JOURNAL OF CRITICAL REVIEWS

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

SYSTEM ARCHITECTURE CONCEPT BASED CROP RECOMMENDATION SYSTEM USING GIS TECHNIQUE

C. Karthik[1]*, K. Vinothkumar[2], S. Vigneshwaran[3]

[1]*Assistant professor, saraswathy college of Engineering and technology, Tindivanam
[2]Assistant professor, VRS college of Engineering and technology, viluppuram
[3]Assistant professor, sri vidya college of Engineering & Technology, viruthunagar

***Corresponding Author:** C. Karthik
*Assistant professor, saraswathy college of Engineering and technology, Tindivanam

Abstract—Recommendation Systems focus on providing the consumers with the closest matches of items that suites the tastes of the consumer. The tastes are identified, based either on the history of purchases made by the consumer (content based) or by accounting for the association born out as a result of their comparison with their closest companions. These companions are the ones who shared similar interests to that of the consumer (collaborative filtering). In this paper, we propose Recommendation System for Crops (CRS),that recommends the appropriate crop(s) for plantation in a particular region, taking into consideration the nature of soil, the climatic conditions, the amount of rainfall, the temperature levels and the water stress impact measures of the region under examination. The system has shown 77% prediction accuracy, proved by validation tests conducted against predictions of classifier models, which are built out of data, gathered by processing image datasets (maps) of Indian territory.

Keywords—Recommendation Systems; Content Based; Collaborative Filtering

INTRODUCTION

Image Processing and Data Mining are the key concepts used here in determining the most suitable crop(s) for plantation in a specific region. The textual data relating to agriculture aren't available at a point in time but rather as average values estimated over a timespan. This makes textual data less consistent in comparison to live GIS image data,which is available through constant monitoring via satellites. The data available as live images are a reliable source of information.The maps of India highlighting the various features, relating to crop growth, have been used as the base data set for a point in time. This data serves as a novel example for analyzing the effectiveness of the system which can be extended for the GIS datasets. Image segmentation using color as the distinctive parameter is performed by clustering technique. K-Means Clustering has been employed as an Image Processing technique to build the usable dataset by overlaying the results obtained from different maps. Classification technique of Data Mining is performed using the feature information of each location as data and the crops favorable for cultivation in that region as the target class. Classifier models based on Ensemble Learning, Support Vector Machines (SVM) with RBF as kernel function, Random Forest, K-Nearest Neighbors and Naïve Bayes were then used for prediction. The Classifier Models used have been briefly discussed in the next section.

CLASSIFICATION ALGORITHMS USED

Classification is the process of assigning class labels to the new data elements, based on the training data elements and the categories to which those training data elements belong to. For e.g., labelling a mail as Spam or not Spam, classifying pictures of animals as Dogs and Cats, etc., based on available feature information.

1. Naïve Bayes Classifier

It uses a probabilistic approach based on Bayes Theorem of Probability to assign class labels to the data elements. It holds a strong assumption that the features are completely independent of each other. The feature vectors in the dataset built have some partial correlation. For e.g., it can be seen clearly that the factors like Temperature and Rainfall are to an extent dependent on each other. More is the rainfall, lesser is the Temperature and vice versa.

2. SVM

Support Vector Machine is a classifier model which tries to construct hyperplanes, with maximal margin for a clear distinction, that segregates the entire space for assigning to different class labels. It makes use of kernel trick, the transformation of lower dimensional data onto higher dimensional space for easier segregation, generally for non-linear separation problem. It doesn't perform well on large datasets, considering the long time required for training.

3. K-Nearest Neighbors

K Nearest Neighbors Classifier, based on the parameter *K*, assigns a class label to the data points, by optingthe class label, which is the majority of all class labels obtained by *K* nearest data points in the training data. It is easy to visualize and is also effective due to the shorter training time needed.

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

4. Bagging

Bagging classifier is a type of ensemble learning method, which makes use of the Bootstrap technique to create random sub samples of dataset (with replacement), train each of them using some other classifier like Decision Trees or K Nearest Neighbors and then use the average or majority predictions for the new data value.

5. Random Forest

Random Forest is similar to bagging model in most sense except for the number of features that are considered for classification. Only a subset of are considered to determine the best split. Several weak classifiers therefore combine here to form the strong classifier.

RELATED WORK

This section gives an overview of the related research work in the context of relevant technologies.

S. Latu [1] showcases the adverse impact of economic development activities on the coastal ecosystems in exemplar developing countries, in the Pacific area, and proposes GIS Visualization strategies for moving beyond subsistence and economic development aspirations to socially, economically and environmentally sustainable development activities.

N. Li et al. [2] talks about how ontology's and semantic technologies offer support to the documentation and retrieval of dynamic information in GI Science by providing flexible schemata instead of fixed data structures which bring down the level of the results.

Neha et.al. [3] developed an ontology for cotton crop in India which can be extended for further making a more robust knowledge base system.

R Jeberson et al. [4] refers to as how GIS web services may be implemented to tackle the natural calamities such as tsunami, flood, earthquake etc.

Y. Jain et al. [5] gives an overview of GIS based agricultural system, which can provide support to the farmers during various phases of farming. A knowledge base is made use of to provide support to the farmer for better reasoning.

V. Kumar et al. [6] proposes a semantic web based architecture to generate agricultural recommendations, using spatial data and agriculture knowledge bases. Knowledge base sends recommendations to the farmers based on climatic conditions and geographical data.

J. Konaté et al. [7] developed a framework for providing recommendation of crops and the recommendation of farming practices based yield, crop life cycle, soil nature, growing season, etc.

Kiran Shinde provides a Fertilizer Recommendation System consists of logic computes all the possible combination of fertilizers to meet the crop requirements and the combination with lowest cost of fertilization will be recommended. It proposes the use of data mining techniques to provide recommendations to farmers for crops, crop rotation and identification of appropriate fertilizer. The results from the recommendation system are optimized with respect to parameter consideration.

With enormous amount of data now available through the Web, opportunities exist to integrate these data to support complex applications. On the other hand, our crop recommendation system is more real time and dynamic as it uses real time images to be processed. The regions in the images are initially segregated and divided into areas of interest which are then mapped to the crop pertaining to that particular region based on past history which plays a major role as future recommendations can be aided from the history of rainfall, soil pattern, fertility , diseases that infected the particular crop, alternate crops that can be cultivated during the interim period and also provides ease of usability to the user by offering language flexibility based on the local region languages as input to be given to the system. Various queries are resolved by the system and result of a query is the recommendation of suitable and possible crops to be grown in that particular region.

SYSTEM IMPLEMENTATION

The simulation of the application covers a wide range of modules done primarily in MatLab, Python, R, Orange and Adobe Photoshop. Photoshop was used to stretch out the images manually to uniform orientation and dimensions. MatLab has been used for the image preprocessing works that was a major part of the project. The audio processing from speech to text and text to speech has used Python and the built-in Google Speech Cloud API. R and Orange software have been used to take advantage of the built-in classification and validation packages. The implementation procedure has been discussed in detail below.

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

1. Indexing the Image Datasets

1.1 Preprocessing of Images

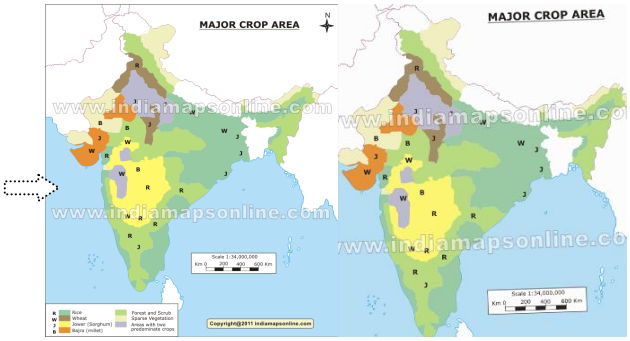


Fig. 1 Map Preprocessing

The dataset of map images played the key role in the application. The India map images that were taken into account were those of soil distribution, climatic conditions, temperature values, rainfall quotient, water stress levels and cropping patterns.

The application demanded them to have perfectly same orientation and dimensions to align the pixels in a way such that each pixel in each map corresponds to the same specific location on all the other contributing map images. The images were carved out by the translation and rotation techniques available in Photoshop software and fitted to the dimensions, 1200x1400 (width x height). It was ensured that only the region into consideration (here, the Indian Territory) came into picture.

1.2 Segmenting Maps based on Color by K-Means Clustering

The preprocessed maps were then subjected to a detailed procedure for extracting the individual portions based on color. The different colors correspond to the possible set of values for the given feature. For instance, for the feature Crop, the crop names likes Rice, Wheat and Bajra may be mapped to colors Red, Green and Blue portions of the map. The individual segments are obtained by K-means clustering and stored in the disk.

1.2.1 Conversion from RGB to Lab Color Space

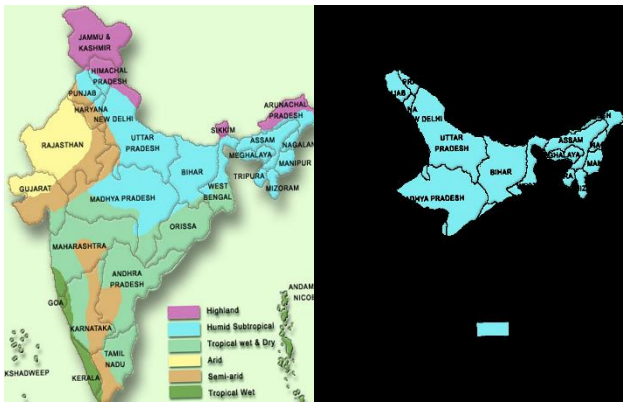
Lab color space is a 3-axis color system with dimension L for luminance and a and b for the color dimensions. The color differences are projected in a better manner using the Lab Color Space in comparison to the RGB model. The conversion to Lab color spaces makes it convenient for handling the chromic part of the image in an efficient manner without any impact on the luminosity.

1.2.2 Subsequent Processing Step

The chrominance part of the image embedded in the second and third dimensions of the original matrix is reshaped into a matrix with number of rows being equal to the total number of pixels(no. of rows x no. of columns) in the image and the number of columns being equal to 2 (for the two color component values).

1.2.3 K-Means Clustering

The system performs K-Means clustering on the processed dimensions based on the value of k provided by the user and returns the cluster identifiers and their corresponding centers. Each of the individual segmented portions are written to the disk for further processing.



ISSN-2394-5125 VOL 06, ISSUE 07, 2019

Fig. 2 Clustered Climate Map of India

1.2.4 Elbow method: Optimal k value

The within group sum of squares has been plotted against the number of clusters. The within group sum of squares is defined as the differences within a group due to presence of foreign members (originally member of another group). The objective is to find the elbow point i.e. the point where the change in within group sum of squares has become stagnant and thereby, increasing the number of clusters doesn't yield any better result.

The value for number of clusters fixed manually was 20 in this case which endorses the results obtained from the elbow method.

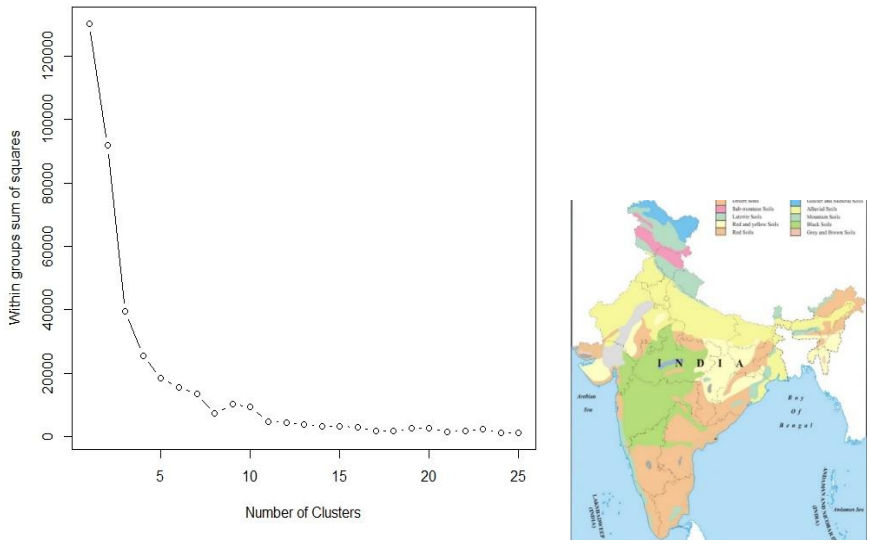


Fig. 3 Elbow Thresholding

1.3 Mapping Clusters to Feature Values

Manually, the list of the cluster identifiers and their corresponding feature values are generated as key value pairs for each of the maps. The program proceeds with the matrix read (which contains the cluster identifiers as index in place of the actual pixel values) as input and produces as result the data set consisting of the values for each of the individual feature values for each pixel. Those pixels which generated NULL values (unmapped values), even for one feature are discarded during this step. The sample dataset has been shown below.

Table 1 sample dataset

Soil	Climate	Temp	Rain	Water	Crop
GlacierSkeletal	Arid	26	20	ExtremleyHigh(>80)	SparseVegetation
GlacierSkeletal	Semi-arid	26	20	Low<10	SparseVegetation
GlacierSkeletal	Arid	26	20	AridLowWaterUse	SparseVegetation
GreyBrown	Arid	26	20	Low<10	SparseVegetation

1.3 Generated Dataset

The generated dataset consisted of 3,83,481 rows of data accounting for 18.9 MBs.

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

2. State-Wise Feature Distribution

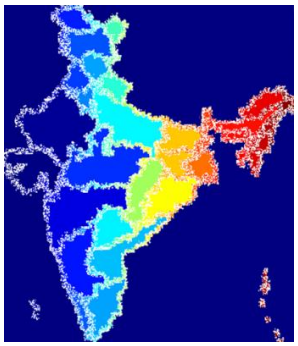


Fig. 4 India Map after Water Shed Transformation

The application also aimed at producing the distribution ratio of various feature percentages in each of the states. For this purpose, the states had to be identified. This decision called for the use of edge detection and watershed transformation followed by segmentation. Then superimposing the state outlines on to the different feature maps produced the distribution in each state. Then, using the superimposed maps, the individual feature proportions were estimated and the percentages for each feature value were computed.

2.1 Watershed Transformation

The watershed transformation is applied here to separate the states, that is obtaining a clear separation based on colors. This involved converting the image form RGB to grayscale. The edge detection algorithm like Sobel was applied to detect the edges and bring about a separation between the states. Then using the watershed transformation, the different states were obtained as contrastingly colored regions.

2.2 Superimposition of features onto State outline

The clustering approach is yet again used to identify the individual states of the map. The states were now available. The features were superimposed on the state outlines to obtain the feature value distribution in that particular state. The individual feature values were separated out and their percentage composition of the entire state space was jotted down.

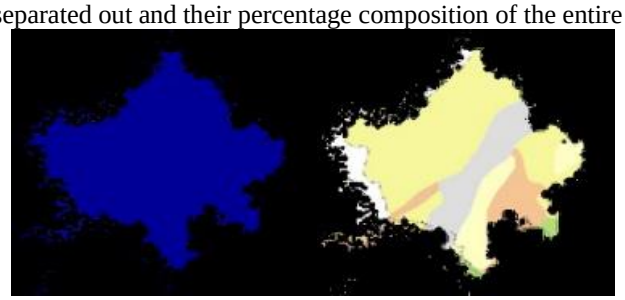


Fig. 5 Rajasthan – Climate Feature Value Extraction

Soil Details:

Mountain	=>	0.054971 percent
GreyBrown	=>	24.0433 percent
GlacierSkeletal	=>	0.14588 percent
Red	=>	11.0174 percent
Alluvial	=>	50.2326 percent
Sub-montane	=>	0 percent
Sub-montane	=>	0 percent
RedYellow	=>	12.5988 percent
Black	=>	1.9071 percent

In addition to the overall statistical distribution of a state, a table stating the maximum proportion value presence of each feature has been built, for e.g. the average rainfall in Rajasthan is generally in ranges of 20 to 60 cm, to which the state is mapped to, in the table, along with the proportion presence.

Rajasthan 20 41.8747

3. Input-Output Modules

3.1 Input modes

3.1.1 City Name - Input speech processing

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

The Google Speech API (in Python) has been used here to recognize the voice input by the user. As already told, the need for voice recognition is for ease of the general user, the farmer in our case who is unaware of the technologies. The local language could add a better value, here, the language supported is English only.

```
Enter the city name:
You said: Coimbatore
Did the entry match: 1 for Yes and 0 for No1
```

The input is given by the user through a microphone which is then converted to textual form by the speech engine. The farmer is prompted to check whether the text is the same as that was spoken. If yes, the contents are then given to the classifier for further processing.

3.1.2 Latitude and Longitude: Input Parameters

A mapping from the latitude and longitude values as input to the appropriate pixel values has been performed. The city name to latitude and longitude conversion is done by means of an index built from the existing dataset.

3.1.3 State Name

This makes use of the state wise feature distribution, the specific portion where the values with maximum proportion value for each feature is taken into account for predictions.

3.1.4 Feature Vector

The values of the features for a new region is given as input to the system. The system using the classifier model predicts the resultant crop(s).

3.2 Output text-to-speech conversion

The result of crop(s) suggested is transmitted as voice back to the user. This is done by using a module in MatLab which is a part of the System's Speech Synthesizer. The appropriate volume levels can be set on the object and speak function propagates the text information as voice back to the user.

SYSTEM ARCHITECTURE

The system architecture is as shown in Fig. 6. The system with the Input/output modules and the processing system has been depicted.

The input to the system is the dataset of map images depicting the different features.

The output from the system is the recommended crop(s) and the state wise feature distribution.

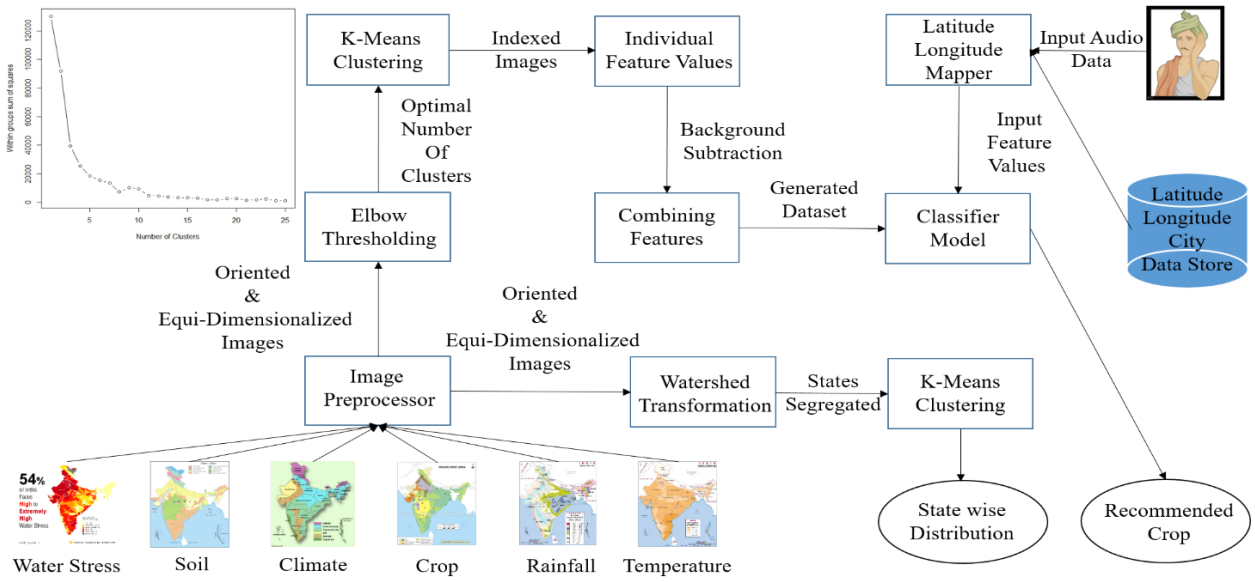


Fig. 6 System Architecture

EXPERIMENTAL RESULTS

An extensive experimentation was done using the generated data set to determine the performance accuracy obtained from each of the classifiers.

The entire dataset was split into training set (67%) and test set (33%). The classifiers used were Naïve Bayes, K-Nearest Neighbors, SVM (Linear), SVM (RBF kernel), Random Forest and Ensemble Learners. The training data set was used to

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

train each of the classifier models. The models built were then tested using the testing data set. The classification report and confusion matrix obtained as a result of the process were recorded. The classification report consisted of the following three different metrics for analyzing the classifier model:

1. Precision
2. Recall
3. F1

The description and the computational procedures for each of the performance measure is shown in Table 2.

The Scikit-learn library in Python has been used for performing the computations for each of the classifiers. The results obtained are tabulated in Tables 3,4,5,6 and 7.

The Confusion Matrix is a N x N matrix, N is the number of class labels, which depicts the True Positive, True Negative, False Positive and False Positive values.

From the results, it can be seen that the ensemble and random forest model perform better than Naïve Bayes (NB) and K Nearest Neighbors (KNN), and gets a score of 77% predictive accuracy. This can be attributed to the dependencies between the features and the large amount of data that disrupts the predictive power of NB and KNN.

Table 2 Computation of Performance Metrics

Performance Metric	Explanation	Computation
Precision	System's ability to **not**	$\frac{\text{True Positives}}{\text{True Positives}}$
Recall	System's ability to identify	$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$
F1	Weighted harmonic mean of	$2 * \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}}$

Table 3 Naïve Bayes Classifier

6404	137	228	1614	254	1146	495
100	4126	231	14	1	1675	14
137	1005	26484	2542	3614	1098	496
564	29	2614	13368	166	780	1
524	46	5874	701	26796	60	117
56	990	1850	627	275	12337	113
994	260	557	172	523	1339	3001

	Precision	**Recall**	**F 1 Score**	**Support**
0	0.37	0.58	0.45	10278
1	0.3	0.9	0.45	6161
2	0.52	0.35	0.42	35376
3	0.58	0.6	0.59	17522
4	0.64	0.63	0.63	34118
5	0.56	0.34	0.42	16248
6	0.27	0.23	0.25	6846
avg/total	**0.53**	**0.5**	**0.5**	**126549**

Table 4 K Nearest Neighbors

6404	137	228	1614	254	1146	495
100	4126	231	14	1	1675	14
137	1005	26484	2542	3614	1098	496
564	29	2614	13368	166	780	1
524	46	5874	701	26796	60	117
56	990	1850	627	275	12337	113
994	260	557	172	523	1339	3001

	Precision	**Recall**	**F 1**	**Support**
0	0.73	0.62	0.67	10278

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

1	0.63	0.67	0.65	6161
2	0.7	0.75	0.72	35376
3	0.7	0.76	0.73	17522
4	0.85	0.79	0.82	34118
5	0.67	0.76	0.71	16248
6	0.71	0.44	0.54	6846
avg	**0.74**	**0.73**	**0.73**	**126549**

Table 5 SVM (kernel = RBF)

6970	270	168	1633	250	616	371
116	4094	447	5	6	1481	12
231	333	27772	1841	3652	1175	372
533	29	2331	14223	110	295	1
513	26	5052	767	27375	75	310
80	1107	1678	692	297	12284	110
1083	396	690	26	79	1154	3418

	Precision	**Recall**	**F 1 Score**	**Support**
0	0.73	0.68	0.7	10278
1	0.65	0.66	0.66	6161
2	0.73	0.79	0.76	35376
3	0.74	0.81	0.77	17522
4	0.86	0.8	0.83	34118
5	0.72	0.76	0.74	16248
6	0.74	0.5	0.6	6846
avg / total	**0.76**	**0.76**	**0.76**	**126549**

Table 6 Bagging (with KNN)

7354	268	156	1268	244	600	388
128	4152	436	7	1	1425	12
277	339	27766	2541	3113	1020	270
1241	26	1636	14230	107	281	1
512	30	5476	792	26906	87	315
279	1144	1707	642	219	12148	109
1072	408	673	27	69	1140	3457

	Precision	**Recall**	**F 1**	**Support**
0	0.68	0.72	0.7	10278
1	0.65	0.67	0.66	6161
2	0.73	0.78	0.76	35376
3	0.73	0.81	0.77	17522
4	0.88	0.79	0.83	34118
5	0.73	0.75	0.74	16248
6	0.74	0.5	0.6	6846
avag/total	**0.76**	**0.76**	**0.76**	**126549**

Table 7 Random Forest

7021	269	152	1618	253	599	366
128	4140	442	3	2	1434	12
221	331	27918	1912	3613	1034	347
536	25	2163	14390	101	306	1

ISSN-2394-5125 VOL 06, ISSUE 07, 2019

486	16	4836	752	27614	102	312
77	1143	1692	706	285	12235	110
1080	406	654	29	69	1145	3463

	Precision	**Recall**	**F 1 Score**	**Support**
0	0.74	0.68	0.71	10278
1	0.65	0.67	0.66	6161
2	0.74	0.79	0.76	35376
3	0.74	0.83	0.78	17522
4	0.87	0.81	0.82	34118
5	0.73	0.75	0.74	16248
6	0.75	0.51	0.61	6846
avg/total	**0.77**	**0.77**	**0.76**	**126549**

CONCLUSION

The use of map image datasets has proved to show significant results. The audio processing at input and output can significantly bridge the gap between the less aware farmers and the powerful technologies available. Future work may focus on obtaining proper datasets that are specific to a smaller bounded location in place of the impractical generalization that has been assumed here by means of image datasets. Better classifier models can be used. Feature ranking and inverse Feature Ranking by having rows as columns and vice versa can be done to determine the best features. These huge volumes of data generated can be classified and handled in a better way using Hadoop based Systems such as Spark, Storm etc. Extending the model with appropriate dataset and parallel processing will be fruitful for the application

REFERENCES

[1] S. Latu. Sustainable Development: The role of GIS and visualisation. EJISDC, 38(5):1–17, 2009.
[2] M. G. Naicong Li, Robert Raskin and K. Janowicz. An ontology-driven framework and web portal for spatial decision support. Transactions in GIS, 16(3):313U329, ″ 2012.
[3] Neha. Building crop ontology for farmers. Master's thesis, Banasthali University, Rajasthan, 2012.
[4] T. eberson Retna Raj, R.; Sasipraba. Disaster management system based on GIS web services. In Recent Advances in Space Technology Services and Climate Change (RSTSCC), 2010.
[5] V. K. Yash Jain, Amita Sharma and S. Chaudhary. Spatial analysis for generating recommendations for agricultural crop production. In India Conference On Geospatial Technologies And Applications (ICGTA-12), 2012.
[6] V. Kumar, V. Dave, R. Bhadauriya, S. Chaudhary. KrishiMantra: Agricultural Recommendation System. ACM - Dev'13, 2013.
[7] J. Konaté, A. G. Diarra, S. O. Diarra, A. Diallo . SyrAgri: A Recommender System for Agriculture in Mali. Mdpi Information-2020(11), pp. 561, 2020. doi:10.3390/info11120561.