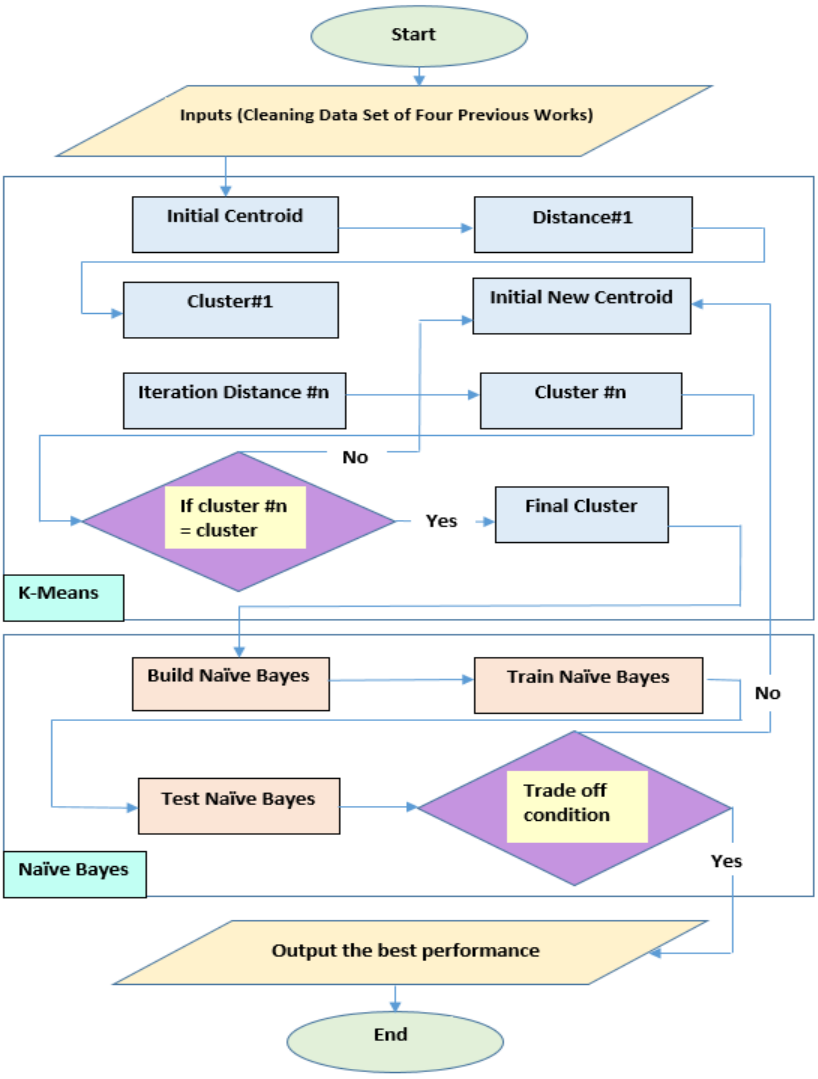

Figure 3. Flowchart for the Hybrid Proposed Framework

5. Results and discussion

This study applying the k-means clustering algorithm to group data then the data is classified using the NB algorithm. Data were taken from the previous works, which were used in this study as a test case. The dataset is tested by ignoring the original label then label the new data by grouping the data using the k-means algorithm. By applying clustering techniques to the original dataset, divide the data into the suitable number of clusters as a test tool as for the data of 145 employees divide it into 3 clusters according to the number of labels. Next predict the results of performance employee data using the NB classification algorithm. Compare the results of classification, clustering, and integration of classification. The result of the dataset employee performance which calculated used data tool weka can be seen in Table3. The results are surprisingly better in terms of ACC, precision, and recall which are calculated by using formulas 1, 2, and 3 as mentioned in the previous section of this paper.

Table 3. Experimental

Dataset	Algorithm	ACC%	Precision%	Recall%
130 Employee	K-Means	70.11%	80%	77%
	Naïve Bayes	65.80%	70%	69%
	K-Means+Naïve Bayes	80%	85%	79%
206 Employee	K-Means	84.50%	86.50%	87%
	Naïve Bayes	80.33%	82.44%	85%
	K-Means+Naïve Bayes	91.29%	89.90%	88.70%
310 Employee	K-Means	96.20%	92%	90%
	Naïve Bayes	95.48%	91.20%	88%
	K-Means+Naïve Bayes	98.56%	99%	99.80%
145 Employee	K-Means	85.70%	87.99%	89%
	Naïve Bayes	82.07%	85%	80.70%
	K-Means+Naïve Bayes	92.24%	90.70%	92%

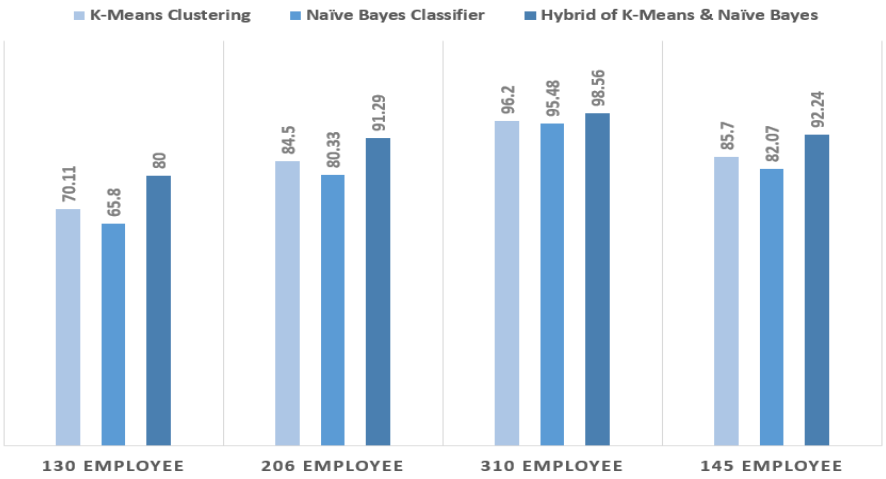

Figure 4. Comparison Results in Accuracy

Depend on the results in Table3 and the chart is shown in Figure4, which are shown above the proper result of K-means is best than NB also the combination of K-Means and NB is best than the other algorithm without combination. Clusters created by K-Means yet use the choice of random centroid because the first step of the K-Means clustering is by choosing a centroid value. The result of the initial random centroid description technique according to the K-Means, which is the original method is simpler and faster. By utilizing the NB classify the next test data. The phase of using the K-Means clustering creates the majority of data that has been grouped based on the original class. This is the impact of NB classification which requires sufficient training data to implement an optimal process of classification. The original K-Means randomly produce an initial centroid which makes the quality of grouping accuracy dependent on the initial centroid. When centroids are incorrect, the accuracy results will be relatively lower. By using an integration of the K-Means clustering algorithm and the NB classification, the process of determining the initial centroid K-Means also influences the accuracy results. However, the impact can be reduced by the addition of the NB classifier which results in better accuracy, although not better than the proposed method.

6. Conclusions and future work

This study proposed the integration of K-Means clustering and NB classification DM techniques in employee performance data to produce higher data accuracy. The proposed method gives better results. Although the initial centroid determination in the K-Means method is carried out randomly, the impact can be reduced by the addition of the NB classifier method resulting in better accuracy and increasing the accuracy of the existing methods. With the results obtained, it can be concluded that the proposed method can improve predictions of employee performance data. The initial centroid determination in the K-Means method affects that the quality of grouping accuracy depends on the initial centroid. For further research, the K-Means Clustering was applied on other algorithms that were used in the literature survey to show the result such as ID3, C4.5, and SVM.

References

[1] M. Xiao, F. Cooke, J. Xuc, H. Bian, "To what extent is corporate social responsibility part of human resource management in the Chinese context? A review of the literature and future research directions," Human Resource Management Review, pp. 30.4: 100726, 2020.

[2] E. M. Mone," Employee engagement through effective performance management: A practical guide for managers", Routledge, 2018.

[3] A. Bogarín, R. Cerezo, C. Romero,"A survey on educational process mining," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no.1, 2018.

[4] S. Bandaru, A. H.C.Nga, K. Deb, "Data mining methods for knowledge discovery in multi-objective optimization: Part A-Survey.", Expert Systems with Applications , pp. 139-159 , 2017.

[5] A.M.Hemeida, S.Alkhalaf, A.Mady, E.A.Mahmoud, M.E.Husseinc, A. M.Baha Eldin, "Implementation of nature-inspired optimization algorithms in some data mining tasks.", Ain Shams Engineering Journal ,vol.11, no. 2, pp. 309-318, 2020.

[6] P. Zschech , R. Horn , D. Ho¨schele , C. Janiesch , K. Heinrich,"Intelligent User Assistance for Automated Data Mining Method Selection," Business & Information Systems Engineering, pp. 1-21, 2020.

[7] C. Romero, S. Ventura, "Educational data mining and learning analytics: An updated survey." , Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery , vol. 10, no. 3, 2020.

[8] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," arXiv preprint arXiv, pp. 1707.02919 , 2017.

[9] A. K. Sahoo, C. Pradhan, H. Das, "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making," Nature Inspired Computing for Data Science. Springer, Cham, pp. 201-212, 2020.

[10] K. Shankar, S. K. Lakshmanaprabu, D. Gupta, A. Maseleno, V. H. C. de Albuquerque, "Optimal feature-based multi-kernel SVM approach for thyroid disease classification," The Journal of Supercomputing , pp. 1128-1143, 2020.

[11] G. Kou, P.Yang, Y. Peng, F. Xiao, Y. Chen, F. E.Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," Applied Soft Computing , vol. 86, 2020.
[12] M. Nabipour , P. Nayyeri, H. Jabani, S. Shahab, A. Mosavi, "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," IEEE Access, pp. 150199-150212, 2020.
[13] M. Richard, "Statistical rethinking: A Bayesian course with examples in R and Stan," CRC press, 2020.
[14] J. J. Dziak, D. L. Coffman, S. T. Lanza, R. Li, L. S. Jermiin, "Sensitivity and specificity of information criteria," Briefings in bioinformatics , pp. 553-565, 2020.
[15] A. Jagannathan, "Determinants of employee engagement and their impact on employee performance" International journal of productivity and performance management ,2014.
[16] M. S. Amin, Y. K. Chiam, K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," Telematics and Informatics, pp. 82-93, 2019.
[17] B. Sasan, and T. Mokfi, "Evaluation and selection of clustering methods using a hybrid group MCDM.", Expert Systems with Applications, vol.. 138, 2019.
[18] A. Qasem and E. Al Nagi , "Using data mining techniques to build a classification model for predicting employee's performance," International Journal of Advanced Computer Science and Applications, pp. 3.2, 2012.
[19] M. John, and C. A. Moturi, "Application of data mining classification in employee performance prediction," International Journal of Computer Applications, pp. 28-35, 2016.
[20] R. Jayadi, lH. Firmantyo,"Employee Performance Prediction using Naïve Bayes," International Journal of Advanced Trends in Computer Science and Engineering, pp. 3031- 3035, December 2019.
[21] M. Nasr, E. Shaaban, A. Samir, "A proposed Model for Predicting Employees' Performance Using Data Mining Techniques: Egyptian Case Study,"2019.
[22] M. M. Fard, T. Thonet, E. Gaussier, "Deep k-means: Jointly clustering with k-means and learning representations," Pattern Recognition Letters, pp.185-192, 2020.
[23] S. Chakraborty, D. Paul, S. Das, J. Xu, "Entropy weighted power k-means clustering," International Conference on Artificial Intelligence and Statistics PMLR, 2020.
[24] H. Yu, G. Wen, J. Gan, W. Zheng, C. Lei, "Self-paced learning for k-means clustering algorithm," Pattern Recognition Letters, pp. 69-75, 2020.
[25] G. Manoj Kumar, and P. Chandra ,"An empirical evaluation of K-means clustering algorithm using different distance/similarity metrics," Proceedings of ICETIT 2019. Springer, Cham, pp.884-892, 2020.
[26] C. Xia, J. Hua, W. Tong, S. Zhong, "Distributed K-Means clustering guaranteeing local differential privacy," Computers & Security , pp. 101-699, 2020.
[27] M. Punniyamoorthy, and R. K. Jeyachitra., "Development of new seed with modified validity measures for k-means clustering," Computers & Industrial Engineering, pp.141: 106290, 2020.
[28] P. Govender, and V. Sivakumar, "Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)," Atmospheric Pollution Research , pp. 40-56, 2020.
[29] S. A. Abdulrahman, W. Khalifa, M. Roushdy, A. M.Salemb, "Comparative study for 8 computational intelligence algorithms for human identification," Computer Science Review, vol.. 36, pp.100237, 2020.
[30] A. Husejinović, "Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers," Periodicals of Engineering and Natural Sciences, vol. 8, no. 1, pp.1-5, January 2020.
[31] S. Kumar, D. Jayadevappa, M. V. Shetty, "A Novel approach for Segmentation and Classification of brain MR Images using Cluster Deformable Based Fusion Approach," Periodicals of Engineering and Natural Sciences, Vol.6, No.2, pp. 237-242, December 2018.
[32] F. Xu, Z. Pan, R. Xia, "E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework.", Information Processing & Management, pp. 102221, 2020.
[33] S. Chen, G. I.Webb, L. Liu, X. Ma, "A novel selective naïve Bayes algorithm," Knowledge-Based Systems, vol.. 192, pp. 105361, 2020.
[34] Y. Choi, G. Farnadi, B. Babaki, G.V. Broeck, "Learning fair naive bayes classifiers by discovering and eliminating discrimination patterns," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34. no. 06. 2020.

[35] S. Kumar, D. Jayadevappa, M. V. Shetty , “A Novel approach for Segmentation and Classification of brain MR Images using Cluster Deformable Based Fusion Approach,”Periodicals of Engineering and Natural Sciences, vol.6, no.2, pp. 237-242, December 2018.
[36] A. Askari, A. d’Aspremont, L. El Ghaoui, "Naive feature selection: Sparsity in naive Bayes." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.
[37] V. H. Nhu,A. Shirzadi , H. Shahabi, ,S. K. Singh., "Shallow Landslide Susceptibility Mapping: A Comparison between Logistic Model Tree, Logistic Regression, Naïve Bayes Tree, Artificial Neural Network, and Support Vector Machine Algorithms," International Journal of Environmental Research and Public Health ,vol.17, no. 8, 2020.
[38] M. M. Musleh, E. Alajrami, A. J. Khalil, Bassem S. Abu-Nasser, A. M. Barhoom, S. S. Abu Naser, "Predicting Liver Patients using Artificial Neural Network," 2019.
[39] R. Jayanthi, and Lilly Florence, "Software defect prediction techniques using metrics based on neural network classifier," Cluster Computing, pp. 77-88, 2019.
[40] G. M. Kumar, and P. Chandra, "An empirical evaluation of K-means clustering algorithm using different distance/similarity metrics," Proceedings of ICETIT 2019, Springer, Cham, pp.884-892, 2020.