Research Article

International Journal of Advanced Computer Research, Vol 11(57)
ISSN (Print): 2249-7277   ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/ IJACR.2021.1152043

KMK based hybrid approach for the performance estimation in case of diabetes data

Yasir Minhaj Khan[*] and Animesh Kumar Dubey
Department of Computer Science, PCST Bhopal, Madhya Pradesh

Received: 20-September-2021; Revised: 25-November-2021; Accepted: 27-November-2021

Abstract

In this paper k-means clustering algorithm has been used with k-points (KMK) selection. It has been applied on the PIMA Indian diabetes dataset. It has been used for distance estimation, centroid selection, effect of data size variations and for the analysis of the complete record. The cluster section has been found to be improved based on k-point selection. It has been used for the assignment of initial centroid. The results indicate that the KMK algorithm is capable in the improvement of centroid selection and distance measures in the assignments of data points. It is due to the better centroid selection mechanism by k-points selection based on the weight measures from the selected dataset. So, the obtained clusters are better in comparison to k-means.

Keywords

K-means, KMK, PIMA, Similarity score, Centroid estimation.

1.Introduction

In the current situation diabetes is the leading cause of death with the ranking of ninth among all diseases worldwide [1]. An estimation of approximately 1.5 million deaths by this disease [1]. The reports suggest that the diseases can be handled in better way if it is detected in the early stages [2−7]. There are several algorithms including machine learning algorithms can be helpful in disease diagnosis and detection including diabetes [8−13]. These algorithms are k-nearest neighbors (KNN), logistic regression (LR), Naïve Bayes (NB), support vector machine (SVM), Decision Tree (DT), random forest (RF), convolutional neural network (CNN), etc. [10−18]. The clustering algorithms are also helpful in the same for these purposes like k-means and fuzzy c-means (FCM). The main aim of these approaches is to detect and predict the diseases in the early stage with higher accuracy. So, the main aim of this paper is to apply the combination of clustering and classification. In this approach k-means and the combination of k-points have been considered for the betterment. Clustering algorithm will work better if the centroid estimation is accurate. So, our main aim is to improve the estimation through k-points.

So, in this paper the main focus of our work is relies on the clustering and classification techniques. So that a hybrid approach can be developed to collaborate the capability of both and which will be helpful in the performance estimation of the diabetes data.

2.Literature survey

In 2021, Khanam and Foo [19], applied the machine learning algorithms, neural networks and data mining techniques on the collected data. They used the data of 768 patients with nine unique attributes from Pima Indian Diabetes (PID) dataset collected from UCI repository of machine learning. The data was pre-processed by WEKA tool. They applied the KNN, LR, NB, SVM, DT, RF and AdaBoost as the machine learning classification algorithms. They used the K-fold cross validation for testing the machine learning performance. A neural network with one, two and three hidden layers was developed. They found the accuracy greater than 70% in all the applied classifiers. The support vector machine and logistic regression provided 77%-78% of accuracy in both training and testing dataset. The neural network model was considered as efficient for analyzing the diabetes with the accuracy of 86%.

[*] Author for correspondence

In 2021 Frimpong et al. [20], proposed a model of feedforward artificial neural network (FFANN). They implemented it with the dense neural network having three hidden layers. Each layer of this network was activated by using the rectified linear unit (ReLu). They considered the dataset of 768 from Pima Indian Diabetes having type-2 diabetes with eight attributes. The evaluation parameters considered were precision, accuracy and recall. They found the highest accuracy of 97.27% in training and 96.09% of accuracy in testing dataset for mainly type 2 diabetes patients. They have compared their results with other existing models of same research area.

In 2021 Rasheed et al. [21], proposed a framework of machine learning for detecting the COVID-19 from the X-ray images. They used the generative adversarial network (GAN) for sample increasing and reducing the overfitting problem. The 198 X-ray images were provided by Joseph Paul Cohen pertaining to affected cases of COVID-19. They also considered the 210 X-ray images of Chest of the healthy individuals for the comparison. The LR and CNN were selected as the machine learning approach. They investigated the dimensionality reduction approach based on the Principal Component Analysis (PCA). The F1-score, recall and accuracy were used as the evaluated performance metrics. The CNN and LR achieved the accuracy of 97.6% and the 95.2%. They found that the proposed methods attained the higher accuracy of 100% when used with CNN and PCA and having variance of 0.99.

In 2021 Jackins et al. [22], took the three different datasets of cancer, diabetes and heart diseases from the wearable devices and the online repository. They applied the Gaussian NB and RF algorithm to estimate the system’s performance against the input data. The comparison of result among two applied algorithms was performed and random forest performs well and gave the accurate results in classifications. The performance measures used in comparison by the algorithm were F1-score, recall and accuracy. The accuracy of 74.03 was achieved in random forest for diabetes dataset, accuracy of 83.85 for heart disease dataset and accuracy of 92.40 for the cancer dataset. Further, these results were compared with density based spatial clustering of applications with noise (DBSCAN) and K-means clustering to identify the effectiveness of the algorithm. It was found that after comparing the results the random forest performed well. They mentioned that as a limitation their model has the more processing time.

In 2021 Marcos-Zambrano et al. [23] proposed a review of different machine learning algorithms applied to the human microbiome. They discussed different supervised, unsupervised and clustering techniques which are helpful in medical field. They discussed the various resources available for dataset and different methods and approaches helpful in detection and prediction of disease.

In 2021 Islam et al. [24] proposed an approach for predicting whether the person is diabetic or not from the retina images. They considered two datasets. One was the larger dataset taken from EyePACS having 80,000 retinal images and a smaller dataset from QBB retina images. They developed a multi-stage CNN. The performance was evaluated with the combination of Gradient Boosting Machines (GBMs). XGBoost (XGB) was used for the GBM implementation. They achieved the accuracy of 84.47 from DiaNet and 80.65 from QBBNet. Due to the limited number of retinal images the expected accuracy was less.

In 2021 Rahman et al. [25] identified some pathogentic processes and pathways arise between Schizophrenia and type 2 diabetes mellitus (T2DM). They analyzed mononuclear cell data from Schizophrenia and T2DM. They used the two GSE18312 and GSE27383 as the dataset of Schizophrenia and performed meta-analysis on them by using the ImaGeo web-utility. The GSE9006 dataset was considered for T2DM. It was pre-processed and analyzed by Web-utility of NetworkAnalyst. They selected the differentially expresses genes (DEGs). They revealed 28 genes dyregulated in T2DM and schizophrenia; that potentially promote the T2DM in schizophrenia patients. They predicted some regulators between T2DM and schizophrenia.

In 2021 Ishaq et al. [26] analyzed the dataset of 194 men and 105 women derived from UCI machine learning repository. They tried to identify the significant features, some data mining techniques for boosting the accuracy for prediction. Decision tree, logistic regression, random forest, stochastic gradient classifier, gradient boosting algorithm, support vector machine, Gaussian naïve bayes, extra tree classifier, Adaptive boosting classifier were employed and compared. To handle the class imbalance problem, they applied the synthetic minority oversampling technique (SMOTE). They considered balanced dataset and evaluated them on the metrics of recall, accuracy, f-score and precision. It has been found that extra tree classifier outperformed from other models with an accuracy of 92%, recall and f-score of 0.933 by using nine significant features of SMOTE technique. They advised to use meta-heuristic for better feature selections arise due to NP-hard nature.

The overall review analysis focus on the efficient prediction and detection of the diseases. Although there are several approaches are already there in the direction for the same purpose. But the major problem is the symptoms variations and the applicability of the approaches.

3.Proposed work

In this paper k-means clustering algorithm has been used with k-points (KMK) selection with the help of KNN algorithm. It has been used for distance estimation, centroid selection, effect of data size variations and for the analysis of the complete record. Steps and the procedure for the KMK are shown below:

KMK algorithm steps:
Step 1: Initialize the k-point centroid.
Step 2: Randomly assign the centroid based on the k-point selection based on the weight vector from the training set.
Step 3: For each individual iterations, determine the minimum and maximum difference based on distance calculation.
Step 4: Iterate the cluster till the n-1 data points.
Step 5: Check the minimum variance and similarity score.
Step 6: Assign the data points in the appropriate clusters.
Step 7: Compute the performances.

For the analysis of the approach, PIMA Indian diabetes dataset was considered. At the initial stage data preprocessing has been performed. Then by the help of KNN algorithm, k-points have been selected for the centroid initialization in the k-means algorithm. For the similarity estimation different distance algorithms have been considered like Euclidean, Pearson Coefficient, Chebyshev and Canberra. Our approach is capable in analyzing the distance performances as well as the data size impact. The complete procedure is shown in *Figure 1*.

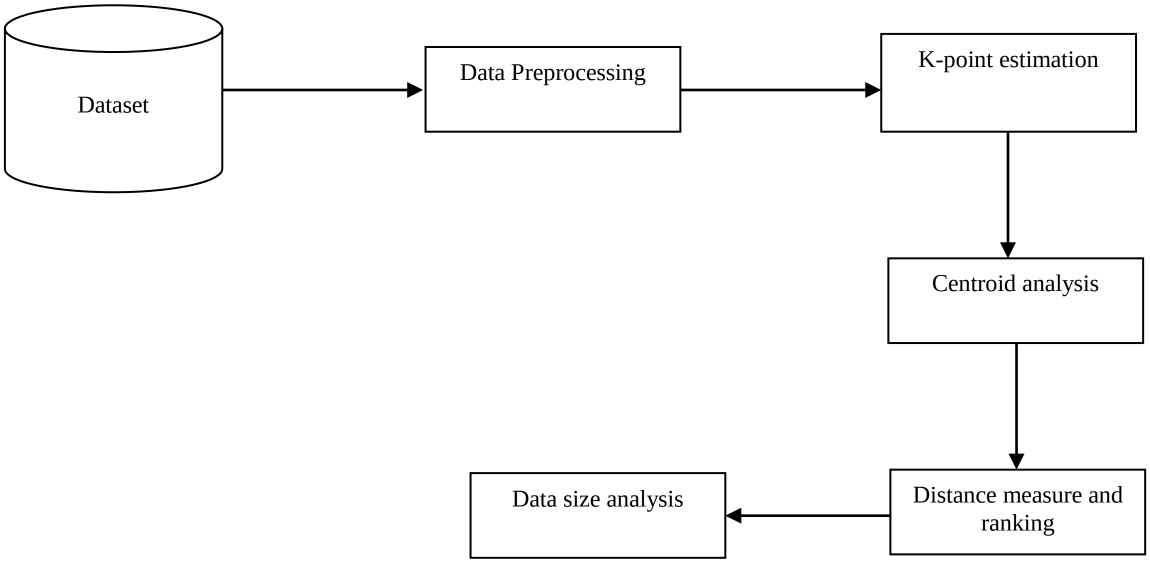

Figure 1 Overall process formation and steps of the approach

4.Results and discussion

Positive predictive value has been considered for the analysis of k-means and KMK. It is calculated by true positive/true positive + false positive. *Figure 2* shows the PPV of k-means based on random variable. *Figure 3* shows the average PPV of k-means based on random variable. *Figure 4* shows the PPV of KMK based on random variable. *Figure 5* shows the average PPV of KMK based on random variable. The results indicate that the maximum accuracy achieved through the k-means algorithm is approximately 82% and from the KMK it is approximately 91%. It is due to the better centroid selection mechanism by k-point selection. So, the obtained clusters are better in comparison to k-means. Hence the performance has been improved from our approach.

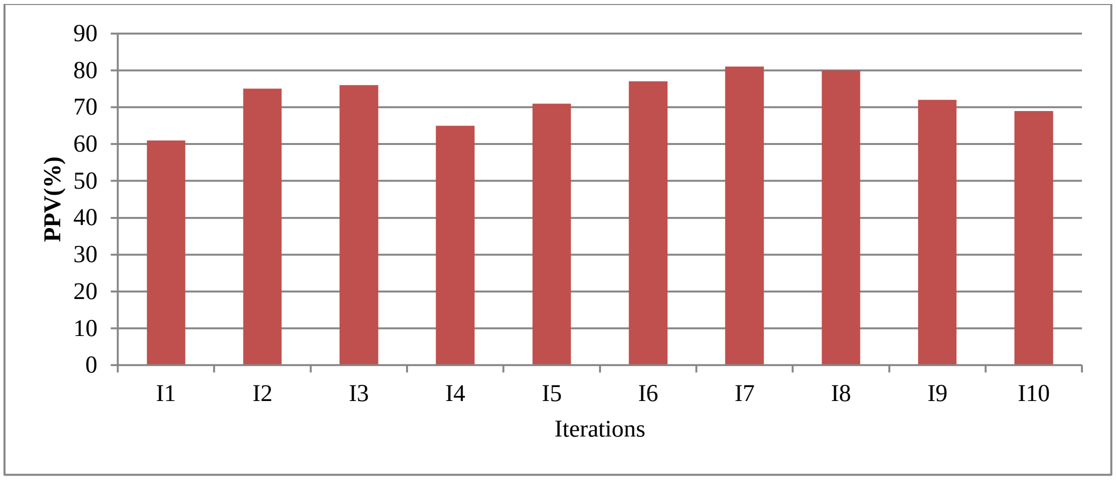

Figure 2 PPV of k-means based on random variable

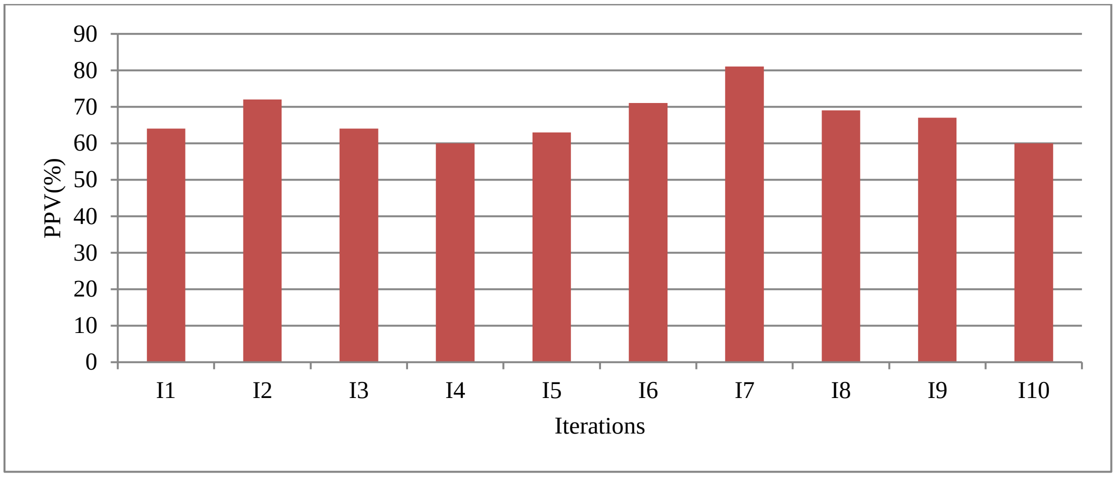

Figure 3 Average PPV of k-means based on random variable

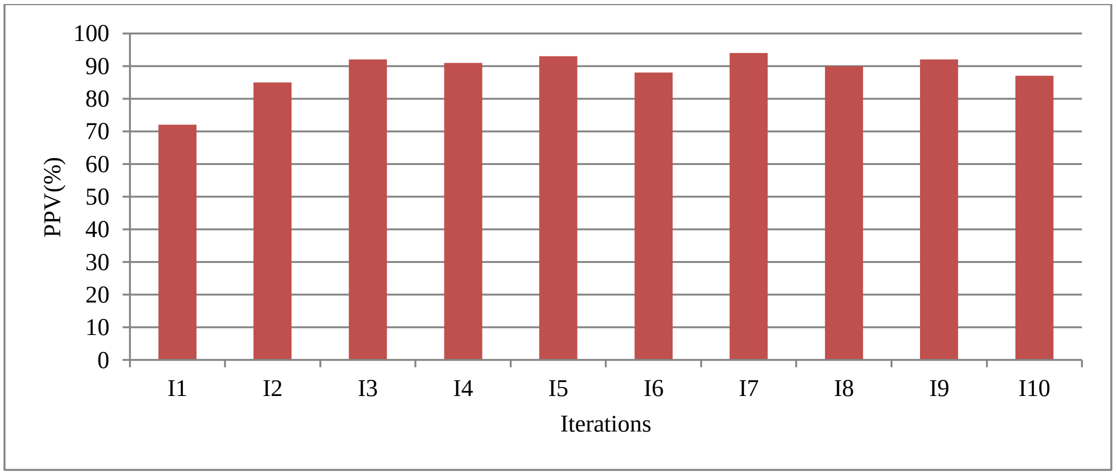

Figure 4 PPV of KMK based on random variable

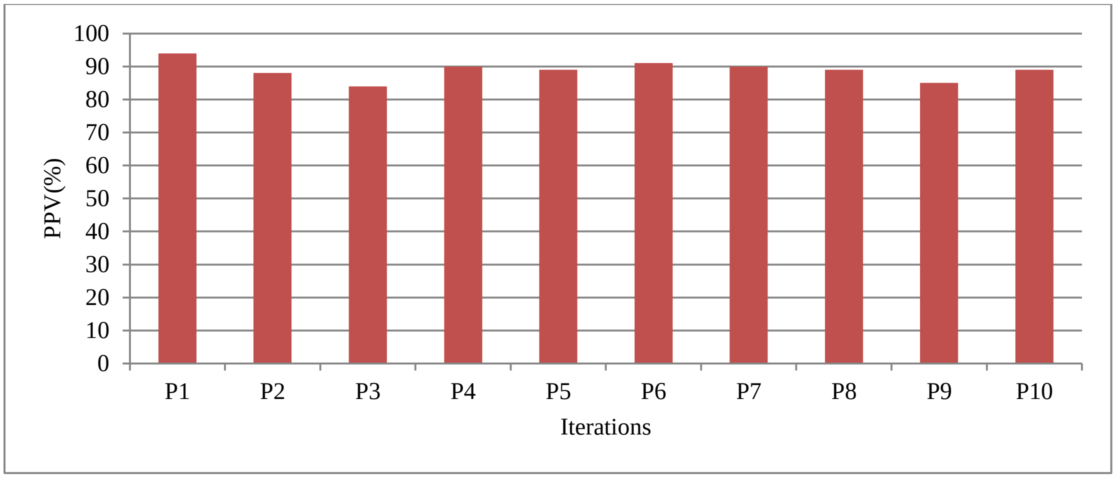

Figure 5 Average PPV of KMK based on random variable

5.Conclusion

This paper explores the k-means algorithm to improve the data point assignment mechanism. For this purpose, k-means algorithm has been combined with k-points in the initial phase. The k points are calculated based on the weight matrix from the complete training data. It is helpful in the segmentation of the complete attributes. So, the initial centroids are capable in improving the similarity score at the time of cluster assignment. The results are found to be prominent in case of KMK approach.

Acknowledgment

None.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

[1] https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 12 January 2022.

[2] Rosy JV, Kumar SB. Optimized encryption based elliptical curve Diffie-Hellman approach for secure heart disease prediction. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(83):1367- 82.

[3] Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. International journal of computer assisted radiology and surgery. 2016; 11(11):2033-47.

[4] Mansour AM. Decision tree-based expert system for adverse drug reaction detection using fuzzy logic and genetic algorithm. International Journal of Advanced Computer Research. 2018; 8(36):110-28.

[5] Dubey AK, Gupta U, Jain S. Epidemiology of lung cancer and approaches for its prediction: a systematic review and analysis. Chinese Journal of Cancer. 2016; 35(1):1-3.

[6] AbdelMaksoud E, Barakat S, Elmogy M. Diabetic retinopathy grading system based on transfer learning. arXiv preprint arXiv:2012.12515. 2020.

[7] Dubey AK, Gupta U, Jain S. Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. International Journal on Advanced Science, Engineering and Information Technology. 2018; 8(1):18-29.

[8] Khandelwal A, Jain YK. An efficient k-means algorithm for the cluster head selection based on SAW and WPM. International Journal of Advanced Computer Research. 2018; 8(37):191-202.

[9] Pei J, Han J, Lu H, Nishio S, Tang S, Yang D. H-mine: hyper-structure mining of frequent patterns in large databases. In proceedings international conference on data mining 2001 (pp. 441-8). IEEE.

[10] Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for mining heterogeneous data with dynamic support. In CSI sixth international conference on software engineering (CONSEG) 2012 (pp. 1-6). IEEE.

[11] Babu DB, Prasad RS, Umamaheswararao Y. Efficient frequent pattern tree construction. International Journal of Advanced Computer Research. 2014; 4(1):331-6.

[12] Li K, Cui L. A kernel fuzzy clustering algorithm with generalized entropy based on weighted sample. International Journal of Advanced Computer Research. 2014; 4(2):596-600.

[13] Dubey AK, Gupta U, Jain S. Medical data clustering and classification using TLBO and machine learning algorithms. CMC-Computers Materials & Continua. 2022; 70(3):4523-43.

[14] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics. 2018:515.

[15] Jamil A, Salam A, Amin F. Performance evaluation of top-k sequential mining methods on synthetic and real datasets. International Journal of Advanced Computer Research. 2017; 7(32):176.

[16] Chahar R, Dubey AK, Narang SK. A review and meta-analysis of machine intelligence approaches for mental health issues and depression detection. International Journal of Advanced Technology and Engineering Exploration. 2021; 8(83):1279- 1314.

[17] Lan GC, Hong TP, Tseng VS. An efficient projection-based indexing approach for mining high utility itemsets. Knowledge and Information Systems. 2014; 38(1):85-107.

[18] Dubey AK, Shandilya SK. Exploiting need of data mining services in mobile computing environments. In international conference on computational intelligence and communication networks 2010 (pp. 409-14). IEEE.

[19] Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021; 7(4):432-9.

[20] Frimpong EA, Oluwasanmi A, Baagyere EY, Zhiguang Q. A feedforward artificial neural network model for classification and detection of type 2 diabetes. In journal of physics: conference series 2021 (p. 012026). IOP Publishing.

[21] Rasheed J, Hameed AA, Djeddi C, Jamil A, Al-Turjman F. A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdisciplinary Sciences: Computational Life Sciences. 2021; 13(1):103-17.

[22] Jackins V, Vimal S, Kaliappan M, Lee MY. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing. 2021; 77(5):5198-219.

[23] Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, et al. Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment. Frontiers in microbiology. 2021; 12:313.

[24] Islam MT, Al-Absi HR, Ruagh EA, Alam T. DiaNet: A deep learning based architecture to diagnose diabetes using retinal images only. IEEE Access. 2021; 9:15686-95.

[25] Rahman MR, Islam T, Nicoletti F, Petralia MC, Ciurleo R, Fisicaro F, et al. Identification of common pathogenetic processes between schizophrenia and diabetes mellitus by systems biology analysis. Genes. 2021; 12(2):237.

[26] Ishaq A, Sadiq S, Umer M, Ullah S, Mirjalili S, Rupapara V, Nappi M. Improving the prediction of heart failure patients’ survival using SMOTE and effective data mining techniques. IEEE access. 2021; 9:39707-16.

Yasir Minhaj Khan is doing M. tech. in Computer Science , PCST RGPV Bhopal (MP) and B.E. in Information Technology RGPV Technical University Bhopal (MP). His area of interest are Data mining optimization, Machine learning, Artificial intelligence.

Email: myasir7987@gmail.com

Animesh Kumar Dubey is working as Assistant professor with the department of Computer Science and Engineering, at Patel College of Science and Technology, Bhopal, India. He has completed his Bachelor of Engineering (B.E.) and MTech. degree with Computer Science Engineering from Rajeev Gandhi Technical University, Bhopal (M.P.). He has more than 15 publications in reputed, peer-reviewed national and international journals and conferences. His research areas are Data Mining, Optimization, Machine Learning, Cloud Computing and Artificial Intelligence.

Email:animeshdubey123@gmail.com