



ADVANCED ANALYTICAL FRAMEWORK FOR CROP YIELD PREDICTION LEVERAGING DIVERSE FEATURE SELECTION METHODS AND MACHINE LEARNING CLASSIFIERS IN VARIED AGRICULTURAL ENVIRONMENTS

¹Dr. M. Lavanya, ²K Vinitha, ³K Vamsi Krishna, ⁴K T Siva Kumar, ⁵P Maruthi, ⁶G Uday Kiran

¹Associate Professor, ²20781A3322, ³20781A3324, ⁴20781A3325, ⁵20781A3337, ⁶20781A3315

¹CSE(AI&ML)

¹Sri Venkateshwara College of Engineering & Technology (Autonomous), Chittoor, India

Abstract : In this paper, we propose a prediction method for major crops in India using a combination of K-means clustering and Modified K Nearest Neighbor (KNN) classification algorithms. Given that agriculture is the primary livelihood for over 40 percent of the state's population and the global population is projected to increase by one third between 2010 and 2050, with a corresponding 60 percent rise in demand for crop production, accurate prediction becomes crucial for maximizing yield.

We utilized MATLAB for clustering using the K-means algorithm and WEKA for classification using Modified KNN. The results demonstrate that our method outperforms traditional data mining approaches. This approach allows us to anticipate crop demand more effectively, thereby enabling farmers and stakeholders to optimize production and meet the growing demands of the population."

Index Terms: Random forest regression, K-means, Machine learning

I. INTRODUCTION

This paper explores the realm of smart agriculture, aimed at bridging the knowledge gap between traditional and educated farmers by leveraging various data-driven techniques. It focuses on estimating aggregate physical production functions for crop yields in specific states, incorporating technological factors and a newly developed weather index as inputs. Regression analysis, coefficient of determination analysis, and Average Error rate calculations were conducted to compare actual results (target) with predictions from our network outputs.

The primary objective is to develop a user-friendly interface for farmers, providing analysis of rice production based on available data. Different data mining techniques were employed to predict crop yields and maximize productivity. Accurate and timely monitoring of crop conditions and estimating potential yields are crucial for operational programs and decision-making processes in agriculture, especially in countries like Ghana.

The study employs linear regression methods to forecast crop yields across different seasons, recognizing the importance of data mining in various economic sectors. Initially utilized by large companies to analyze consumer data for profitability, data mining methods have expanded into sectors like agriculture and biofuel industries, aiding decision-making processes.

Corn production information is well-established, but various factors like planting date, fertilization, tillage, crop rotation, and weed control practices can influence yield and profitability. The paper also addresses the increasing global energy demand,

focusing on designing and implementing a system to control motor performance using Short Message Service (SMS) via cell phones. This system allows remote control of motor functions and provides status updates via mobile phones, enhancing operational efficiency.

The integration of IoT systems in smart farms enables connectivity among diverse agricultural devices, facilitating intelligent agricultural services based on shared expert knowledge. Additionally, understanding crop price trends is crucial for agri-business profitability, with research focusing on climate-harvest relationships and price forecasting.

Spatial resolution considerations are vital for capturing environmental variability, particularly in mountainous regions, although high-resolution data are limited in availability globally.

II. LITERATURE SURVEY

A. System Architecture

In Figure 1, the raw data consists of the production figures for major crops, rainfall data, groundwater levels, and cultivation areas in India. Considering the high production levels of rice, maize, ragi, sugarcane, and paddy in India, these crops are identified as the major crops of the state. However, since the raw data is incomplete, preprocessing steps are necessary to make it usable.

Data preprocessing involves converting raw data into a meaningful and understandable format. Typically, data preprocessing consists of three main steps: data cleaning, data transformation, and data reduction.

2.1. Data Cleaning: This step involves removing incomplete or erroneous data to ensure the dataset's integrity. Incomplete data can skew analysis and predictions, so it's essential to clean the dataset before further processing.

2.2. Data Transformation: Transformation involves mapping the data into a uniform format. For instance, data may be available at different time granularities such as hourly, monthly, or yearly. Transforming the data into a consistent format, such as year-wise data, enhances its usability and consistency.

2.3. Data Reduction: Data reduction simplifies the dataset by transforming it from an unorganized form into a more manageable and structured format. One common technique for data reduction is clustering, particularly using the K-means algorithm. In K-means clustering, the number of clusters (K) is chosen beforehand. In this case, five clusters are used, representing different levels of production (very low, low, medium, high, very high). After applying the K-means algorithm, the dataset is clustered based on minimum distances, resulting in labeled output. This labeled dataset can then be used for supervised prediction tasks.

Once the dataset is in a supervised format, algorithms such as Fuzzy, KNN, and Modified KNN can be applied for prediction tasks. These algorithms leverage the labeled dataset to make predictions about crop yields, based on factors such as production, rainfall, groundwater levels, and cultivation area.

III. METHODOLOGY

K-Means clustering

The agricultural data is clustered using the K-Means algorithm, which is an unsupervised clustering algorithm. The data is classified into clusters, with 'k' representing the number of clusters. Initially, centroids are assumed to be the first two values in the dataset. Then, the distance between each data point and the cluster center (centroid) is calculated using the Euclidean formula.

$$\text{Distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Assign each data point to the cluster center that has the minimum distance from the data point among all the calculated centroids. Then, recalculate the new cluster centers until there is no change in the clusters from the previous iteration. The agricultural input data, such as rainfall, groundwater, cultivated area, and output crop production, are clustered into categories of very low, low, moderate, high, and very high obtained from the output of the K-means algorithm.

IV. IMPLEMENTATION

In this module, a dataset of groundwater levels spanning the past ten years in India is utilized. The data is converted into a dataframe and preprocessed to eliminate records with zero values in all columns. From the columns 'MONSOON', 'POMKH', 'POMRB', and 'PREMON', the average value is calculated. The middle value is determined, and any values below this middle value are considered as zero, while those above are considered as one. These two values are then applied to create a new column called 'class factor'.

These classifications are shown in Table 3 for input data and Table 4 for output crop production.

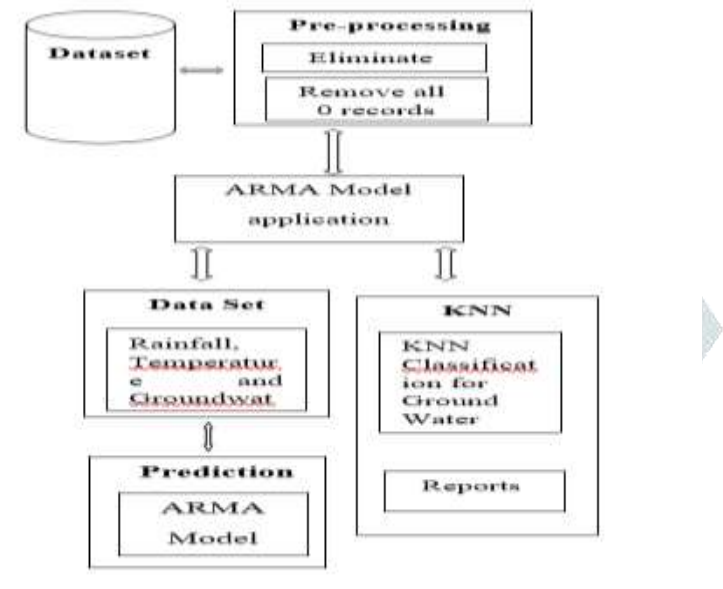


Figure1: **WORK FLOW GROUND WATER LEVEL CLASSIFICATION BASED ON KNN MODEL**

Next, KNN classification is employed to predict groundwater levels using the 'class factor' column. For a test run, a K value of 6 is chosen, and the model is predicted. Subsequently, the accuracy of the KNN model is calculated and displayed.

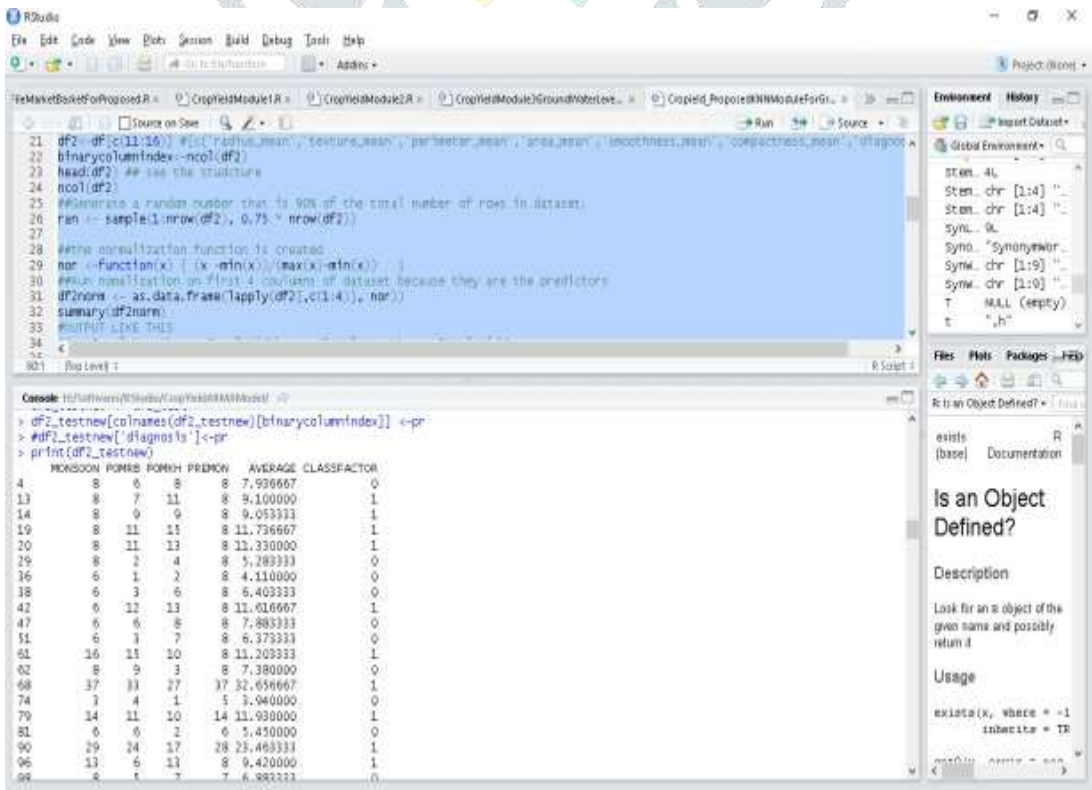


Figure2: KNN CLASSIFICATION FOR GROUND WATERLEVEL DATA

```

21 df2 <- df[c(1:15), #["Paddy_mean", "texture_mean", "perimeter_mean", "area_mean", "smoothness_mean", "compactness_mean", "diagonal",
22 binarycolumnIndex=ncol(df2)]
23 head(df2) ## see the structure
24 ncol(df2)
25 #Generate a random number that is 90% of the total number of rows in dataset.
26 ran <- sample(1:nrow(df2), 0.75 * nrow(df2))
27
28 ##the normalization function is created
29 nor <-function(x) { (x -min(x))/(max(x)-min(x)) }
30 ##Run normalization on first 4 columns of dataset because they are the predictors
31 df2norm <- as.data.frame(lapply(df2[,c(1:4)], nor))
32 summary(df2norm)
33 #OUTPUT LIKE THIS
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

V. ACCURACY OF THE KNN MODEL

5.1 K Nearest Neighbor(KNN)

The nearest neighbor algorithm assigns to a test pattern the class label of its closest neighbor. Let there be n training patterns, $(X_1, \theta_1), (X_2, \theta_2), \dots, (X_n, \theta_n)$, where X_i is of dimension d and θ_i is the class label of the i th pattern. If P is the test pattern, then if $d(P, X_k) = \min$

$\{d(P, X_i)\}$ where $i = 1 \dots n$. Pattern P is assigned to the class

θ_k associated with X_k

Steps involved:

1. Determine the parameter k , which represents the number of nearest neighbors.
2. Calculate the distance between the query instance and all the training samples.
3. Sort the distances and determine the nearest neighbors based on the K -th minimum distance.

In WEKA version 3-6, the training dataset is first selected (Figure 2), and then the corresponding classifier is chosen, with the output being displayed (Figures 3 and 4).

5.2 Modified K Nearest Neighbor

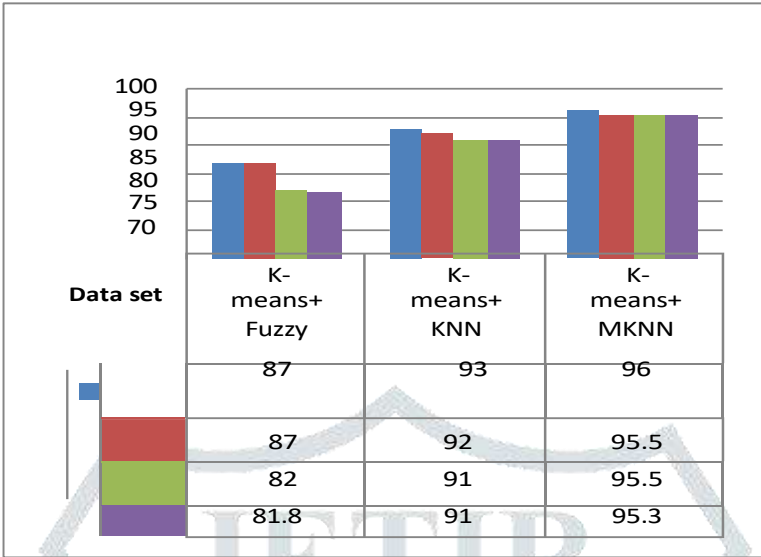
A classification method is proposed to enhance the performance of K-Nearest Neighbor, called Modified K-Nearest Neighbor (MKNN). This method utilizes robust neighbors in the training data. Inspired by this approach, MKNN computes the fraction of neighbors with the same label to the total number of neighbors.

The main idea behind this method is to assign the class label of the data based on K validated data points from the training set. First, the validity of all data samples in the training set is computed. Then, a weighted KNN is performed on any test samples.

VI. RESULT & ANALYSIS

In MATLAB, K-Means clustering was initially performed using parameters such as rainfall, area, and groundwater for analysis. Subsequently, the results from MATLAB were associated with WEKA to conduct further analysis using K-Means with Fuzzy, K-Means with KNN, and K-Means with Modified KNN algorithms. During experimentation with these three algorithms, the precision, measured by correctly classified instances, varied.

After analyzing the output from each algorithm, it was determined that K-Means with Modified KNN yielded the best results among the three experimented algorithms. The analysis results are presented graphically to illustrate the superiority of K-Means with Modified KNN over the other algorithms.



The proposed work incorporates fuzzy logic to estimate crop yield, which operates on a set range rather than discrete values. Therefore, errors in predicted rainfall data do not pose significant issues as long as the difference between actual and estimated values is not drastic. The model demonstrates its capability to successfully predict crop yield for a given year when the rainfall and temperature values for previous years are known.

Similarly, the model successfully predicts groundwater levels for a given year when the values from previous years are available. Moreover, the project employs KNN classification to classify groundwater level dataset records, enabling prediction models for future test datasets. This approach facilitates the analysis of past groundwater levels to predict future levels.

In the future, logistic regression can be applied to further classify the data, enhancing the accuracy and reliability of predictions. This integrated approach leverages various techniques to improve the understanding and prediction of agricultural variables, ultimately aiding decision-making processes in agriculture.

VII. CONCLUSION

In conclusion, this study has focused on predicting major crop yields in India using K-Means clustering and Modified KNN classification algorithms. The analysis encompassed three types of algorithms: fuzzy logic, KNN, and Modified KNN. Through rigorous experimentation and evaluation, it was determined that Modified KNN outperformed the other algorithms in terms of accuracy and precision.

Looking ahead, future research endeavors will delve into exploring various bio-inspired methods for crop yield prediction. This entails conducting a comparative study to assess the accuracy and effectiveness of each algorithm in predicting crop yields. By leveraging bio-inspired methods, such as genetic algorithms, swarm intelligence, and neural networks, we aim to enhance the predictive capabilities and robustness of crop yield prediction models.

Ultimately, these advancements hold promise for revolutionizing precision agriculture practices, empowering farmers with more accurate predictions and insights into crop yield variability. By continuously refining and advancing predictive models, we can contribute to the sustainability and efficiency of agricultural production, ensuring food security and prosperity for future generations.

REFERENCE

[1] D Ramesh, B Vishnu Vardhan. "Crop yield prediction using weight-based clustering technique." International Journal of Computer Engineering and Applications, Volume IX, Issue IV, April 15.

[2] Spyridon Mourtzinis, Francisco J. Arriaga, Kipling S. Balkcom, and Brenda V. Ortiz. "Corn yield prediction model uses simple measurements at a specific growth stage." Published on 2 July 2013.

[3] J. Arriaga, Kipling S. Balkcom, and Brenda V. Ortiz. "Corn Grain and Stover Yield Prediction at R1 Growth Stage." Spyridon Mourtzinis, Francisco. Published on May 3, 2013.

[4] Richard Kidd Perrin. "Analysis and prediction of crop yields for agricultural policy purposes." Iowa State University, 1968.

- [5] Simone M.S. Costa and Caio A. S. Coelho. "Crop yield predictions using seasonal climate forecasts." Centro de Previsão de Tempo e Estudos Climáticos, Instituto Nacional de Pesquisas Espaciais, Cachoeira Paulista – SP, Brasil.
- [6] Askar Choudhury, James Jones. "Crop yield prediction using time series models." Illinois State University.
- [7] Mrs. K.R. Sri Preetha M.E., S. Nishanthini, D. Santhiya, K. Vani Shree. "Crop yield prediction." International Journal On Engineering Technology and Sciences – IJETS, March 2016.
- [8] Mohammadhossein Hajiyan. "Early Prediction of Crop Yield." School of Engineering, University of Guelph, May 1, 2012.
- [9] M. Lavanya, Smart Production and Manufacturing System Using Digital Twin Technology and Machine Learning Springer Nature Journal, 42979-023-01976.

