

Statistical features learning to predict the crop yield in regional areas

Pinaka Pani Ramanahalli¹, Hemanth Kollegal Siddamallu², Ravi Kumar Yelwala Basavaraju³

¹School and Computer Science and Applications, REVA University, Bangalore, India

²School of Computer Science and Applications, REVA University, Bangalore, India

³Department of Computer Science and Engineering, Christ (Deemed to be University), Bangalore, India

Article Info

Article history:

Received Mar 9, 2021

Revised May 16, 2022

Accepted Jun 9, 2022

Keywords:

Classification

Clustering

Computer science

Machine learning

Statistical features

ABSTRACT

The plethora of information presented in the form of benchmark dataset plays a significant role in analyzing and understanding the crop yield in certain regions of regional territory. The information may be presented in the form of attributes makes a prediction of crop yield in various regions of machine learning. The information considered for processing involves data cleaning initially followed by binning to reduce the missing data. The information collected is subjected to clustering of data items based on patterns of similarity, The data items that are similar in nature is fed to the system with similarity measure, which involves understanding the distance of data items from its related data item leading to hyper parameters for analyzing of information while calculating the crop yield. The information may be used to ascertain the patterns of data that exhibit similarity with nearest neighbor represented by another attribute. Thus, the research method has yielded an accuracy of 89.62% of classification for predicting the crop yield in agricultural areas of Karnataka region.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Pinaka Pani Ramanahalli

School of Computer Science and Applications, REVA University

Bangalore-560064, India

Email: rppani.mca@gmail.com

1. INTRODUCTION

The research has its intent towards identification of crop yield in regional areas of Indian territory. The research also has its focus on benchmark datasets consisting of various attributes that is required to ascertain the crop yield in agricultural areas of Karnataka regions. The information collected from Gandhi Krishi Vigyan Kendra (GKVK) Bangalore has given useful input in the form of attributes of agricultural land. The machine learning (ML) considered the actual information in terms of crop yield in the past 5 years based on attributes mentioned in Table 1. This information is helpful in understanding the crop yield in the present year based on the quality of soil present in that region and the amount of rainfall indicated in dataset. The crop yield prediction [1]–[6] can be done for next subsequent years based on the temperature and humidity of the soil also makes significant contribution to the system. Machine learning mentioned in [7]–[11] has its objective towards identifying the crop yield in various regions by employing the erected methods on benchmark dataset. The dataset includes the information of soil, temperature, humidity, pH value and various other criteria that helps the system to predict the percentage of crop yield in regional areas of Indian plateau. Figure 1 indicates the strategy used to classify the data of a dataset into different classes of information.

Machine learning techniques are usually classified into supervised and unsupervised techniques. Supervised machine learning starts from prior knowledge of the desired result in the form of labeled data

sets, which allows to guide the training process as per [12]–[16], whereas unsupervised machine learning works directly on unlabeled data. In the absence of labels to orient the learning process, these labels must be “discovered” by the learning algorithm [1]. In this technical report, we discuss the desirable features of good clustering [17]–[24] results, recall Kleinberg’s impossibility theorem for clustering, and describe a taxonomy of evaluation criteria for unsupervised machine learning. We also survey many of the evaluation metrics that have been proposed in the literature. We end our report by describing the techniques that can be used to adjust the parameters of clustering algorithms, i.e. their hyper-parameters.

Table 1. Parameters used for prediction of crop yield

Sl.No	Parameters used	Percentage of Nutrients in Soil	# Times used in research data
1	Temperature	23 c	24
2	Soil type	Black	17
3	Rainfall	25 cms	17
4	Crop information	2 times/year	13
5	pH-value	83	11
6	Humidity	29	11
7	Area of production	2	8
8	Fertilization	31	7
9	Normalized difference vegetation index (NDVI)	56	6
10	Nitrogen	19	6
11	Potassium	73	5
12	Zinc	44	3
13	Magnesium	18	3
14	Sulphur	91	2
15	Boron	86	2
16	Calcium	93	2
17	Carbon	86	2
18	Phosphorous	74	2
19	Climate	Rainy	1
20	Time	6 months	1
21	Manganese	63	1

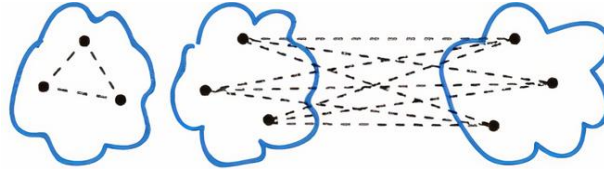


Figure 1. Architecture of proposed correlation of similarity learning for crop yield prediction

2. RELATED WORK

The research has been carried out by identifying the research gaps in the study of processing of soil data and classification of soil data into different classes of information based on evolved algorithms implementation. These evolved method has hinted erected research to come-up with a distance metric based algorithm for extraction of soil data and classification of same on benchmark dataset along with customized dataset [5], [6]. The paper evaluates the process of creating and selecting the attributes with machine learning methods for classification of data items through the research [22]–[28]. The research papers throw light in the areas of machine learning that is very helpful for processing and classifying the data items into different classes of data such as eroded soil or suitable for cultivation based on machine learning strategies. These ML techniques has driven a great advantage over specific dataset collected form Kaggle and it has pitched the direction as to how the machine learning methods has put forth for the purpose of classification [7]–[11], [13]–[17], [29]–[33]. The research papers such as [3]–[6], [12]–[14] has thrown light in to the areas of agriculture for the purpose of classification of soil data into different classes of information with machine learning and deep learning. The research presents a features that is helpful in training the system for extraction of data and classification of data into different classes of information [6], [12], [13]. The concluding remarks that shall be drawn from the research paper for the purpose of processing of features for data items into different classes of information and segregation of data into different classes of suitable data is done in [14]–[16], [29]. The research also has certain important information for processing and segregation of data into different classes of information is dine with ML. A survey on various data mining techniques has led to the outcome of proposed method for classification of soil data as per [18], [20]–[22].

3. PROPOSED ALGORITHM

The statistical information employed in this research work clusters the data items into group based on similarity measure and predicts the percentage of crop yield in various regions are described in algorithm statistical learning for crop yield prediction (SL-CYP). The algorithm has a better efficiency with other contemporary approaches in terms of pushing the data samples close to each other and pulls the samples far away from one another for samples with threshold values. The SL-CYP has yielded good performance in terms of measuring the crop yield in regional areas of India.

Algorithm: predicts the crop yield from attributes

Algorithm: statistical learning for crop yield prediction (SL-CYP)

Description: predicts the crop yield with similarity measure

Output: yields the classification results with performance

Begin

```

Step 1  [Initialization]                Solve (1)
Step 2  [Parameter Tuning]
Step 2.1                Solve (2), (3) and (4)
Step 2.2                Combine Step 2.1
Step 3  [Optimization]
Step 3.1
Step 3.2                Solve (5), (6) and (7)
Step 3.3                Combine (8)
Step 3.4                Repeat Step 2 and Step 3
                          Subtract (9) from step 2.2 and 3.2

```

End

4. METHOD

The proposed method involves various phases of prediction of crop yielding from various regions. The phases include data collection, pre-processing, processing of data with statistical information, features extraction for training the system to learn as to how the prediction or classification of data into different classes is to be made for a specific information present in data items. The crop yield prediction is achieved with the help of nutrient information and its abundance in wide range of location that is helpful to farmers in the presence of good nutrients in soil.

4.1. Collection of data

The research has focused its attention towards prediction of crop yield in various regions of Indian Plateau. The regions of Indian plateau have significant difference in their rainfall and temperature along with various other parameters like nutrients in soil. The farmer's crop yield prediction can be assessed by analyzing the moisture content present in the soil alongside the type of crop suitable for specific regions of Indian Plateau. The Figure 2 indicates the regions of Indian Plateau consisting of different states such as Karnataka and others which is suitable for particular type of soil. Based on the information available and collected from agricultural department of Karnataka, the research experimental conclusions have been drawn. The assessment of soil quality and its suitability in various regions has been predicted with the help of statistical approach incorporated in algorithm. Furthermore, Figure 2(a) indicates the nation where the soil details are considered for research purpose. Similarly, Figure 2(b) presents the region within India (Karnataka) from where the details of soil is subjected for processing, Figure 2(c) presents the nutrients of soil in regions of Karnataka used as a reference for assessing the soil nutrients and its usefulness in predicting crop yield in these regional areas of Karnataka within India.

4.2. Data pre-processing

The data preprocessing involves cleaning of data. The purpose of cleaning of data is needed to remove any redundant data from data items present in it. The repeated data or blank data is cleaned and normalized to the standard format for processing of data items prior to the estimation of crop yield in regional areas of Indian Plateau. Further, data pre-processing involves cleaning of missing data items from attributes specified in dataset. The challenging issues faced in this research involves data cleaning with missing data. These missing data is eliminated from the dataset by ignoring the tuples that corresponds to attributes. The second approach that has been employed in this research is cleaning of noisy data from a set of attributes of data items. In order to eliminate these cleaning of noisy data items, we have incorporated this proposed ideology with binning the data items. The process of binning not only removes the noise present in attributes, it also smoothens the vibrational aspects present in it.

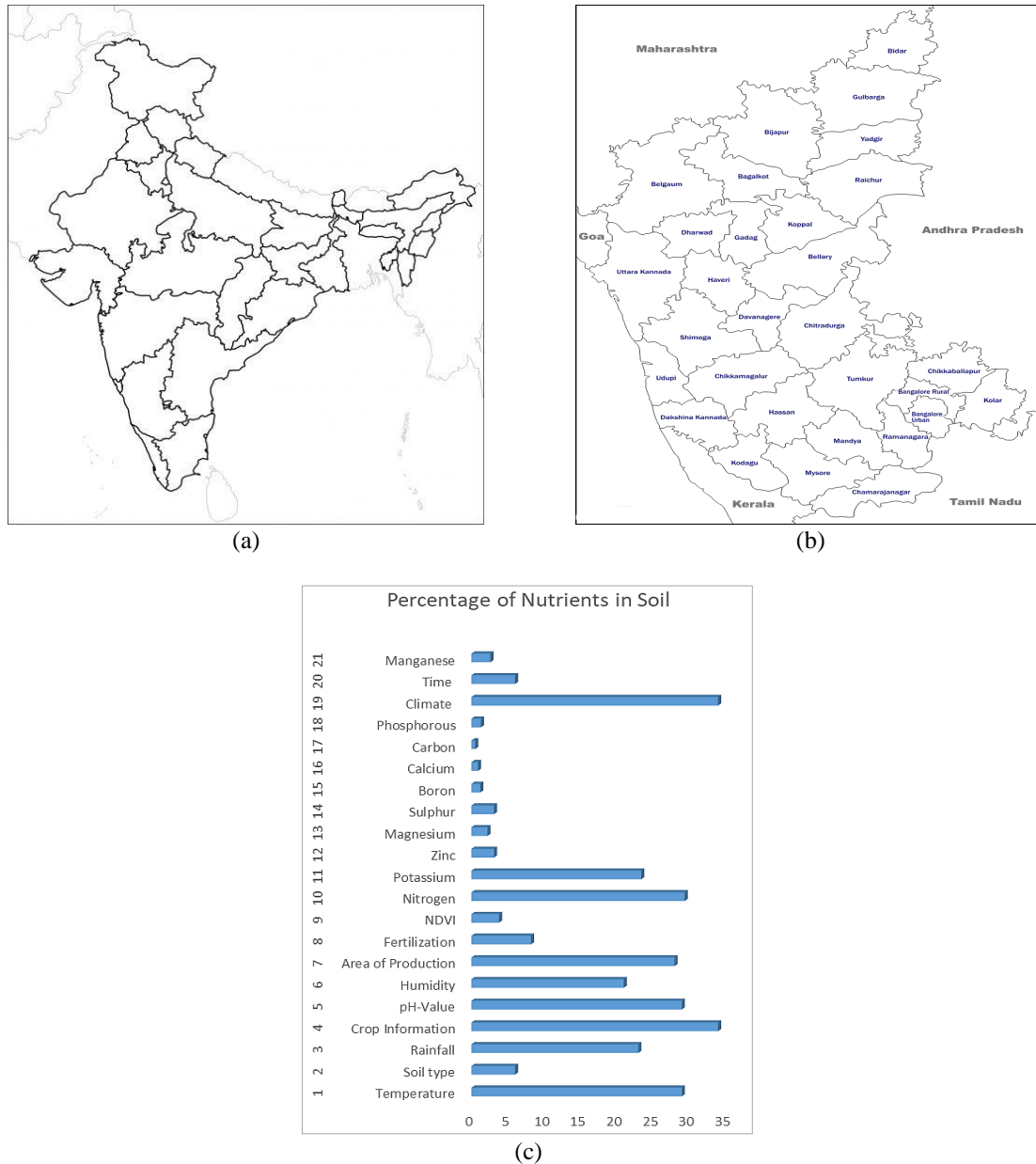


Figure 2. Maps representing the regions of India which includes Karnataka and the nutrients used for prediction of crop yield: (a) indicates the geographic map of India (b) represents the geographic map of Karnataka, and (c) indicates the parameters considered for prediction of soil nutrients

4.3. Expansion of patterns for similarity with clustering

The attributes of dataset have been explored for measuring the patterns of similarity and comparing the co-related data items with similarity metrics. The similarity metrics also helps the system to identify the attributes that exposes the hidden similarity with other attributes in support of clustering. The erected research also focuses its attention towards identifying the attributes that has certain significance in terms of distance metrics as well as patterns of data items. The similarity metrics may also be used in terms of distances such as Euclidean similarity distance metric learning and various other aspects of patterns of data items. The patterns of data items have been used in combination of distance metrics along with pattern matching. The patterns that are similar to each other up to certain threshold are pushed closer to each other and patterns that are less than the threshold values are pulled as far as possible to help improvise the clustering of data items.

The erected research has thrown light on patterns of similarity that may be established with distance metrics such as Euclidean distance. The Euclidean distance metric learning has yielded good classification

accuracy alongside with clustering. As shown in (1) has its significance in terms of measuring the similarity patterns of data items. The system pushes the attributes of data items that are more than the threshold closer to each other and pulls the attributes that are dissimilar far away from one another.

$$d(p, q) = \sqrt{(p_2 - p_1)^T + A(q_2 - q_1)} \quad (1)$$

The erected research has been formulated to optimize the performance of measuring the similarity of patterns such as (2) to (5).

$$\max S(A) = S_1(A) + S_2(A) - S_3(A) \quad (2)$$

$$S_1(A) = \frac{1}{N_k} \sum_{m=1}^n \sum_{t_1}^k (p_m - q_{t_1})^T W W^T (p_m - q_{t_1}) \quad (3)$$

$$\text{tr} \left(W^T \frac{1}{N_k} \sum_{m=1}^n \sum_{t_1}^k (p_m - q_{t_1})^T (p_m - q_{t_1}) W \right) \quad (4)$$

$$\text{tr}(W^T O_1 W) \quad (5)$$

Similarly, (3) may be written to identify the pattern similarity of data items in the form of (6). The equation (6) represented in the form of simplified manner

$$S_2(A) = \frac{1}{N_k} \sum_{m=1}^n \sum_{t_1}^k (p_{t_1} - q_m)^T W W^T (p_{t_1} - q_m) \quad (6)$$

$$\text{tr} \left(W^T \frac{1}{N_k} \sum_{m=1}^n \sum_{t_1}^k (p_{t_1} - q_m)^T (p_{t_1} - q_m) W \right) \quad (7)$$

$$\text{tr}(W^T O_2 W) \quad (8)$$

Similarly, the data items of dataset needs to be trained with other parameters, the patterns are tuned by keeping one of the variable constant and the other is tuned, likewise the other parameter is kept constant and the present variable is tuned to extract similarity of patterns that may be suitable for parameter optimization and further helps the systems to learn the features of similar patterns to be pushed closer to each other for clustering and pulled other patterns that are dissimilar far away from one another.

$$S_3(A) = \text{tr} \left(W^T \frac{1}{N_k} \sum_{m=1}^n (p_m - q_n) (p_m - q_n)^T W \right) \quad (9)$$

$$\text{tr}(W^T O_3 W) \quad (10)$$

$$\max S(W) = \text{tr}[W^T (O_1 + O_2 - O_3) W] \quad (11)$$

The optimization function helps the system to analyze and understand the features closer to each other.

$$(O_1 + O_2 - O_3)\omega = \lambda\omega \quad (12)$$

5. RESULTS AND DISCUSSION

The research process has given its contribution in the form of implementation with python. The implementation of these objectives has been achieved with certain built-in functionalities like Numpy, Matplotlib and various other functions are used for assessing the performance of system in a better manner. The result of processing with evolved approach has yielded a good classification accuracy of 89.62% with a benchmark dataset collected from GKVK Bangalore. The challenging issues that the research has faced in terms of cleaning and finding similarities of attributes of information has led to good performance measures as mentioned in Figures 3 to 6.

5.1. Metrics used

The three metrics are considered for assessing the prediction results of erected approach with existing methods. These are precision, recall and F-Measure. The precision is calculated with two important parameters such as true positive and false positive. As shown in (11) defines paves the way to identify the

performance of the proposed method while predicting the result of clustering alongside classification accuracy. Similarly, (12) recall considers two significant parameters such as true positive and false negative while f-measure as per (13) calculate uses both results of precision and recall for measuring the prediction results of the erected approach. The equations (13) to (15) determines the performance of a system.

$$Precision = \frac{TruePositive}{TruePositive+FalsePositive} \tag{13}$$

$$Recall = \frac{TruePositive}{TruePositive+FalseNegative} \tag{14}$$

$$fmeasure = \frac{(2*precision*recall)}{precision+recall} \tag{15}$$

5.2. Comparison of performance

The performance of erected method over actual data items vs. predicted results has shown its significance in terms of precision and recall. The research result obtained in graphical representation indicates the value of precision mapped into recall for analysis of proposed method. The efficacy of the proposed method indicates the result is better than other contemporary methods as it is mentioned in related work from [17]–[19], [29], [30].

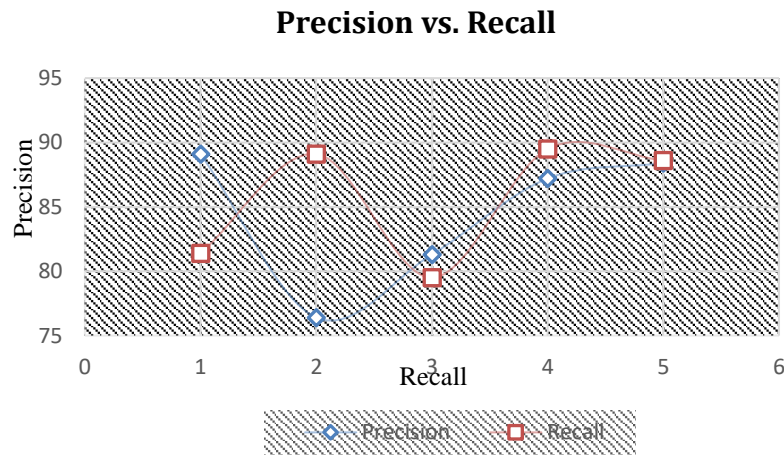


Figure 3. Prediction result of proposed method with respect to ground truth data is shown for five attributes like temperature, soil type, rainfall, crop information, and pH value

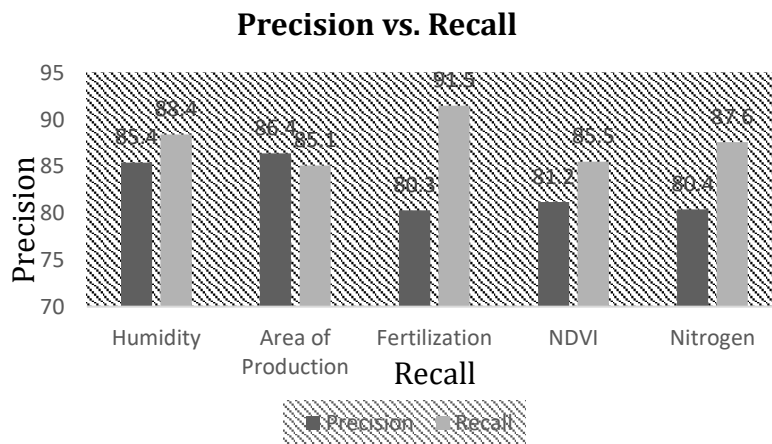


Figure 4. Prediction result of proposed method with respect to ground truth data is shown for five attributes like humidity, area of production, fertilization, NDVI, and nitrogen

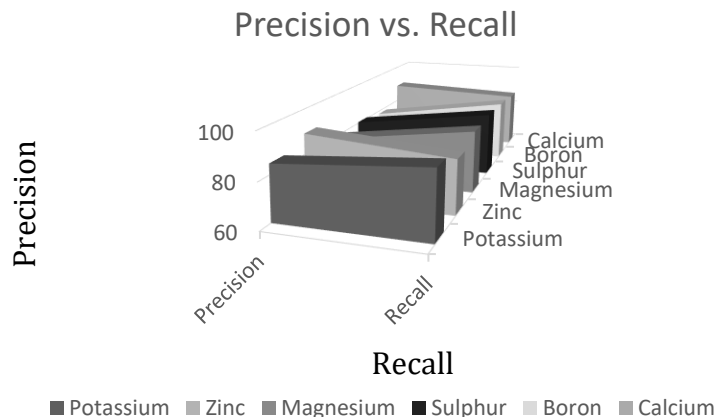


Figure 5. Prediction result of proposed method with respect to ground truth data is shown for five attributes like potassium, zinc, magnesium, Sulphur, boron, and calcium

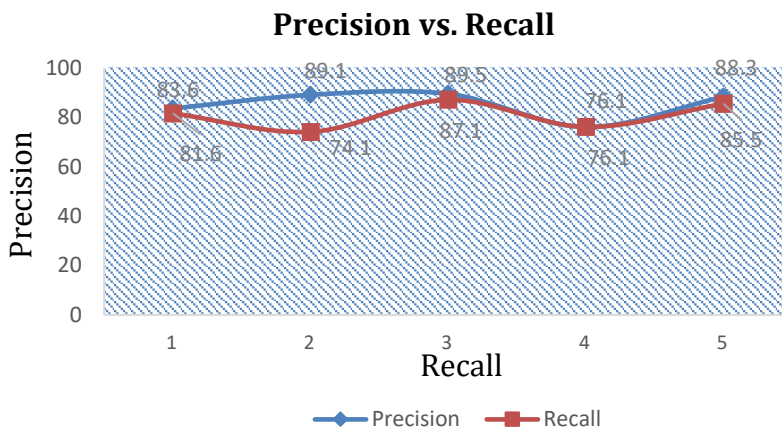


Figure 6. Prediction result of proposed method with respect to ground truth data is shown for five attributes like carbon, phosphorous, climate, time, and manganese

6. CONCLUSION

The evolved research has produced good classification accuracy alongside clustering while measuring the performance with existing methods. Since, the research has yielded good classification accuracy of 89.62% over a dataset which is collected from agricultural department GKVK. The dataset is considered as a benchmark, as it considers 21 attributes as different conditions of agriculture where we find it useful for ascertaining the crop yield. The proposed statistical features learning for crop yield prediction has been considered as a state-of-the-art technique, as it provides good efficacy over a dataset and yields good results of classification in comparison with other methods.

ACKNOWLEDGEMENTS

The authors are delighted to acknowledge Chancellor of REVA University, Dr. P. Shyama Raju and Director, School of CSA, REVA University, Dr. S. Sendhil for their continuous support and guidance in developing this research work.





REFERENCES

[1] D. Greene, P. Cunningham, and R. Mayer, "Unsupervised learning and clustering," in *Machine Learning Techniques for Multimedia*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 51–90.
 [2] V. Estivill-Castro, "Why so many clustering algorithms," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 1, pp. 65–75, Jun. 2002, doi: 10.1145/568574.568575.
 [3] J. Kleinberg, "An impossibility theorem for clustering," in *NIPS'02: Proceedings of the 15th International Conference on Neural Statistical features learning to predict the crop yield in regional areas (Pinaka Pani Ramanahalli)*




- Information Processing Systems*, 2002, pp. 463–470.
- [4] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education, Inc., 2005.
 - [5] G. Gan, C. Ma, and J. Wu, *Data clustering: Theory, algorithms, and applications*. Society for Industrial and Applied Mathematics, 2007.
 - [6] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of Intelligent Information Systems*, vol. 17, pp. 107–145, 2001, doi: 10.1023/A:1012801612483.
 - [7] M. G. P. S. and B. R., “Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms,” *Applied Artificial Intelligence*, vol. 33, no. 7, pp. 621–642, Jun. 2019, doi: 10.1080/08839514.2019.1592343.
 - [8] M. Lango and J. Stefanowski, “Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data,” *Journal of Intelligent Information Systems*, vol. 50, no. 1, pp. 97–127, Feb. 2018, doi: 10.1007/s10844-017-0446-7.
 - [9] K. R. Y. Basavaraju, C. K. Narayanappa, and P. Dayananda, “Weighted full binary tree-sliced binary pattern: An RGB-D image descriptor,” *Heliyon*, vol. 6, no. 5, May 2020, doi: 10.1016/j.heliyon.2020.e03751.
 - [10] R. K. Y. Basavaraju and C. N. R. Kumar, “Harmonic rule for measuring the facial similarities among relatives,” *Transactions on Machine Learning and Artificial Intelligence*, vol. 4, no. 6, Dec. 2018, doi: 10.14738/tmlai.46.2221.
 - [11] R. K. Y. Basavaraju and C. K. Narayanappa, “Triangular similarities of facial features to determine: The relationships among family members,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 22, no. 3, pp. 323–332, May 2018, doi: 10.20965/jaciii.2018.p0323.
 - [12] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. United States: 1988, 1988.
 - [13] C. C. Aggarwal, *Data mining*. Cham: Springer International Publishing, 2015.
 - [14] J. Handl, J. Knowles, and D. B. Kell, “Computational cluster validation in post-genomic data analysis,” *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, Aug. 2005, doi: 10.1093/bioinformatics/bti517.
 - [15] Q. Zhao, “Cluster validity in clustering methods. Dissertations in forestry and natural sciences,” University of Eastern Finland, 2012.
 - [16] T. Calinski and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.
 - [17] G. H. Ball and D. J. Hall, “ISODATA, a novel method of data analysis and pattern classification,” Technical report AD0699616, 1965.
 - [18] J. A. Hartigan, *Clustering algorithms (Wiley series in probability and mathematical statistics)*, 1st Edition. Wiley, 1975.
 - [19] L. Xu, “Bayesian ying-yang machine, clustering and number of clusters,” *Pattern Recognition Letters*, vol. 18, no. 11–13, pp. 1167–1178, Nov. 1997, doi: 10.1016/S0167-8655(97)00121-9.
 - [20] M. E. Celebi and K. Aydin, *Unsupervised learning algorithms*. Cham: Springer International Publishing, 2016.
 - [21] R. R. Sokal and F. J. Rohlf, “The comparison of dendrograms by objective methods,” *TAXON*, vol. 11, no. 2, pp. 33–40, Feb. 1962, doi: 10.2307/1217208.
 - [22] S. Theodoridis and K. Koutroumbas, *Pattern recognition*, Pattern Re. 2003.
 - [23] M. Cord and P. Cunningham, *Machine learning techniques for multimedia*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.
 - [24] C. C. Aggarwal and C. K. Reddy, *Data clustering: Algorithms and applications*. Chapman and Hall/CRC, 2016.
 - [25] J. W. Perry, A. Kent, and M. M. Berry, “Machine literature searching X. Machine language; factors underlying its design and development,” *American Documentation*, vol. 6, no. 4, pp. 242–254, Oct. 1955, doi: 10.1002/asi.5090060411.
 - [26] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
 - [27] J.-Y. Hsieh, W. Huang, H.-T. Yang, C.-C. Lin, Y.-C. Fan, and H. Chen, “Building the rice blast disease prediction model based on machine learning and neural networks,” *Easy chair the world of scientists*, pp. 1–8, 2019.
 - [28] A. Bahl *et al.*, “Recursive feature elimination in random forest classification supports nanomaterial grouping,” *NanoImpact*, vol. 15, Mar. 2019, doi: 10.1016/j.impact.2019.100179.
 - [29] J. C. Dunn†, “Well-separated clusters and optimal fuzzy partitions,” *Journal of Cybernetics*, vol. 4, no. 1, pp. 95–104, Jan. 1974, doi: 10.1080/01969727408546059.
 - [30] X. L. Xie and G. Beni, “A validity measure for fuzzy clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, 1991, doi: 10.1109/34.85677.
 - [31] R. K. Y. Basavaraju and C. N. R. Kumar, “Eye center localization using cascaded corner detection and geometrical measurements algorithm,” in *2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15)*, Dec. 2015, pp. 1–5, doi: 10.1109/ITACT.2015.7492678.
 - [32] R. K. Y. Basavaraju and C. N. R. Kumar, “Local binary pattern: An improved LBP to extract nonuniform LBP patterns with Gabor filter to increase the rate of face similarity,” in *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)*, Aug. 2016, pp. 1–5, doi: 10.1109/CCIP.2016.7802878.
 - [33] R. K. Y. Basavaraju, “Assessment of facial homogeneity with regard to genealogical aspects based on deep learning approach,” *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 3, pp. 1550–1556, Apr. 2021, doi: 10.17762/turcomat.v12i3.962.

BIOGRAPHIES OF AUTHORS




Pinaka Pani Ramanahalli     with MCA degree from Sri Venkateswara University, Tirupati. He has over 11 years of teaching experience and two years of experience in IT Industry where he involved in development of software applications using STRUTS framework. At present he is working as an Assistant Professor at REVA University, Bengaluru since 2010. Currently he is doing research to improve crop yield using Machine Learning algorithms. He can be contacted at email: rppani@gmail.com.



Hemanth Kollegal Siddamallu    received Ph.D. Degree in the year 2014 from Mangalore University, B.Sc Degree and MCA Degree in the years 2006 and 2009, respectively from University of Mysore. Currently working as Associate Professor, faculty of Computer Science and Applications, REVA University, Bangalore, India. His area of interest includes data mining, artificial intelligence and deep learning. He can be contacted at email: hemanth.ks@reva.edu.in.



Ravi Kumar Yelwala Basavaraju    received B.E. from VTU, Belagavi and M.Tech. Degree in Computer Science and Engineering from VTU, Belagavi, Karnataka, India, in 2007 and 2011, respectively. The Ph.D. degree in Computer Science and Engineering from Visveswaraya Technological University, Belagavi, and Karnataka, India in 2021. He is currently working for Computer Science and Engineering at Christ University, Bangalore, India from November 2021. His research interests include the applications of artificial intelligence, machine learning and robotics. He can be contacted at email: ravikumarybdhanya@gmail.com.