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, Baghdad, Iraq

ABSTRACT

For businesses to be successful, they need to be able to predict an employee's future performance. Because the employee is the most important component of an organization, and the failure or success of an organization is depending on an employee's performance, this has become a major concern for decision-makers and managers in almost all types of businesses when it comes to putting plans in place to locate a team of experts. As a result, management gets personally invested in the accomplishments of these workers. In particular, to ensure that the most convenient employment is given to the most qualified candidate at the most appropriate moment. Analytical forecasting is a current trend in human resources. Data mining is helpful in the realm of predictive analytics. Clustering of K-Means and classification of Nave Bayes (NB) are included in the proposed framework for better outcomes in the analysis of employee performance information, running in tool of WEKA, in order to make better-informed decisions on the performance of their employees, combining NB and K-Medoids improves the ACC of forecasting performance of employee. When compared to K-Means and NBs. The proposed methodology enhances employee performance prediction accuracy.

Keywords: NB, K-Means, Performance

Corresponding Author:

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

1. Introduction

Managers in practically every industry, including government, commercial enterprise, and higher education, have turned to human resources as a top priority [1]. Corporate employers are concerned about selecting the right staff for the right time slots. After hiring personnel, in order to keep the top performers, management was concerned about their performance [2] by implementing performance systems. DMis a young and promising knowledge finding and information domain [3].

When you use database management (DM) techniques, you can learn how to get information out of the database and get knowledge out of it. People who work for DM do a lot of different things, like clustering and classifying. The classification approaches are supervised learning algorithms that divide data into limited labels. It's one of classifying raw data using the most important DM methods[8] [9]. In order to foresee future data patterns, classification algorithms frequently use models [10]. NB [11] is one of the most well-known classification algorithms. [12] Classification of the target class is another way used to anticipate it. It is based on probabilities, but it also gives a specific approach for developing other algorithms of learning [13]. As a result, the results of this classification are more exact, productive, and relevant to new data introduced into the dataset [14].

A successful firm is one that is able to recognize and reward its employees for their efforts. Because each employee performs differently, not all employees should be treated the same [15]. By utilizing its algorithms, DM can support the business. The clustering algorithm and the NB of the data mining technique can be employed to find the important proof features of an organization's future forecast [16]. For future prediction,

clustering is the most often utilized method for grouping data into groups having similar proof characteristics as a way to optimize the degree of inter-class similarity or decreased [17].

2. Statement of problem

Evaluation of employee performance is one of the more challenging processes because it aims to inform each employee about his or as well as making decisions about her future career advancement and salary. It also identifies parts of the a location of business that has to be upgraded or modified. Resources of human in the public sector of most other firms utilize typical evaluation methods that don't allow them to accurately measure their results. As a result, several past studies have used algorithms of supervised classification in data mining to build a performance forecasting model for their staff. Multiple factors that may or may not be beneficial in class prediction are included in the category-dependent variable dataset are included in supervised classification problems. Unsupervised learning analyzes an unlabeled dataset and attempts to uncover a latent structure in it, such as grouping. Using clusters in the data set, this research shows how to improve the classifier's performance by employing k-means clustering.

3. Literature survey

In this 2012 study, data mining was applied to construct a staff efficiency can be predicted using a grading model. Data Mining (DM) was employed in the building of classification model to create rules that was gathered via a questionnaire to 130 IT personnel. Many tests were conducted to validate the constructed model using Three algorithms were evaluated using Hold-out (60 percent) and Fold Cross-Validation: ID3 accuracy of 50 percent and 43.7 percent, C4.5 (J4.8) accuracy of 60.5 percent and 56.2 percent, and Nave Bayes accuracy of 65.8% and 68.7% [18]. Employee performance data from the Kenya School of Government's Human Resources Department was used to compile this study in 2016. Classification was carried out using three different algorithms from DM, including NB, C4.5, and ID3 to find the best one. The data was collected over a five-year period and consisted of 206 assessment reports based on 14 performance criteria. There are two datasets for this collection of information. When comparing the accuracy of multiple categorization systems, it was found that ID3 was 64.5 percent accurate, NB was 80.33 percent accurate, and C4.5 (J4.8) was 82.60 percent accurate. In 2019, this research proposed the forecasting of an employee's performance in a company using the categorization of the NB technique, which was used to construct the prediction model. Details about 310 employees were included in the analysis. Among the dataset's parameters are 28. According to the data, NB accurately evaluated 95.48 percent of the cases [20]. In 2019, this article looks at the possibility of using classification algorithms to actual data acquired from the Ministry of Civil Aviation of Egypt during a survey of 145 employees to construct a predictive employee performance model. After data preparation and preprocessing, classification begins. Applying algorithms, NB, DT, and SVM to create an employee's performance prediction model, it was shown that the best results for ACC of C4.5 (J48), Nave Bayes and SVM were equal to 79.31 percent, 82.07 percent, and 86.90 percent correspondingly [21].

4. Methods and materials

4.1. Clustering of K-means

It is a popular clustering method. This technique is the most often used method for research and industrial clustering [22]. k centroids, one for each cluster, are the primary focus of k-means analysis [23]. These centroids must be positioned in a novel way because of numerous reasons. As a result, the safest choice is to keep them as far apart as possible [24].

In a given data collection, each data point corresponds to a new center and represents the next phase in the process. The first step is taken if there is no need for an early group. In this situation, the cluster center should be updated from the beginning. Figure1 depicts the K-Means in broad description [24] [25] after finding . We'll use these new centroid coordinates to create a new link between our existing datasets and the one nearest to us. We've formed a loop, which implies the k centroids are gradually shifting their positions until they can't move any further. In other words, centroids are no longer shifting. After a certain point, the total can no longer be reduced any further using this procedure. There is a result in Figure1 [26] of a group of clusters that are compact. This algorithm's goal is to reduce to the minimum possible value [27]. It is most commonly utilized in

applications such as market research, forecasting, etc. The K-means Algorithm [28] [29] has numerous advantages and downsides, some of which are listed in Table1.

Table 1. A Comparison of K-means' Pros and Cons Clustering

Advantages	Disadvatages
Simple to implement.	The number of clusters is difficult to anticipate (K-Value).
K-Means may be more computationally efficient than hierarchical clustering when there are many variables (if K is small).	The first seeds have a significant impact on the eventual outcome.
If you're looking for the best clusters, k-Means may be the way to go.	The sequence in which the data is presented has an effect on the final outcome.

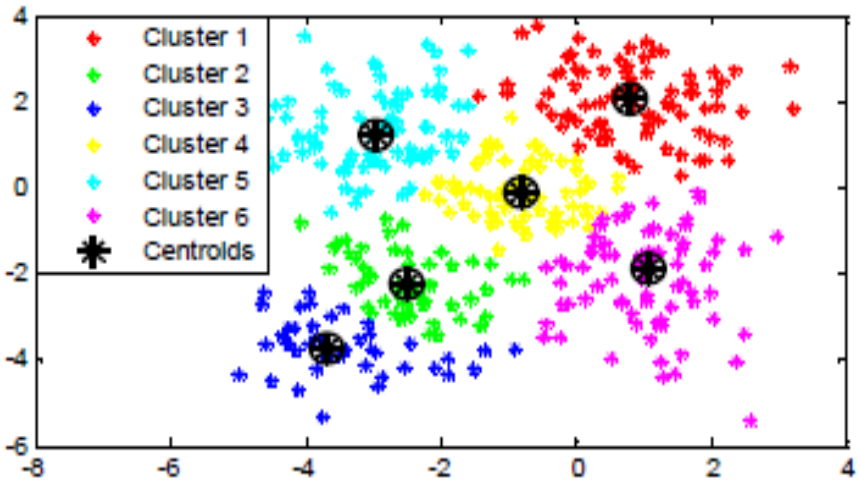

Figure 1. K-Means Clustering [31]

4.2. Classifier of naïve bayes

It is a statistical categorization model that is based on maximum posterior hypothesis and bayes theorem. This system of categorization is so widely used because it is so simple to use [30]. According to probability computations, NB can forecast the chance of membership in a class of tuple data that will be entered into a certain class. In the field of machine learning, this method is frequently employed because it is known to be accurate with simple computations. NB relies on a strong and straightforward premise of independence in its construction. There are three clusters in the NB classifier, and each one is further subdivided into more precise categories (Figure 2). Using Nave Bayes algorithms has several advantages and disadvantages, some of which are listed in Table2 [35] [36].

Table 2. The Benefits and Drawbacks of the Nave Bayes Classifier

Advantages	Disadvatages
This classifier is extremely accurate if the independent assumption holds.	If the independent assumption is not met, performance will be very poor.

It's simple to implement because all that needs to be done is calculate the likelihood.	When the probability of a feature in a class is zero, smoothing is an over-head and a must-do step.
Text classification, for example, benefits from its high dimensionality.	Due to the product of several small probabilities (e.g. 0.053), vanishing value is also a concern.

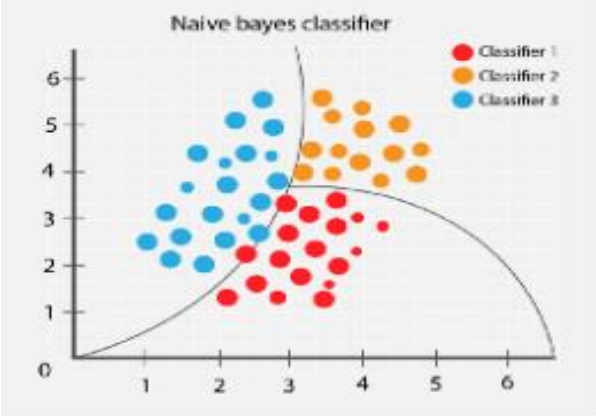

Figure 2. Naïve Bayes Classifier

4.3. Measures of performance

We examined the performance of the classifiers using several criteria, including Accuracy (ACC), Precision, and Recall. An algorithm's accuracy in a given test range is measured by the classifier's accuracy coefficient (ACC). Confusion matrices are used to determine the relevant parameters such as (TP) True Positive : a positive example that has been labeled, (FN) False Negative: incorrectly labeled as a bad example, (TN) True Negative : a negative example that is labeled , and (FP) False Positive: example that should have been classed as negative but was instead is now considered positive. Formula 1 can be used to compute the ACC. According to the formula2 below, precision is defined as the percentage of advised products that were ingested by the consumer. Recall is defined as the following formula3 as to which items the user ingested out of all those advised. Provide enough information for the work to be duplicated. Only relevant adjustments should be stated when referencing previously published methods [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. Proposed Method

In order to enhance the categorization model's accuracy, we merged the NB with K-means. When this classification is used with K-means clustering, promising results have been obtained, making NB one of the most successful learning techniques. A flow-chart for the hybrid model method is shown in Figure 3. K-Means clustering and NB classifier are the first two algorithms. K-Means demonstrates numerous procedures for cleaning a cluster analysis dataset of any noise or erroneous employee data. This information is derived from past studies that have been mentioned in reviews of literature. Following data preparation, as a result, we've worked on creating classification and clustering models.A set of characteristics has been chosen to assess the efficacy of prior studies. Personal information, educational background, and professional experience are all included in the qualities. This information was utilized to forecast employee performance. To determine training data, the clustering procedure utilizing the K-Means algorithm is conducted initially in the first phase. There

are numerous levels of performance evaluation for employees, in the form of the adjectives “excellent,” “very good,” “good,” “average,” or “bad,”when it comes to collecting performance data. The NB method will be used to test the data in the second phase. In the K-mean, the centroid is the first step. Using formula4, this centroid calculates the distance between the centroid and data points.

$$d(a,b) = \sqrt{(a_1 - b_n)^2 + \cdots + (a_n - b_n)^2} \quad (4)$$

Then, for each row of data, explain the class cluster. When all of the data has been grouped, aggregated values for each metric and each cluster is calculated. If the centroid value for each variable is not equal to the average values for that variable, the distance calculation must be repeated until the average data for each variable equals the centroid value. Use of the NB classifier and in this hybrid technique, the K-means clustering algorithm is used. is explained. It's still necessary to improve NB's prediction of employee performance even when it's good at classifying employees. This hybrid strategy is a blend of clustering and classification. Naïve bayes can proceed data with a lot of dimensions and can learn quickly. The hybrid approach's clustering method necessitates the original dataset's layout and penalty factor parameters, and the number of clusters, and the NB kernel to the training dataset. To put it another way, the number of clusters k is the focus of this framework. The calculation time is sped up by removing superfluous and redundant features. Data is partitioned into k clusters using K-means, which maintains its original dataset structure in the process. The final step is to apply NB [39] [40].

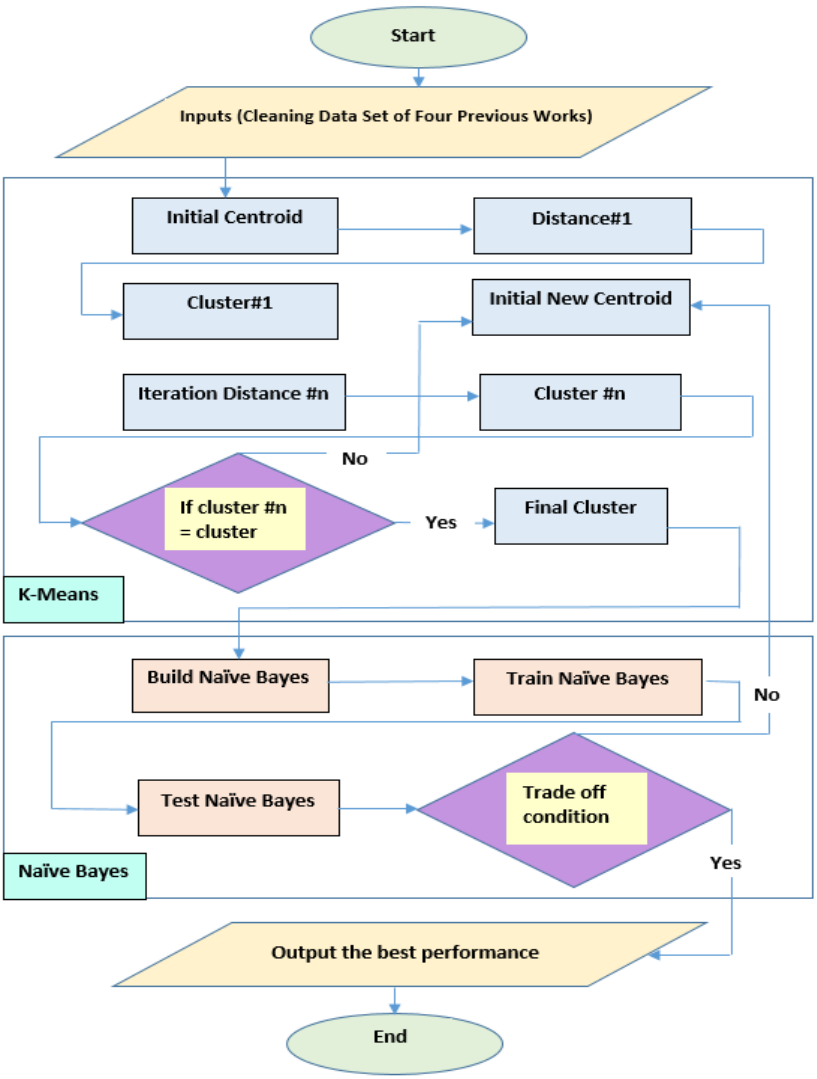

Figure 3. depicts the proposed hybrid framework's flowchart

5. Results and discussion

Clustering of K-means is used to assemble the information, and then the NB algorithm is used to classify it. Data was obtained from prior publications, which were employed as a test case in this study. When the dataset is put to the test, the previous label is ignored and the new data is labeled with k-means clustering. As a test tool, use to the original dataset using clustering techniques and the data should be broken down into the appropriate number of groups, such as dividing the data of 145 employees into three clusters based on the total number of labels. Next, make use of the NB algorithm for classification to anticipate the findings of employee performance data. Investigate how categorization might be improved through clustering and integration. Table 3 shows the findings of the employee performance data that was computed using theWeka. Using formulas 1, 2, and 3 from the earlier section of this research, in terms of ACC, precision, and recall, the results are shockingly superior.

Table 3. Result

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%

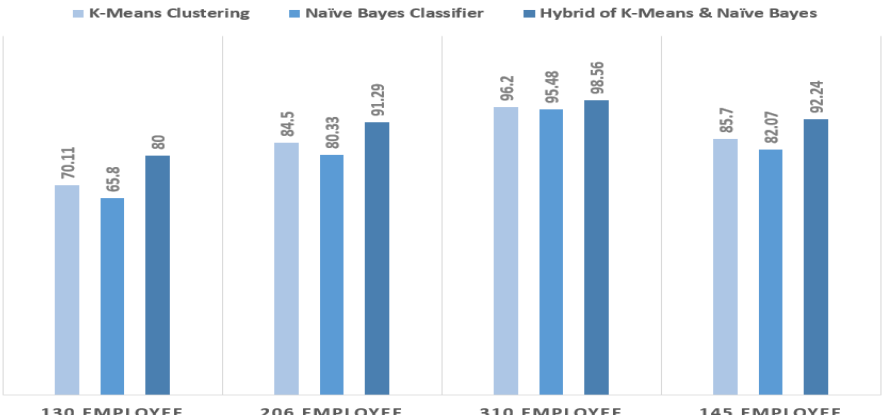

Figure 4. The comparison shows which method is more accurate

According to the data in Table 3 and the graph in Figure 4, K-means is superior to NB in terms of accuracy, and the combination of NB and K-Means outperforms the other algorithm on its own. Because the initial step in K-Means clustering is to choose a centroid value, clusters generated by this algorithm still employ a random centroid. The original method, which uses K-Means to describe a random centroid, yields a more complex and time-consuming outcome. In order to classify the next set of tests, use the NB. The most of the data according to the original class, they've been placed together in the phase where K-Means clustering is used. This is the effect of NB classification, which necessitates a large amount of training data in order to perform an optimal classification procedure. K-Means algorithm generates an strating centroid at random, making the quality of grouping accuracy reliant on it. The accuracy results will be lower if the centroids are inaccurate. K-Means and the naïve bayes are integrated so that the procedure for calculating the centroid's initial position K-Means has an impact on the accuracy. The impact can be decreased, though, by using the NB classifier, which provides greater accuracy, though not as good as the proposed technique.

6. Conclusions and future work

In order to improve data accuracy, this study advocated combining K-Means and NB. Data mining algorithms with data on employee performance. The proposed method yields more accurate findings. Although the K-Means technique's initial centroid determination is done at random, the impact can be decreased by using the NB classifier approach, which improves ACC and increases the accurateness of the current techniques. Depend on the results, it can be inferred that the suggested strategy can enhance data on employee performance forecasts. In the K-Means approach, the initial centroid determines the quality of grouping accuracy, which is dependent on the starting centroid. K-Means Clustering was utilized to illustrate the results of other algorithms employed in the literature study, such asSVM, C4.5, and ID3 for future investigation.

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] D. A. Nur Wulandari, et al., "An Educational Data Mining For Student Academic Prediction Using K-Means Clustering And Naïve Bayes Classifier," Journal Pilar Nusa Mandiri, Vol. 16, No. 2, Septemper 2020.
[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.