The Design of Hybrid Crop Recommendation System using Machine Learning Algorithms

Viviliya B, Vaidhehi V

Abstract: Crop Recommendation System for agriculture is based on various input parameters. This paper proposes a hybrid model for recommending crops to south Indian states by considering various attributes. The recommender model is built as a hybrid model using the classifier algorithm such as Naïve Bayes, J48 and association rules. Based on the appropriate parameters, the system will recommend the crop. Technology-based crop recommendation system for agriculture helps the farmers to increase the crop yield by recommending a suitable crop for their land with the help of geographic and the climatic parameters. The proposed hybrid recommender model is found to be effective in recommending a suitable crop.

Keywords: Recommendation System, Naive Bayes, J48, association rules, Hybrid model.

I. INTRODUCTION

India is an agricultural country with the second-highest land area of more than 1.6 million square-kilometers under cultivation [1]. Most of the Indian population is involved in agriculture and the Indian economy is largely dependent on agriculture. India possesses a power potential to be a superpower in the field of agriculture. Agriculture promotes economic growth and provides a way for rural development. The usage of technology-enhanced practices in agriculture is found less in India. Due to such old practices, the yield of the crop is not as estimated that made the farmers lose hope in the agriculture process. Therefore, many farmers have neglected agriculture and started other occupations. In some cases, suicidal attempts by the farmers have also increased [2].

Thus the need for technical assistance in this field is highly recommended. As agriculture is carried out from ages, past agricultural past data can be easily collected and used for recommendation. Therefore, this research work is to help the farmers in choosing a suitable crop for their land by generating the list of recommendations. The recommendation system is designed based on the necessary attributes such as the geographic and the climatic parameters with the region, soil type, nutrient contents of the soil, temperature and groundwater level the crop. Thus, the farmers need not waste their time and lose money by cultivating the other crops which are not suitable for their land. This research will help the farmers to increase the yield of agricultural production by suggesting an appropriate crop for the agricultural land.

There is no universal system to assist farmers in agriculture. And there is a need to reduce the suicide of farmers. The objective of the paper is to build a recommender system to increase the crop yield. The system helps the farmers in selecting a suitable crop for their agricultural land based on the required parameters.

The system is to design and develop a recommendation model to generate recommendations for crops based on geographical and climatic parameters using machine learning algorithms.

This research paper proposes a hybrid crop recommendation system using classifiers such as Naïve Bayes, J48 and association rules. The performance of the proposed model is found to be effective in recommending the crop based on various input parameters. This paper considers the geographical characteristics of south Indian states.

The various steps involved in this research work are data collection, preprocessing, building recommender model, training and testing the recommender model. The first step of preprocessing feature selection. As the dataset is an imbalanced dataset, the second step of preprocessing is to balance the dataset. The proposed recommender model is designed, trained and tested to recommend a suitable crop to the farmer.

This research is to propose a hybrid recommender model that recommends the most suitable crop for the given inputs. The performance of the model is validated on various parameters and it is found that the hybrid recommender model is effective in recommending the items to the user.

The paper structure is as follows section II Literature Survey, Section III Methodology with dataset collection and crop recommendation using hybrid model, Section IV Results and Discussion, Section V Conclusion and References.

II. LITERATURE REVIEW

Mansi Shinde et.al [3], crop recommendation is implemented using the random forest algorithm based only on the soil quality which is determined based on NPK values of the soil. Jasmine Hsu et.al [4], crop recommendation is implemented using linear Support Vector Classifier, Decision trees, and random forest algorithms based on the data collected from National Oceanic and Atmospheric Administration (NOAA) which includes climate and weather parameters. Nilesh Dumber et.al [5] recommendation system is designed using a genetic algorithm (GA), association rule mapping (ARM) and the knowledge discovery in the database (KDD) using weather forecast report, soil and land use survey of India, soil report.

Revised Manuscript Received on December 05, 2019.

* Correspondence Author
Ms. Viviliya B received her BCA degree from Bishop Heber College, Bharathidasan University, TamilNadu.
Prof. Vaidhehi V received her MSc degree from Bharathidasan University, TamilNadu and MPhil degree from Kamaraj University, TamilNadu.
The Design of Hybrid Crop Recommendation System using Machine Learning Algorithms

Based on the soil is implemented by Rohit Kumar Rajak et al. [18]. Crop Recommendation system using an ensemble approach is implemented by S. Padumalar et al. [19] by considering only soil type using a random tree, k-nearest neighbor and naive Bayes are combined as an ensemble approach. The design and implementation of crop and fertilizer recommendation systems are discussed by Hao Zhang et al. [20] considering meteorological data, soil data, and crop data. The usage of genetic algorithms to maximize crop yield and to sustain soil fertility is explained in detail by Oladapo J. Olakulehin et al. [21] using different attributes such as soil depth, availability of water, drainage system, aeration, pH, mineral compositions, organic matter, and soil organisms. The performance of various classification algorithms in Data mining is elaborated by Bhuvana et al. [22] which helps to identify the most appropriate algorithm for the agriculture domain.

Through the literature survey, it is identified that there is no universal recommender system for the farmers with a huge dataset and technology is a barrier for the uneducated farmers. Hence, this research paper can help farmers to increase crop yield with the recommended crops. It is noted from the literature that the crop recommendations are predominantly based on soil type. Very few research work has considered multiple parameters for generating suitable recommendations. The crop recommendation system for other countries is existing whereas few research activities are based on Indian based crop recommendation system. The existing literature shows that the design of the crop recommendation system is implemented only for selected crops or selected geographic locations. Therefore this research paper is based on the design of the crop recommender system based on multiple parameters for south Indian states which can recommend multiple crops.

III. METHODOLOGY

The various steps involved in this research work are data collection, preprocessing, building recommender model, training and testing the recommender model which is shown in “Fig. 1”.

Fig. 1. Research Overview
A. Dataset Collection

Dataset is taken from the Indian government agricultural portal www.data.gov.in. The dataset contains State Name, District Name, Crop Year, Season, Area, Production and Yield information from 1997 to 2015. Attributes such as Rainfall, Groundwater level, Temperature, Water availability and Soil type are taken from [23]. Attributes such as Fertilizers, Nitrogen, Phosphorous and Potassium Values, Organic Carbon value and Soil pH value are taken from [24].

The details of all the attributes are listed as follows. Groundwater level provides information about the groundwater level of the particular district. Temperature gives information about the temperature of the particular district. Rainfall is the amount of rainfall of the particular district. Water availability specifies the water availability of the particular district and also about the amount of water required for the particular crop for cultivation. Soil type is about the various types of soil such as sandy loam, clay loam, black, red, alluvial and lateritic.

Organic Carbon gives information about the amount of carbon present in the soil. N, P and K values provide information about the amount of nitrogen, phosphorus, and potassium present in the soil respectively. District Name tells the name of the district in that particular state. Crop Year contains the previous ten-year records of the crops. Season tells the season in which the crop is sown. The Area specifies the area of the cultivation land in hectares. Production is the yield measured in tons. Crop specifies the name of the crop to be recommended to the farmer.

Time locality and Space locality specifies the latitude and longitude information about the region. Fertilizer gives information about the fertilizer to be used in that soil. Soil pH specifies the soil nutrient content present in the agricultural land. Sowing time specifies the crop sowing period. Crop type specifies whether the given crop is Food Crop, Cash Crop or Plantation Crop.

B. Data Preprocessing

The basic preprocessing tasks such as data integration, attribute selection, attribute extraction, and data reduction are implemented in this work as given in “Fig. 2”.

![Data Preprocessing Steps](Image)

**Fig. 2. Data Preprocessing Steps**

**Data Integration**

Data is collected from different resources are integrated into a single dataset. The different attributes from different sources are collected and integrated together as a dataset to build the recommender model. The dataset is a highly imbalanced dataset and its description is given in Table I.

<table>
<thead>
<tr>
<th>Particulars</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of classes (crops)</td>
<td>124</td>
</tr>
<tr>
<td>Total number of records</td>
<td>246092</td>
</tr>
<tr>
<td>Minimum number of instances per class</td>
<td>4</td>
</tr>
<tr>
<td>Maximum number of instances per class</td>
<td>15082</td>
</tr>
<tr>
<td>Number of attributes</td>
<td>20</td>
</tr>
</tbody>
</table>

**Table-1: Imbalanced dataset**

**Attribute Selection**

Based on the correlation measure, all attributes are evaluated. Out of 20 attributes, 15 attributes are selected such as Groundwater level, Temperature, Rainfall, Water availability, Soil type, Organic Carbon, N, P and K Values, District Name, Crop Year, Season, Area, Production and Crop are considered as significant based on their correlation value. These attributes are used in building machine learning models.

Identified attributes are Groundwater level, Temperature, Rainfall, Water availability, Soil type (sandy loam, clay loam, black, red, alluvial, lateritic), Organic Carbon, N, P and K Values, District Name, Crop Year, Season, Area, Production, Crop, Time locality, Space locality, Fertilizer type, Crop type, Sowing time (Duration) and Soil pH value.

Finalized attributes are Groundwater level, Temperature, Rainfall, Water availability, Soil type (sandy loam, clay loam, black, red, alluvial, lateritic), Organic Carbon, N, P and K Values, District Name, Crop Year, Season, Area, Production and Crop.

**DATA REDUCTION**

Data amputation is implemented to reduce the noise in the data. In this work, instance amputation is done to balance the imbalanced dataset. Steps for balancing the dataset is stated as:

S:1 – Identify the minimum and maximum instance from the dataset

S:2 – For each class label Ci, calculate size [Ci]

S:3 – Set the step count as 500 and the total number of instances for each Ci from minimal to maximal instances of the dataset

S:4 – Calculating the instances

The description of the dataset after data reduction is given in table II.

<table>
<thead>
<tr>
<th>Particulars</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of classes (crops)</td>
<td>24</td>
</tr>
<tr>
<td>Total number of records</td>
<td>172568</td>
</tr>
<tr>
<td>Minimum number of instances per class</td>
<td>4116</td>
</tr>
<tr>
<td>Maximum number of instances per class</td>
<td>15082</td>
</tr>
<tr>
<td>Number of attributes</td>
<td>15</td>
</tr>
</tbody>
</table>

In this way, the minimum instance is set to 4116 and the number of classes reduced from 124 to 24. This preprocessing is to check whether all the states and districts are covered in the dataset.
At this point of 4004 instances, though the number of crops is reduced from 124 to 24, all the districts in India are covered. So there will be no mismatch while recommending the crops. The details of every iteration of class balancing are shown in table III.

### Table- III: Iteration of Class Balancing

<table>
<thead>
<tr>
<th>Number of classes (crops)</th>
<th>Minimum number of instances per class</th>
</tr>
</thead>
<tbody>
<tr>
<td>124</td>
<td>4</td>
</tr>
<tr>
<td>52</td>
<td>504</td>
</tr>
<tr>
<td>47</td>
<td>1004</td>
</tr>
<tr>
<td>42</td>
<td>1504</td>
</tr>
<tr>
<td>38</td>
<td>2004</td>
</tr>
<tr>
<td>36</td>
<td>2504</td>
</tr>
<tr>
<td>35</td>
<td>3004</td>
</tr>
<tr>
<td>28</td>
<td>3504</td>
</tr>
<tr>
<td>24</td>
<td>4004</td>
</tr>
<tr>
<td>20</td>
<td>4504</td>
</tr>
</tbody>
</table>

### Attribute Extraction

Based on the results of feature selection, the selected features and instances are extracted and stored as .csv files. As the dataset is still imbalanced, instances for southern states alone are considered for this research work. The description after feature extraction is shown in Table IV.

### Table- IV: Dataset after Feature Extraction

<table>
<thead>
<tr>
<th>Particulars</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of classes (crops)</td>
<td>24</td>
</tr>
<tr>
<td>Total number of records</td>
<td>34716</td>
</tr>
<tr>
<td>Minimum number of instances per class</td>
<td>4116</td>
</tr>
<tr>
<td>Maximum number of instances per class</td>
<td>15082</td>
</tr>
<tr>
<td>Number of attributes</td>
<td>15</td>
</tr>
</tbody>
</table>

### C. Hybrid Crop Recommendation System

Crop recommendation is generated as given in “Fig. 3”. Naive Bayes and J48 classifiers are considered in this research work. The final recommendations are generated by association rules based on the results of the classifiers. Classification is a process of assigning an entity to an already defined class by analyzing the features.

**Fig.3. Hybrid Recommender Model**

**Naive Bayes Algorithm:**

The Naive Bayes classifier works on a simple concept. But it is comparatively intuitive. The Naive Bayes classifier is based on the Bayes rule of conditional probability. It makes use of the attributes defined in the dataset. All the attributes are analyzed individually as they are equally important and independent of each other. In addition, Naive Bayes performs better than many other algorithms that are more complex.

**J48 Decision Tree:**

It is a predictive machine-learning model. This decides the target value of a new sample. The target value is nothing but the dependent variable. The value decided is based on the different other attribute values of the present data which is available. Different attributes of the decision tree are denoted by the internal nodes of the decision tree. The possible values of the attributes in the observed sample are denoted by the branches. The final value of the dependent variable is denoted by the terminal node. The dependent variable is the attribute that is to be predicted. That will be predicted based on all the other attributes values. The independent variables in the dataset are the other attributes that help in predicting the value of the dependent.

The following simple algorithm works as the base for the J48 decision tree classifier. Whenever it encounters a training set, it identifies the attribute that discriminates the various instances most clearly. The feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. The target value obtained can be assigned and branches can be terminated for the data instances falling in same value category. This is for the cases with no ambiguity. For the other cases, the other attribute which gives the highest information gain value is used.

**Rule Generated:**

Based on the results of the classifier algorithm Naive Bayes and J48 and using the association rule the list of recommendations is generated by the model. The crops are recommended on the basis of the generated rules which are based on a hybrid recommender model.

### D. Training the Model

The models are designed using a 10 fold cross-validation method. This method is considered to be more effective than the data split method. In this 10 fold cross-validation method, the dataset is divided into ten equal sets and in each fold of training the model nine datasets are used to train the model and one dataset is used to test the model. The average learning of all the 10 folds helps the model to learn the data effectively.

**E. Testing the Model**

The designed model is tested for various performance measures. The following parameters such as accuracy, TP Value, FP Value, Precision, Recall, F-Measure, MCC, ROC Area, and PRC Area are considered in this work.

**ACCURACY:**

Accuracy is the primary metric for evaluating classification models. The best accuracy of 100% indicates that all the predictions are correct. It is calculated as given in equation(1),

\[
\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}
\]

(1)

In terms of a true positive, false positive, true negative and false negative, it is calculated as in equation(2),

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

(2)
RESULTS AND DISCUSSION

The machine learning models used in this research work are designed, trained and tested in Weka. The application is designed using HTML, CSS, JavaScript, and Python. The machine learning models are integrated into the application. The user interface of the application is shown in “Fig. 5”.

The results of this research work are stated in two steps. In the first step, results of the preprocessing are discussed and in the second step, the performance of the recommender system is discussed.

A. Results of the preprocessing

The preprocessing tasks are implemented as stated previously. The objective of this research work is to recommend a suitable crop in south India. Thus, after the preprocessing steps, the crops are reduced based on the number of instances. The total crops in the imbalanced dataset are 124. But it is reduced to only 24 crops after the preprocessing steps. Even after reducing the crops from 124 to 24 based on instances, the preprocessed dataset also covers all the states and districts in India. So, the dataset covers all the regions as required. Thus the preprocessing steps did not affect the coverage of instances of the required region. This is shown in Table V.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)
\]

TP RATE:
True Positive is the positive proportion of the data which are the instances correctly classified as a given class.

FP RATE:
False Positive is the positive proportion of the data which are the instances incorrectly classified as a given class.

PRECISION:
It is the ratio of the relevant instances obtained from the retrieved instances. It is also called a positive predictive value.

RECALL:
It is the ratio of the total amount of relevant instances that were actually retrieved from the dataset. It is also called as sensitivity.

F-MEASURE:
It is based on the machine learning Weka evaluation. For a two-class classifier, 3 f-measures will be generated. For the proposed method, f-measure for NB classifier, f-measure for J48 and a weighted f-measure for these two algorithms are generated.

MCC:
Matthews Correlation Coefficient is the measure of two-class classifications and its quality. The regression coefficient of the problem is the geometric mean.

ROC AREA:
The receiver operating characteristic curve (ROC) is used to denote the connection in a graphical representation. It is calculated as

\[
\text{ROC} = \frac{\text{Total number of True Positives}}{(\text{True Positive} + \text{False Negative})} \quad (3)
\]

PRC AREA:
The Precision-Recall Curve (PRC) is a useful measure. When the classes are very imbalanced PRC is used to calculate the success of prediction. Result relevancy is calculated by precision and the amount of truly relevant results is calculated by recall in information retrieval.

Table-V: Region cover before and after preprocessing

<table>
<thead>
<tr>
<th>ATTRIBUTES</th>
<th>BEFORE PRE-PROCESSING</th>
<th>AFTER PRE-PROCESSING</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>District</td>
<td>646</td>
<td>646</td>
</tr>
<tr>
<td>Seasons</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Crop</td>
<td>124</td>
<td>24</td>
</tr>
</tbody>
</table>

B. Performance of the Recommender System

After the preprocessing steps, the machine learning models are designed using preprocessed dataset. 10 fold cross-validation is used to train the model. The performance of the model is measured on various parameters such as accuracy, TP Value, FP Value, Precision, Recall, F-Measure, MCC, ROC Area, and PRC Area. The performance is recorded and tabulated in Table VI.

Table- VI: Performance Measures of the Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>NB</th>
<th>J48</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>69.90</td>
<td>95.93</td>
</tr>
<tr>
<td>TP Rate</td>
<td>0.700</td>
<td>0.959</td>
</tr>
<tr>
<td>FP Rate</td>
<td>0.024</td>
<td>0.003</td>
</tr>
<tr>
<td>Precision</td>
<td>0.801</td>
<td>0.883</td>
</tr>
<tr>
<td>Recall</td>
<td>0.700</td>
<td>0.959</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.716</td>
<td>0.864</td>
</tr>
<tr>
<td>MCC</td>
<td>0.682</td>
<td>0.863</td>
</tr>
<tr>
<td>ROC Area</td>
<td>0.975</td>
<td>0.998</td>
</tr>
<tr>
<td>PRC Area</td>
<td>0.840</td>
<td>0.964</td>
</tr>
</tbody>
</table>

The accuracy of the classifier is based on the correctly classified instances. The performance of the model is evaluated based on accuracy. So, if the accuracy is high, then the performance of the recommender model built is with better recommendations. The performance of the model is easily depicted as a graph in “Fig. 4”.

Fig. 4. Performance of the Classifiers
The Design of Hybrid Crop Recommendation System using Machine Learning Algorithms

The trained model is integrated into the application and the user inputs are shown in “Fig 5”. Users can specify the option to select the details of the state, District, Rainfall, Ground Water Level, Temperature and the Season and all necessary input parameters. The list of recommendations is generated as primary and secondary based on the hybrid recommender model running in the backend.

The result of the recommendations is shown in “Fig. 6”. The output of the recommended crops as the primary crop and the alternate crop.

V. CONCLUSION

The crop recommendation system recommends crops to the user based on the previous data and the current geographic and climatic parameters of the user. Based on the input, a suitable crop will be recommended. As of now, it is built with only two classifier algorithms and association rule which helps the farmers in choosing the appropriate crop for their agriculture land by the proposed recommender model which in turn will increase the yield and the profit.

This implementation considers 15 attributes and 2 algorithms for the recommendation and the system can be enhanced with more classifier algorithms. Also, the ensemble approach of classifying can be explored. Though the system is developed with a huge amount of dataset, there are some limitations like the farmer cannot locate their exact place with the latitude and the longitudes. The system allows the user to enter the place by selecting their state and the district. Recommendation for the seeds and fertilizers can also be added into the model as the future enhancement.

REFERENCES

1. https://en.wikipedia.org/wiki/Agriculture_in_India
2. https://en.wikipedia.org/wiki/Farmers'_suicides_in_India


23. Dr. P. Chandra Shekara, Dr. N. Balasubramani, Dr. Rajeev Sharma, Dr. Chitra Shukla, Dr. Ajit Kumar, Bakul C. Chaudhary, “Farmer’s Handbook on Basic Agriculture”, German Federal Ministry for Economic Cooperation and Development (BMZ), Second Edition: August 2016.


AUTHORS PROFILE

Ms. Viviliya B received her BCA degree from Bishop Heber College, Bharathidasan University, TamilNadu. She is currently pursuing her MCA degree in CHRIST (Deemed to be University), Bengaluru, Karnataka. She is a Web developer and her areas of interest include Artificial Intelligence and Machine Learning.

Prof. Vaidhehi V received her MSc degree from Bharathidasan University, TamilNadu and MPhil degree from Kamaraj University, TamilNadu. She is pursuing her Ph.D. at Jain University, Bengaluru, Karnataka. Presently working as Associate Professor in CHRIST (Deemed to be University), Bengaluru, Karnataka. Her areas of interest include Machine Learning and Recommender Systems.