Journal of the Maharaja Sayajirao University of Baroda
ISSN :0025-0422

# ANALYSIS OF SOUTH INDIAN AGRICULTURE PRODUCTION DATA USING MACHINE LEARNING CLASSIFICATION TECHNIQUES

**Ms.K.S.Leelavathi**, Research Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, Tamil nadu
**Dr.M.Rajasenathipathi,** Head of the Department, Department of Computer Technology, Nallamuthu Gounder Mahalingam College, Pollachi, Tamil nadu

**ABSTRACT**: Southern states of India consists all type of landscapes and soils with numerous irrigation technologies for growing all kinds of agricultural plants. An agricultural yield from South India fulfils our country needs and also generates huge revenue from the exports. In the previous research works in the agricultural mining and machine learning process focus the crop disease prediction and maximization of cropping yields. This work tries to find out the efficient classification technique for classification and analysis of season based yield production in the southern states of India. Naïve Bayes (NB), REPTree, AdaBoost, IBk and Random Tree classifiers from Weka are taken to classify the agriculture production data set based on seasons. An experimental result shows that the REPTree classifier produces 85.23% of accuracy in the 10 cross folds validation.
Keywords : South Indian Agri production, Classification, Season

## I. INTRODUCTION

In India, Agriculture is the one of main source of income to the rural area peoples and also the country. Especially in southern states of Indian landscapes like plains, coastal areas, hill areas, deltas and plateaus helps to growing numerous count of agriculture crops. Pulses, rice, millets, coconuts, fruits and vegetables category of the crops are important varieties in the south indian farming.

Agriculture based income is the only source of micro level farmers. Various kinds of problems faced by the farmers in the agriculture cropping process and preserving the yields like drought, heavy rains and floods, diseases affected in the plants and yields, unhealthy soils and changed climatic conditions. Indian government supports the farmers and encourage to increase the agriculture lands through agricultural loans and subsidies, assistance though agriculture research institutes and department of agriculture. But Indian agriculture system needs better system to forecast the agriculture yields, maximize the production, preserve and market the yields. Data mining helps to find the suitable crops for cropping, protect the crops from diseases and maximize the yields.

Agriculture based data mining techniques helps to farmers in the basis of soil classification and soil fertility diagnosis, analysis and forecasting of rain fall and weather, identification of suitable crops with disease detection in earlier stages, optimized usage of pesticides and insecticides and yield maximization.

Machine Learning (ML) techniques / algorithms/ models covers a learning process with the objective to learn from data set(training) to perform a task. In ML, set of attributes (features/variables) are called data. A feature can be nominal (enumerated like country name, gender), binary, ordinal (e.g., high or medium or low), or numeric (integer, real number, etc.). Performance metrics (statistical and mathematical models) are used to measure the performance of ML model on data (set of attributes). Finalized ML trained model utilized to cluster or classify or predict the test data (set of attributes). Tasks of ML categorized into learning type and learning models or the learning models hired to implement the particular task.

Depends on the data (set of attributes) and learning system, ML tasks are categorised as supervised and unsupervised. Normally data consists the sample inputs with its corresponding outcomes in the supervised learning and its goal is to form a generic rule to relate the inputs to outputs. In some cases, reinforcement learning required to overcome input data and target output with missing data. In the supervised learning, trained model based on training data set is used to predict the output (nominal labelled data) in the test data set. In unsupervised learning, training and

test data sets are unlabelled and also have no distinction between these data sets. The learner practices input data with the objective of ascertaining concealed patterns.

In Agriculture, machine learning is the emerging concept for crop, soil and water management. It helps to predict the crop yield, maximization of crop production, quality checking, leaf / flower disease detection, livestock update and guidance for farm based enterprises. Agriculture based machine / deep learning analysis based on the management of crops, water, soil and live stock. Crop Management consists of Yield Prediction, Disease Detection, weed detection, crop quality, species recognition, maximization of yield and minimization of natural resource utilization.

This paper attempts to find out the prominent machine learning technique to classify the season based yield production in the southern states of India.

## II. REVIEW OF LITERATURE

The existing works mostly concentrates rain fall, temperature and soil details for agriculture data analysis. Ramesh et al [2] analysed East Godavari district based agriculture data (production of agriculture products from 1965 to 2009) using clustering and regression methods with rainfall as dependent variable and area, year and production as independent variable to find the high accuracy level in capabilities of yield production. Verheyen et al [3] utilized DM techniques to analyse the characteristics of soils using K-Means clustering with inclusion of GPS technology. Alberto et al [4] implements machine learning techniques like Linear Regression, SVM, KNN and Regression trees for analysis crop yield prediction. Pantazi et al [5] utlizing unsupervised machine learning techniques to predict the wheat yield with field variation. Soil data from sensors and crop growth details based on satellite images from different soils. Gandhi N et al [6] forecast rice yield production in different climatic conditions using machine learning techniques with the help of SVM. Region of cultivation, normal and extreme temperatures are taken as the parameters for rice yield in Kharif season. Geetha [7] issued overview of horticulture based agribussiness model and approach by utlizing information mining. The author deals different mining systems for horticulture, issues, agribusiness with parameters like overall and seasonal rainfall, region and year.

## III. PROPOSED METHODOLOGY

The proposed methodology attempts to find the proper classification technique to classify the Indian southern states agriculture production based on seasons. From the machine learning models, Naïve Bayes from bayes, IBk from Instance based, Adaboost from ensemble learning, RepTree and Random Tree from tree based models are used to learn and classify the southern Indian states agriculture production data set.

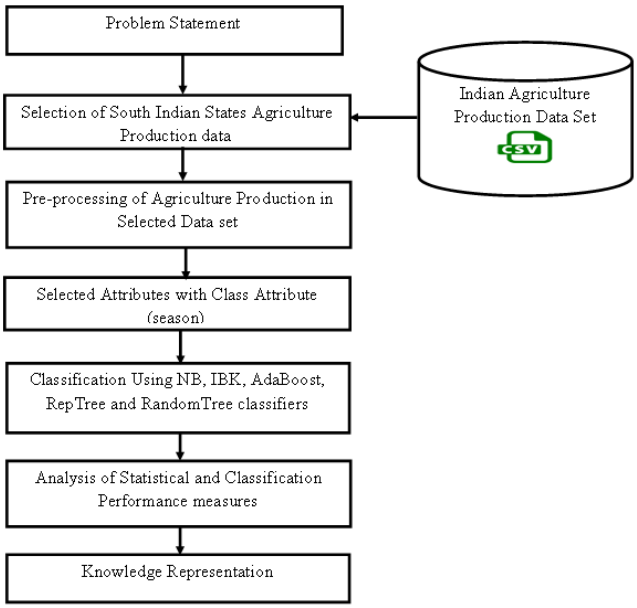


Fig.1. Framework of Classification on Agri yield data

**About the data set:**

The data set is taken from the Indian government data portal with all states agriproduction from the period of 1993 to 2017. From the data set, south indian states are extracted and 43376 instances are selected with 7 attributes. Andrapradesh, Telungana, Karnataka, Kerala, Tamilnadu and Pondicherry states agriculture production details are taken with 7 attributes such as state and district name, crop year, season, name of the crop, total area and production.

The seasons attribute is selected as class attribute. The various crop yield production details are framed based on the seasons such as Kharif, Rabi, Winter, Autumn and Summer.

**Classifiers:**

Naïve Bayes, Adaboost, IBk, RepTree and Random Tree classifiers techniques are used to classify the agriculture production data based on the season.

- IBk is an instance based classification technique utilize the distance measures (euclidean distance - kNN) to find k close instances in the agriculture production data set based on season.
- Naive Bayes assumes that the presence of a unambiguous feature from the class is unrelated to the existence of other features of the class.
- Reptree utilizes decision tree(C4.5 Algorithm) and produce discrete / continuous outcome for classification / regression.
- Random Tree consists set of independent decision trees (generated from different data samples and its subsets of the dataset) and it selects most frequent tree for learning and classification of the data. Selection of most frequent tree helps to reduce the over fitting problems.
- Adaptive Boosting (AdaBoost) is an meta and iterative ensemble method works like a RandomForest but it works with 2 leaf decision tree. Adaboost splits the trees into groups based on the decisions and includes the significance for each tree. For the final classification, selects the group which one consist of largest sum by the RandomForest and perform the classification.

**Performance Measures:**

Classification performance parameters such as true positive, false positive, precision, recall and f-measure values and statistical performance such as kappa and error values like mean absolute, root mean square, relative absolute and root relative squared are taken to analyse the classification performance based on the season (class attribute) in yield production of the southern states of India.

## IV. RESULTS AND DISCUSSION

Naïve Bayes, Adaboost, IBk, RepTree and Random Tree classifiers techniques are used to classify the agriculture production data based on the season. The obtained results from Weka are taken to discuss to find out the prominent machine learning technique.

| Classifier Used | Model Building Time (in Sec) | Kappa statistic | Mean absolute error | Root mean squared error (RMSE) | Relative absolute error (RAE) | Root relative squared error (RRSE) |
|---|---|---|---|---|---|---|
| Naïve Bayes | 0.17 | 0.4104 | 0.1601 | 0.3279 | 73.09 % | 99.09% |
| RepTree | 0.89 | 0.7264 | 0.0697 | 0.2045 | 31.83% | 61.78% |
| AdaBoost | 0.41 | 0 | 0.3112 | 0.3896 | 142.09% | 117.73% |
| IBK | 0.01 | 0.1927 | 0.1785 | 0.4223 | 81.49% | 127.63 % |
| Random Tree | 0.17 | 0.7272 | 0.0603 | 0.2431 | 27.52% | 73.45 % |

Table 1. Statistical measures of various classifiers

The above table describes, IBK classifier takes 0.01 seconds to build the model for south indian agri yield production data set based on the season as class attribute. Naïve Bayes and Random Tree classifiers consumes 0.17 seconds, AdaBoost consumes 0.41 seconds and RepTree classifier takes 0.89 seconds for model building. It shows IBK takes minimum time and RepTree takes highest time to build the model for yield data set based on season.

In the statistical measures, RepTree and Random Tree classifiers outperforms other classifiers such as Naïve Bayes, AdaBoos and IBK. Based on Kappa Statistical measures, RepTree and

RandomTree classifiers closely matched the data label (season – class attribute) as ground truth and expected accuracy. Based on Mean absolute error, RepTree and RandomTree classifiers produces very less error range that denotes both classifiers shows good results in the magnitude of difference between prediction and true values of the observation. Based on RMSE, RAE and RRSE, RepTree and RandomTree classifiers outperforms other classifiers.

Based on the statistical measures, Tree based classifiers classifies the south indian agri yield data based on the season than other classifier. RepTree classifiers perform well than Random Tree classifier. But it takes more time to build the model.

| Classifier Used | Correctly Classified Instance | | Incorrectly Classified Instance | | True Positive | False Positive | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|---|---|---|
| | Total | % | Total | % | | | | | |
| Naïve Bayes | 27010 | 62.27% | 16366 | 37.73% | 0.623 | 0.155 | 0.738 | 0.623 | 0.642 |
| RepTree | 36971 | 85.23% | 6405 | 14.77% | 0.852 | 0.123 | 0.850 | 0.852 | 0.851 |
| Adaboost | 24623 | 56.77% | 18753 | 43.23% | 0.568 | 0.568 | 0.322 | 0.568 | 0.411 |
| IBK | 24025 | 55.39% | 19351 | 44.61% | 0.554 | 0.339 | 0.560 | 0.554 | 0.557 |
| Random Tree | 36885 | 85.04% | 6491 | 14.96% | 0.850 | 0.110 | 0.851 | 0.850 | 0.851 |

Table 2. Performance measures of various classifiers

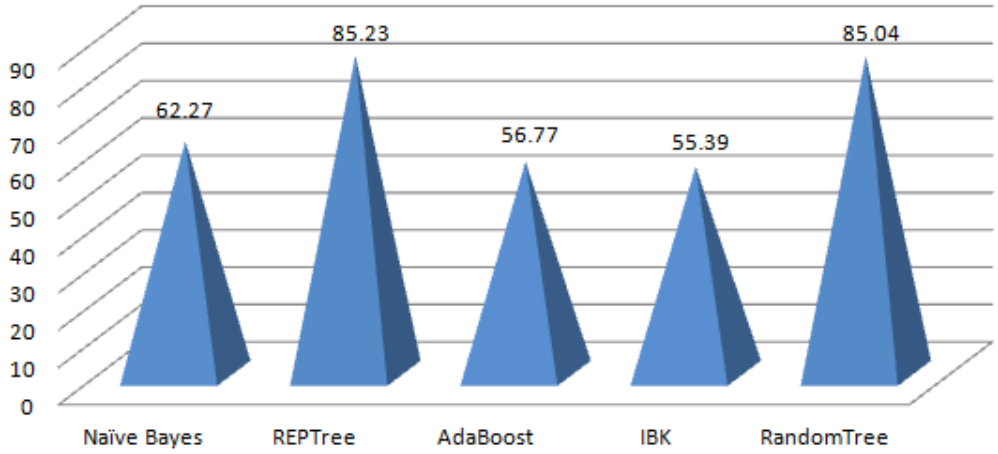


Chart 1 – Classification Performance

The above table 2 reveals that the RepTree classify the Agriyield data set based on the season with high accuracy(85.23%) than other classifiers. Random tree classifier closely followed the RepTree classifier with 85.04% accuracy. Naïve Bayes, AdaBoost and IBK did not performs well with classified accuracy level 62.27%, 56.77% and 55.39% respectively.

From the 43376 instances, RepTree classified 36971 instances correctly with 85.23% accuracy and 6405 instances are not correctly classified by the RepTree.

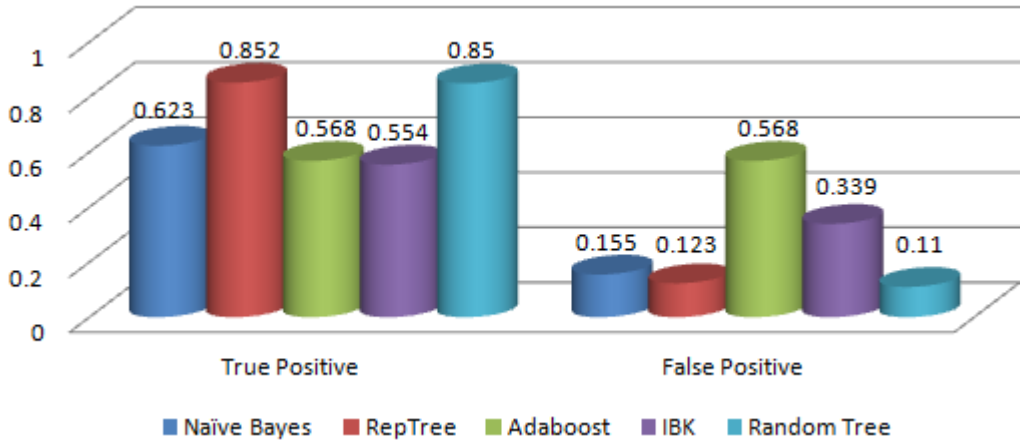


Chart 2 – TP and FP Performance

True Positive (TP) and False Positive (FP) denotes the correctly and incorrectly predicted positive classes respectively. RepTree and Random tree classifiers predict the positive classes better than other classifiers.

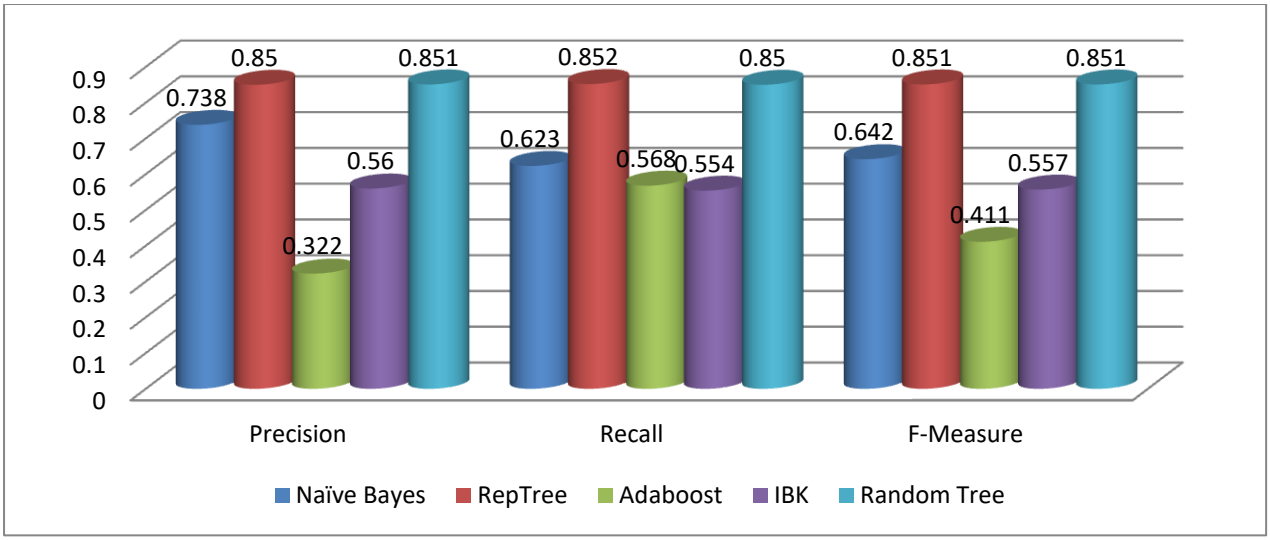


Chart 3 – Precision, Recall and F-Measure based performance

## V. CONCLUSION AND SCOPE FOR FUTURE ENHANCMENT

In this paper, different categories of machine learning techniques are utilized on the South Indian Agriculture production data set to analyse and findout the most effective classification technique. NB, REPTree, AdaBoost, IBK and Random Tree classification techniques are used to classify the data set based on the season. REPTree and RandomTree classifiers performs well when compared to other classifiers such as NB, AdaBoost and IBK.

RepTree classifier performs well than Random Tree classifier. But it takes more time to build the model. The proposed work to be extend with this work to enhance the classification accuracy of RepTree by combining clustering approach and also focus to reduce the model build time

## References

1. MMasrie, A Z M Rosli, R Sam, Z Janin and M K Nordin, Integrated optical sensor for NPK_Nutrient of Soil detection,IEEE 5th ICSIMA 2018, Nov-2018-*28-30*, Thailand
2. D.Ramesh and B.Vishnuvardhan, DM Technique and Appl. to Agri. Yield data In: Intl. Jrnl of Advanced Research in Computers and Comm. Engineering, 2013
3. Verheyen and Deckers, High resolution cont. soil classifying using morphological soil profile descrption, Geoderma. 2001;101:31–48.
4. G S Alberto, F S Juan and O Bustamante. Predictive ability of ML methods for massive crop yield prediction. Span J Agric Res. 2014;12(2):313–28.
5. Pantazi, Alexandridis and Mouazen, Wheat yield prediction using ML and advanced sensing tech., Computer Electronics Agriculture. 2016;121:57–65.
6. Gandhi N, Armstrong and Tripathy, “Rice Crop Yield Prediction in India using SVM”, IEEE The 13th Intl. Joint Conf. On CS and Soft. Eng. (JCSSE), Thailand, 2016.
7. Geeta M C S, “A survey on DM Techniques in Agriculture”, Intl. Jrnl of Innovative Research in Comp. and Comm. Eng., vol. 3,2015