International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 08 | August -2020 ISSN: 2582-3930

Efficient Agricultural data analysis for recommending crop cultivation using Data Mining Techniques

Divya Raikar[1], Prof. Bharathi H.N.[2]

[1] Student, Department of Computer Engineering, K.J. Somaiya College of Engineering, Vidyavihar, Mumbai, India
[2] Department of Computer Engineering, K.J. Somaiya College of Engineering, Vidyavihar, Mumbai, India

---

Abstract - In agriculture sector, farmers and agribusinesses undergo innumerable complexities to make decisions regarding the crop selection to grow based on environmental conditions like temperature, rainfall, information, prices, exports, suicides details are the major factors that affect the productivity of the crop. Various crops are grown in India which includes staples crops like rice and wheat. Other crops are also grown like pulses, potatoes, sugarcane, oil seeds, etc. Despite major achievements in technology, agricultural area faces a major problem in estimating the correct choice of crops. An essential issue for agricultural planning is to accurately estimate yield and crop. Data mining techniques can be used as effective solution for this problem. The aim of this paper is to analysis of the agricultural data and extracting information using data mining techniques to recommend crop to farmers

***Key Words*:** Data Analysis, Classification, Data Partitioning, K-Nearest number, Naïve Bayesian, Decision Tree, Support Vector Machine, Deep Neural Network

1. INTRODUCTION

India is an agricultural country. Agriculture is considered as the backbone of India from where most of the people derive their livings. Agriculture has a lot of impact on the economy of the country. In the past decades, there has been a lot of natural calamities that has occurred which has directly or indirectly hampered the production of the crops across various regions of the country. On the basis of cultivation seasons in India, crops can be divided as Karif, Rabi, Whole Year, summer, winter and autumn. Along with the environmental conditions like rainfall and temperature, the external factors like cost rate, export, area, etc. conditions also affects the productivity. But, over past few years, India has seen many farmers committing suicide due to number of conflicting reasons like-high debt burden, poor government policies, crop failure, environmental issues, etc.

Data mining is the study of extracting useful information from the dataset. Data mining techniques is used to analyze the process of hidden patterns which is then converted into relevant information. These efficient analysis using data mining algorithms can help farmers to take decisions. Data mining follows the following steps: Extracting or storing the data, transforming and loading the dataset, managing the data set. Mining consumption accounts for a significant portion of mining time due to its high computation and input/output intensity. The present study focuses on the effect of various parameters on production of major crops. In this study, the aim is to predict the crop production by applying the following algorithms- K-Nearest Neighbour, Decision tree, Naïve Bayesian algorithms, Support Vector Machine and Deep Neural Network.

The possibility of applications of data analysis in the agricultural sector is very promising as till date very few good researches have been done in this sector in perspective, mainly in predicting the trend and productivity. Predicting the crop the feature attributes is one of the challenges in the field of agriculture. Along with this large population, rising urban and rural income is has become the driving force for agricultural products demand. Also, external demand is driving export from agricultural sector. Thus, as the time grows, the robust demand starts increasing which in turns creates opportunities in agro allied services. Government of India is aiming to double farmer's income by 2022. If accurate prediction of crop is given, it will help farmers. Crop prediction models can be utilized in preciseness of agriculture to enhance field production in order to satisfy growing demand.

At this stage when lots of farmers are drowned in debt, it is very necessary to provide them an efficient source of information to guarantee a proper cultivation according to their area demographics and economical conditions, hereby reducing the suicide rate. Hence a successful implementation will enable us to predict the future outcome of the yield which can in turn help the government to look into any scarcity of the crop. The efficiency of the crop yield prediction system depends on how accurately these all features have been extracted and how these classifiers are implemented. In this paper, it will summaries the results obtained by various algorithms with their accuracy and recommendations

2. RELATED WORK

Farming is the backbone of Indian economy. The farming communities in India are facing a multiple problems to maximize their agriculture productivity There are approaches developed using data mining algorithms .Using Multiple Linear Regression technique [1] can be used to estimate the future yield prediction in tea cultivation with respect to climatic change The parameters selected for the study were temperature, rainfall, relative humidity, sunshine and evaporation for the four regions (South Bank, North Bank, Upper Assam and Cachar) of Assam. Considering tea production as dependent variable and difference of average maximum and minimum temperature, total rainfall , and Average Sunshine hours as independent variables from April to November, multiple linear regression analysis was conducted using 1977- 2006 data (30 years) to predict the crop productivity. The conclusion drawn at the end is that tea production estimation equations developed for the four regions were validated for the future yield prediction (2007, 2009 and 2010) and developed model can be used to predict tea production for each region. Comparison study of classification algorithms [2] was used to focus on finding general trends for

effect of various climatic parameters on the rice crop yield of the Tropical Wet and Dry climatic zone of India for Kharif season. This paper analyzed the correlation between the climatic factors and rice crop yield. Data is visualized in Ms Excel using the scatter plots and then algorithm is executed in open source data mining tool-WEKA. It also aims to predict rice crop yield through the use of the historical data of various factors affecting the Tropical Wet and Dry climatic zone of India by applying data mining techniques Knn and Decision Tree.

The various parameters of data sets, data pre-processing, data mining techniques was studied to build prediction model using Scikitlearn and py-earth[3] on Rice, Wheat and Maize dataset. Regression analysis was used as a predictive modelling technique which used- Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) for analyzing the effect of crop production of Rice, Wheat and Maize over years. Multiple Linear regression algorithm worked good on Maize dataset while Multivariate adaptive regression Splines worked good on Rice and Wheat dataset. A single algorithm did not work on the different crop dataset.

A predictive Apriori algorithm [4] is studied using open source data mining tool (WEKA), i.e., association rule mining for determining the effect of daily temperature and rainfall on paddy yield. The rules resulted from Predictive Apriori algorithm, it stated that there were variation in paddy yields over the period of study, out of which variation recorded for rainfall were not related to the variation noticed in the output of paddy crop yields. Also the paddy yield is directly related to temperature in vegetative phase and inversely effects on the maturation phase. It was because the irrigation facility was good enough to eliminate the adverse effect of rainfall. Although the prediction mechanism is done using data mining algorithms, integrating it with neural network will aim at achieving higher prediction accuracy to determine crop yield to both farmers and other decision makers.

Problem analyses in distribution channel and storage of food product, and precession farming decision making is done using big data analysis with hadoop framework [5]. Pentaho BI allows generating interactive report and highly customizing dashboard that help front end to easy understand and make decision. It gives real time analyzer report about weather, soil, and current status of market and storage capacity and real time demand of market. Map Reduce is a processing technique and a program model for distributed computing. The Map Reduce split data in two parts one is key and value .First it map data with relative data and then reduce that with the distributing with deferent nodes. All nodes mapped with the key nodes store data .Whatever data in the nodes by using deduce function that contain analytical function to reduce them and give fast and reliable data set as output .By using this technique we get data fast and reliable view that will input to the BI application that then generate the 3D data view for decision making and analytical reports. For the decision making, data visualization and chart report are not enough to forecast the predicting result.

The analysis of the agriculture data is carried out to find the optimal parameters to maximize the crop production using data mining techniques like PAM, CLARA, DBSCAN and Multiple Linear Regression [6]. The data mined contains feature set like crop, soil and climatic data, and analysing new, non-experimental data optimizes the production and makes agriculture more flexible to climatic change. The study includes data mining techniques like PAM, CLARA and DBSCAN to obtain the optimal climate requirement of wheat like optimal range of best temperature, worst temperature and rain fall to achieve higher production of wheat crop

In agriculture sector, the entire planning is dependent on type of crop. If wrong crop is cultivated, it can affect the productivity. Data mining techniques are necessary for accomplishing practical and effective solutions for this problem, so then the farmers can use this information which will help them in making critical farming decisions. Hence, in this project the aim is to efficiently analyze the data analysis for recommending crop cultivation different using Data Mining Techniques

3. METHODOLOGY

Initially, the agricultural data will be collected from various sources, which will include details of crops, rainfall, regions, production etc. Pre-processing of the dataset should be done. The processing of the data is carried out in following steps- Data collection, Date pre-processing, Feature extraction and Feature Normalization. Data Mining Techniques, such as the k-NN, Decision Tree and Naïve Bayesian is used to recommend crop cultivation. Further, prediction of the crop production will be done using neural network approach and SVM. Further, the comparison of all these algorithms will be done, and prediction model will be built on the algorithm that gives highest accuracy.

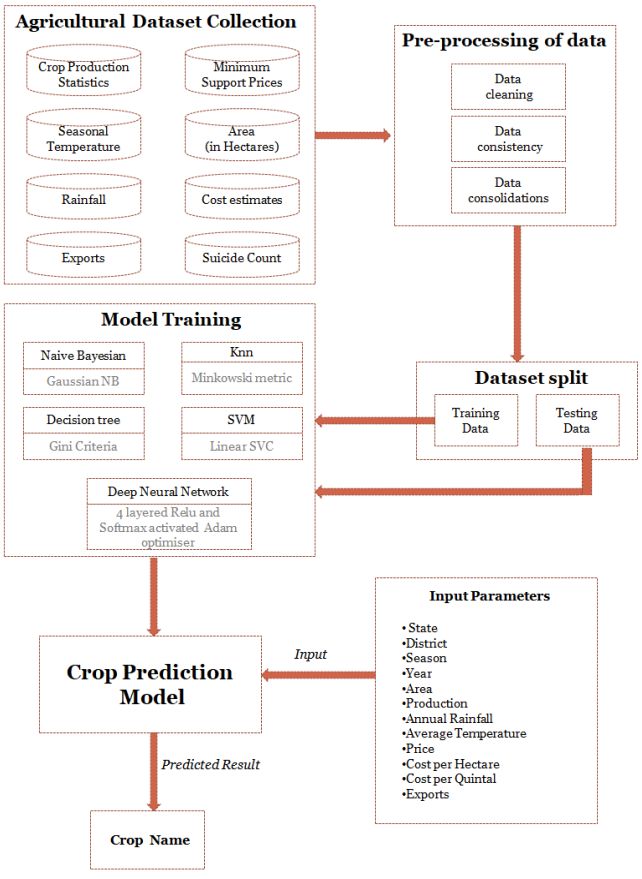


Fig. 1: System Architecture for Prediction of Crop

The architecture is a system that unites its parts or components into a coherent model. The prediction model contains following components- Agricultural dataset collection, pre-processing, data training and data testing, model training using specified algorithms. The crop information base consists of farm data like crop varieties, crop year, area and seasonal (parameter like Kharif, Rabi, summer, winter, autumn and seasonal crops), rainfall, temperature, export and prices. The knowledge-based additionally contains of zones furthermore district information as well. The processing of the data is carried out in following steps- Data collection, Date pre-processing, Feature extraction and Feature Normalization. Data Mining Techniques, such as the k-NN, Decision Tree and Naïve Bayesian, Multiclass Neural network and SVM is used to recommend crop cultivation.

A. Data Collection

The dataset of District-wise, season-wise crop production statistics [7] is publically available for non-commercial use which contains records of Indian Government from year 1997 to 2015 over the period of 18 years. It consists of all states of India, the districts, year of production, crop covered area (Hectare), season (Rabi, Kharif, winter, summer, annual and autumn), production (Tones), yields and crops by contributors- Ministry of Agriculture and Farmers Welfare, Department of Agriculture, Cooperation and Farmers Welfare and Directorate of Economics and Statistics (DES) [8]. The dataset of All India Seasonal and Annual Mean Temperature Series contains time series, which shows a warming over India during the recent years- 1901 to 2017 in degree Celsius [9]. The sub-division wise rainfall dataset containing month wise all India rainfall data measured in mm for each month and season has been provided in the data from 1901 to 2015[14]. Minimum Support Price (MSP) is a form of market intervention by the Government of India to insure agricultural producers against any sharp fall in farm prices. The minimum support prices are announced by the Government of India at the beginning of the sowing season for certain crops on the basis of the recommendations of the Commission for Agricultural Costs and Prices (CACP) [11, 12]. MSP is price fixed by Government of India to protect the producer - farmers - against excessive fall in price during bumper production years. The dataset contains minimum support prices for crops in rs per quintal. The data containing All India and State-wise area in hectares under Principal Crops (food grains, cereals, oil seeds, tobacco, potato, black pepper, chillies, natural rubber, bananas, ginger, coconut and turmeric) from 2001 to 2016[13] and cost estimates of some principal crops in Five Major Producing States During 2008 to 2013[17] is gathered. Export dataset contains Indian agricultural exports data from 1997 to 2019[17]. This data containing details of farmer's suicides from 2010 to 2014 (From: Ministry of Agriculture & Farmers Welfare [15].

Table -1: Agricultural Dataset details

Dataset	Details
District-wise, season-wise crop production statistics	Contains of all states of India, the districts, year of production, crop covered area (Hectare), season (Rabi, Kharif, winter, summer, annual and autumn), production (Tonnes), yields and crops by contributors from 1997 to 2015
Seasonal and Annual Mean Temp Series - India	The dataset contains All India Seasonal and Annual Mean Temperature Series. The time series shows a warming over India during the recent years- 1901 to 2017 in degree Celsius. [8]
Subdivision wise Rainfall	The dataset contains month wise all India rainfall data measured in mm. The sub-division wise rainfall and its departure from normal for each month and season has been provided in the data from 1997 to 2015
Minimum support price of crops	The minimum support prices are announced by the Government of India at the beginning of the sowing season for certain crops on the basis of the recommendations of the Commission for Agricultural Costs and Prices from 2002 to 2020
Cost Estimate of some principal crops	The data refers to cost estimates of some principal crops in Five Major Producing States during 2008 to 2013 containing all actual expenses
All India level Area Under Principal Crops	The data refers to All India and State-wise area in hectares under Principal Crops from 2001 to 2016 in hectares
Exports	This data contains Indian agricultural exports from 1997 to 2019[16

B. Data Preprocessing

- Data Cleaning and Consistency- The dataset contains missing information. To eliminate the null values, data cleaning is done in the beginning.
- Data Consistency- The data collected are of different years and are not same across the databases. To make it consistent, data consistency. The crops, years, missing data, varying formats are made consistent in a consolidated dataset
- Data Consolidation- The data from the above datasets are collected and cleaned and then merged so that it can be analyzed correctly. The dataset size is 3,66,711 records with 13 attributes

C. Data training and testing set

Model Training and Testing is controlled through data partitioning which in turn provide helps to achieve high efficiency and good scalability of distributed execution. It aims at improvising the testing efficiency load by equally distributing data among the models.

All the consolidated data are divided into training and testing set using the train_test_split from sklearn model selection. Data is split into 80-20 ratio

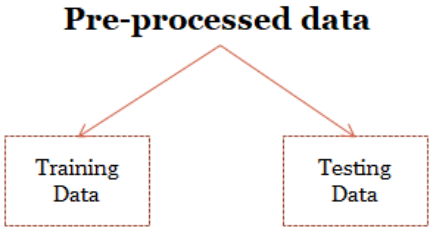


Fig. 2: Agricultural Data Seperation for training and testing

D. K-Nearest Neighbor

KNN is a supervised learning model and instance based lazy algorithms. The dataset is divided into 8:2 proportions to carry out the training and testing phase. From dataset, features and the labels are constructed. For the purpose of label encoding, sklearn is used. Using the KNeighbour Classifier module, the KNN classifier object is implemented by passing the argument number of neighbors in the KNN function. The power metric is set to Minkowski metric. Since the computation is standalone, the number of parallel jobs is set to

None. Using the panda data frame, the iloc indexer is used for integer location based indexing selection by position for data frame output for training and testing.

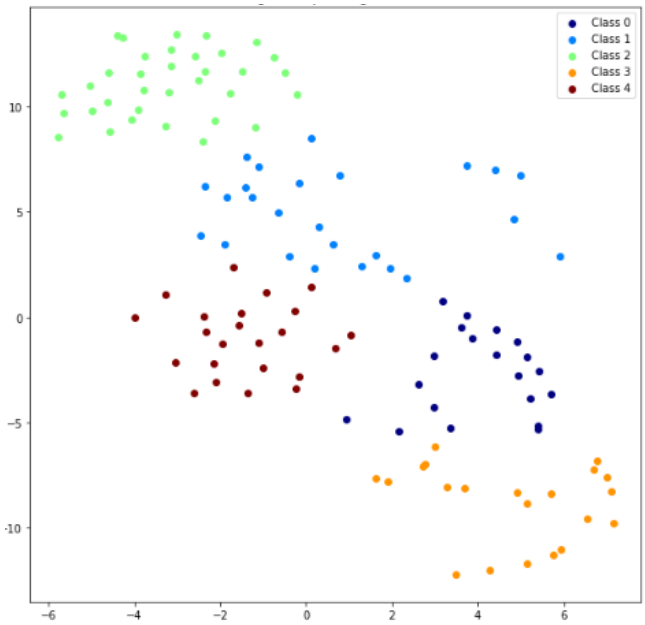
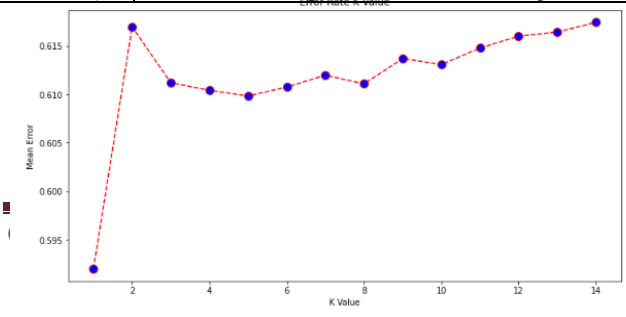


Fig. 3: Crop clustering using training features

On the given dataset, conversion to 2 dimensions using tSNE is done to keep the neighbors in original space closer to lower dimensions. The cluster number was chosen by plotting the data first. On applying the KNN clustering, we have found out that the crops are falling into the following clusters- Fruits and fertile plants, Cereals and vegetables, Ground crops, Dry crops and Oilseeds, Pulses and some vegetable

Table -2: Crop clustering using data features

Cluster	Crops
0 (Fruits and fertile plant)	'Groundnut', 'Papaya', 'Onion', 'Orange', 'Mango', 'Ragi', 'Colocosia', 'Tomato', 'Grapes', 'Arcanut (Processed)', 'Arhar/Tur', 'Arecanut', 'Barley', 'Mesta', 'Dry ginger', 'Lemon', 'Jute & mesta', 'other oilseeds', 'Sapota', 'Pineapple'
1 (Cereals and vegetables)	'Citrus Fruit', 'Garlic', 'Other Vegetables', 'Jack Fruit', 'Korra', 'Drum Stick', 'Tea', 'Bhindi', 'Varagu', 'Brinjal', 'Other Fresh Fruits', 'Linseed', 'Wheat', 'Beans & Mutter(Vegetable)', 'Cabbage', 'Cowpea(Lobia)', 'Pome Fruit', 'Sweet potato', 'Rajmash Kholar', 'Cashewnut Processed', 'Kapas', 'Ginger', 'Maize', 'Turmeric', 'Pome Granet'
2 (Ground crops)	'other fibres', 'Pump Kin', 'Lab-Lab', 'Perilla', 'Peach', 'Ash Gourd', 'Apple', 'Ber', 'Ricebean (nagadal)', 'Water Melon', 'Cond-spcs other', 'Blackgram', 'Turnip', 'Other Citrus Fruit', 'Lentil', 'Bottle Gourd', 'Snak Guard', 'Peas (vegetable)', 'Litchi', 'Plums', 'Redish', 'Carrot', 'Jobster', 'Pear', 'Cauliflower', 'Sannhamp', 'Cucumber', 'Yam', 'Beet Root', 'other misc. pulses', 'Bean', 'Other Dry Fruit', 'Bitter Gourd', 'Ribed Guard'
3 (Dry crops and Oilseeds)	'Pulses total', 'Potato', 'Rice', 'Coffee', 'Gram', 'Rubber', 'Atcanut (Raw)', 'Cotton(lint)', 'Oilseeds total', 'Soyabean', 'Jute', 'Total foodgrain', 'Jowar', 'Bajra', 'Guar seed', 'Banana', 'Coconut ', 'Rapeseed &Mustard', 'Tapioca', 'Moth', 'Sugarcane'
4 (Pulses and some vegatable)	'Bottle Gourd', 'Sannhamp', 'other misc. pulses', 'Other Dry Fruit', 'Pump Kin', 'Cauliflower', 'Lentil', 'Bitter Gourd', 'Other Citrus Fruit', 'Cucumber', 'Snak Guard', 'Water Melon', 'Cond-spcs other', 'Turnip', 'Peas (vegetable)'



Error Rate: The data are plotted against 15 k values. Each k nearest neighbor classifiers generates an error rate for each k fold values. On running the models for 15 neighbors, the result states that the model is best fitted for 5 neighbors which give the least error rate.

Fig. 4: Error rate k values

E. Naïve Bayesian

Naïve Bayesian is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. It is a supervised learning algorithm. Gaussian Naïve Bayesian network is implemented since the features have continuous values. Using Gaussian distribution, it helps to minimize the sum of squared errors. Model is implemented in GaussianNB estimator of sklearn class. The data set is passed through the NB classifier, the score utility function is used from the scikit learn. It returns the mean accuracy on the given test data and labels.

$$P(C|x) = P(x|C)P(C) / P(x) \quad (1)$$

$$P(C|x)=P(x_1|C) \times P(x_2|C) \times \ldots\ldots P(x_n|C) \times P(C) \quad (2)$$

Equation (1) and Equation (2) defines P(C|x) as posterior probability, P(x|C) as likelihood, P(C) as class prior probability and P(x) predictor prior probability

F. Decision Tree

Decision Trees (DTs) are a supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the above extracted data features. Using sklearn library, DecisionTreeClassifier is used since it is capable of performing multi-class classification on a dataset. In this project, the classifier uses gini criterion which is generally a measurement of the likelihood of an incorrect classification of a new instance of a random variable, if that new instance were randomly classified according to the distribution of class labels from the data set. The data set is passed through the Decision Tree classifier, score utility function is used from the scikit learns. It returns the mean accuracy on the given test data that are used to train the features are significant enough to classify the crops into the clusters appropriately.

The results show that high area under cultivation Lower annual rainfall, Increase in temperature, Increase in price and Increase in cost per quintals are reasons for increase in number of suicides. These analyses are more insightful to interpret that if these factors are studied well and if correct measures are taken like- proper irrigation facilities, increase in export, decrease in cost, utilization of the area properly, the suicide rate can be minimized

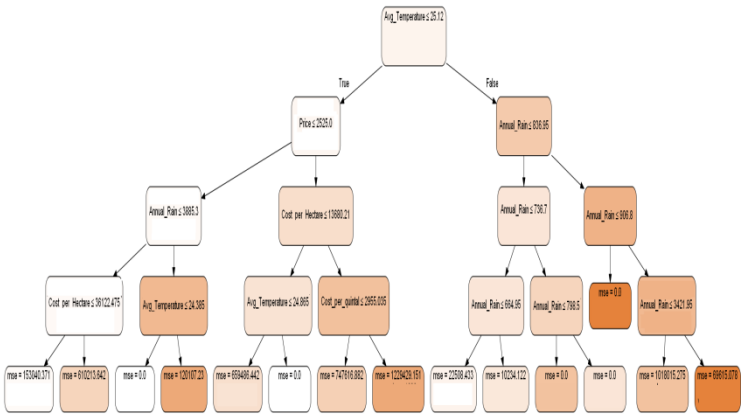


Fig. 5: Decision Tree Classifier on Agricultural Dataset

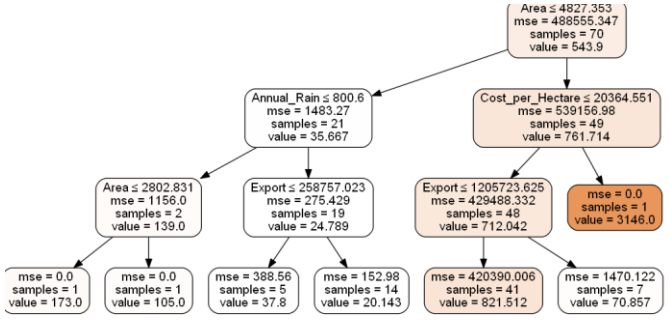


Fig. 6: Decision Tree Left Node on Agricultural Data

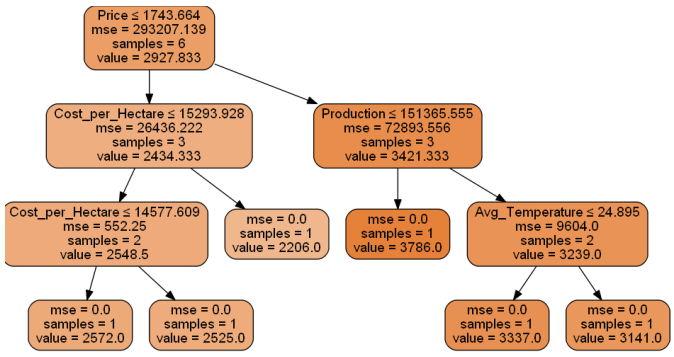


Fig. 7: Decision Tree Right Node on Agricultural Data

G.SVM

Support vector machines (SVMs) are supervised algorithm for both classification and regression. It is a predictive analysis data-classification algorithm that assigns new data elements to one of labeled classes. It is mostly used as a binary classifier; it assumes that the data in question contains two possible target values. Whereas in multiclass SVM, augments SVM to be used as classifier on a dataset that contains more than one class (grouping or category)

The main challenge faced by an analyst is to convert text/categorical data into numerical data and still make an algorithm/model to work. To convert categorical data into numerical value, Label-Encoder method is used. These encoders are a part of Scikit-learn library and are used to convert text or categorical data into numerical data which the model expects and perform better with. LabelEncoder encode the target labels with a value between 0 and n_classes-1 where n is the number of distinct labels. If a label repeats it assigns the same value to as assigned earlier. Once the label encoding is done, the training and test data is spliced using these encoded values.

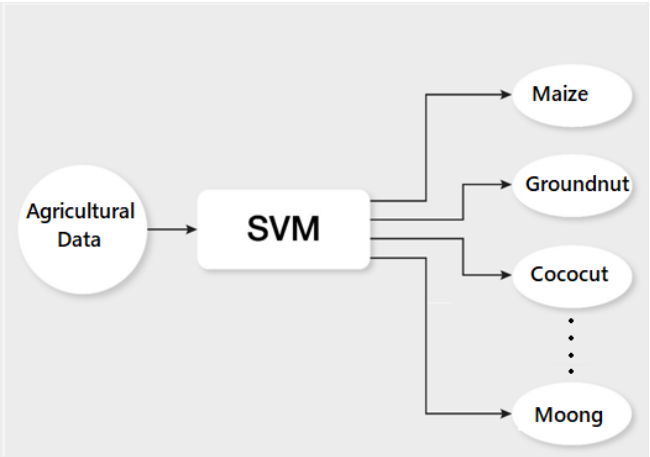


Fig. 8: Multiclass SVM Classifier on Agricultural Data

H. Multiclass Neural Classifier

A neural network is a supervised learning algorithm which means that we provide it the input data containing the independent variables and the output data that contains the dependent variable. For instance, in our example our independent variables are state, district, area, production, season, temperature, rainfall, export, price and suicides. The dependent variable is crop name.

In this model, a fully-connected network structure with four layers is used. Fully connected layers are defined using the Dense class. Here, the number of neurons or nodes in the layer as the first argument, and the activation function using the activation argument is specified. The rectified linear unit activation function referred to as ReLU on the first four layers and the Softmax function in the output layer. The rectified linear activation function reduces the gradient problem, hereby allowing models to learn faster and perform better. Final layer is the so-called Softmax activation function, which results in a multiclass probability distribution over our target classes. It returns a probability distribution over the target classes in a multiclass classification problem. The model expects rows of data with 13 variables (the input_dim=13 argument)

- The first hidden layer has 8 nodes and uses the ReLU activation function.
- The second hidden layer has 10 nodes and uses the ReLU activation function.
- The third hidden layer has 10 nodes and uses the ReLU activation function.
- The fourth hidden layer has 10 nodes and uses the ReLU activation function.
- The output layer has 123 nodes and uses the Softmax activation function.

The above model is optimized using an efficient stochastic gradient descent algorithm “Adam”. This is a popular version of gradient descent because it automatically tunes itself and gives good results in a wide range of problems. It optimizes algorithm for stochastic gradient descent for training deep learning models. AdaGrad keeps one learning factor

throughout the learning, whereas RMSProp uses the last best learning factor while training. Adam combines the both of the best properties of the AdaGrad and RMSProp algorithms to provide an optimization algorithm that can handle sparse gradients on noisy problems.

4. RESULTS

Upon analysis of the database we have found that recommended crops in the following states are:-

Table -3: Crop recommendations per states

ANDAMAN & NICOBAR ISLANDS	Coconut ,Rice, Banana , Arecanut, Sugarcane, Dry ginger ,Tapioca , Dry chillies, Sweet potato, Cashewnut, Turmeric, Black pepper, Other Kharif pulses
ANDHRA PRADESH	Coconut, Sugarcane, Rice, Maize, Cotton(lint), Banana, Groundnut, Mango, Gram, Dry chillies, Onion, Mesta, Urad
ARUNACHAL PRADESH	Rice, Maize, Dry ginger, Oilseeds total, Potato, Sugarcane, Small millets, Wheat, Dry chillies, Turmeric, Pulses total, Ginger, Coffee
ASSAM	Coconut, Rice, Sugarcane, Banana, Jute, Potato, Rapeseed &Mustard, Pineapple, Wheat, Orange, Arecanut, Papaya, Ginger
BIHAR	Sugarcane, Rice, Wheat, Maize, Jute, Potato, Mesta, Masoor, Banana, Moong, Khesari , Rapeseed &Mustard, Onion
CHANDIGARH	Wheat, Potato, Rice, Maize, Onion, Rapeseed &Mustard, Urad, Sunflower, Masoor, Moong, Gram, Kapas, Garlic
CHHATTISGARH	Rice, Khesari, Gram, Sugarcane, Maize, Wheat, Soyabean, Potato, Small millets, Groundnut, Urad, Onion, Arhar/Tur
DADRA AND NAGAR HAVELI	Sugarcane, Rice, Coconut, Other Rabi pulses , Ragi, Arhar/Tur, Wheat, Urad, Other Kharif pulses, Jowar, Banana, Small millets, Niger seed
GOA	Coconut, Rice, Sugarcane, Other Vegetables, Cashewnut, Other Fresh Fruits, Groundnut , Other Rabi pulses, Banana, Mango, Pineapple, Arecanut, Ragi
GUJARAT	Sugarcane, Cotton(lint), Wheat, Groundnut, Banana, Onion, Rice, Potato, Bajra, Castor seed, Maize, Rapeseed &Mustard, Arhar/Tur
HARYANA	Wheat, Sugarcane, Rice, Cotton(lint), Bajra, Rapeseed &Mustard, Potato, Barley, Gram, Onion, Other Vegetables, Maize, Jowar
HIMACHAL PRADESH	Maize, Wheat, Rice, Potato, Barley, Sugarcane, Peas & beans (Pulses), Dry ginger, Onion, Small millets, Urad, Rapeseed &Mustard, Other Kharif pulses
JAMMU & KASHMIR	Maize, Rice, Wheat, Rapeseed &Mustard, Potato, Bajra, Barley, Urad, Other Vegetables, Sesamum, Peas & beans (Pulses), Other Kharif pulses, Onion
JHARKHAND	Rice, Potato, Maize, Wheat, Ragi, Arhar/Tur, Onion, Sugarcane, Gram, Rapeseed &Mustard, Masoor, Barley, Cucumber
KARNATAKA	Sugarcane, Rice, Coconut, Maize, Jowar, Ragi, Arecanut, Cotton(lint), Onion, Banana, Groundnut, Dry ginger, Gram
KERALA	Coconut, Tapioca, Rice, Banana, Sugarcane, Mango, Arecanut, Rubber, Black pepper, Cashewnut, Dry ginger, Papaya, Pineapple
MADHYA PRADESH	Wheat, Soyabean, Gram, Sugarcane, Rice, Maize, Onion, Potato, Rapeseed &Mustard, Cotton(lint), Jowar, Banana, Bajra
MAHARASHTRA	Sugarcane, Cotton(lint), Jowar, Soyabean, Rice, Maize, Wheat, Bajra, Gram, Arhar/Tur, Groundnut, Banana, Moong
MANIPUR	Rice, Banana, Pineapple, Potato, Sugarcane, Maize, Other Fresh Fruits, Dry chillies, Cabbage, Dry ginger, Papaya, Peas & beans (Pulses), Cauliflower
MEGHALAYA	Rice, Potato, Banana, Dry ginger, Jute, Maize, Tapioca, Mesta, Arecanut, Sweet potato, Pineapple, Turmeric, Cotton(lint)
MIZORAM	Rice, Maize, Sugarcane, Other Kharif pulses, Other Rabi pulses, Potato, Soyabean, Sesamum, other oilseeds, Rapeseed &Mustard Tapioca, Cotton(lint), Tobacco
NAGALAND	Rice, Sugarcane, Maize, Potato, Soyabean, Rapeseed &Mustard Tapioca, Small millets, Peas & beans (Pulses), Ginger, Wheat, Dry ginger, Niger seed
ORISSA	Rice, Sugarcane, Maize, Sweet potato, Groundnut, Potato, Onion, Moong, Ragi, Urad, Jute, Arhar/Tur, Dry chillies
PUDUCHERRY	Coconut, Sugarcane, Rice, Banana, Tapioca, Mango, Groundnut, Cotton(lint), Urad, Brinjal, Moong, Bajra, Ragi
PUNJAB	Wheat, Rice, Sugarcane, Cotton(lint), Maize, Barley, Rapeseed &Mustard, Guar, seed, Moong, Gram, Arhar/Tur, Groundnut, Peas & beans (Pulses)
RAJASTHAN	Wheat, Bajra, Rapeseed &Mustard, Maize, Soyabean, Gram, Guar seed, Barley, Sugarcane, Cotton(lint), Groundnut, Onion, Moth
SIKKIM	Maize, Rice, Other Vegetables, Wheat, Small millets, Rapeseed &Mustard, Potato, Soyabean, Other Kharif pulses, Urad Other, Cereals & Millets, Other Fresh Fruits, Barley
TAMIL NADU	Coconut, Sugarcane, Rice, Tapioca, Banana, Groundnut, Maize, Jowar, Cotton(lint), Onion, Ragi, Urad, Turmeric
TELANGANA	Coconut, Rice, Sugarcane, Cotton(lint), Maize, Groundnut, Onion, Turmeric, Jowar, Dry chillies, Soyabean, Arhar/Tur, Gram
TRIPURA	Rice, Sugarcane, Potato, Jute & mesta, Mesta, Jute, Maize, Wheat, Cotton(lint), Rapeseed &Mustard, Other Kharif pulses, Oilseeds total, Arhar/Tur
UTTAR PRADESH	Sugarcane, Wheat, Rice, Potato, Bajra, Maize, Rapeseed &Mustard, Gram, Barley, Masoor, Peas & beans (Pulses), Arhar/Tur,Onion

UTTARAKHAND	Sugarcane, Wheat, Rice, Total foodgrain, Potato, Ragi, Small millets, Maize, Other Cereals & Millets, Onion, Barley, Dry ginger, Soyabean
WEST BENGAL	Coconut, Rice, Potato, Jute, Sugarcane, Wheat, Rapeseed &Mustard, Maize, Sesamum, Groundnut, Mesta, Masoor, Urad

Implementing a User interface to integrate the backend operation for smoothening of work efficiency-

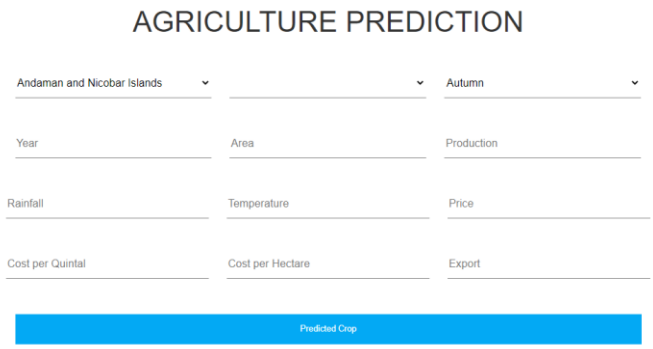


Fig. 9: User Interface of Agricultural Crop Recommendation.

The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data. In the example below 5 different algorithms are compared:

1. K-Nearest Neighbors
2. Decision Tree
3. Naïve Bayesian
4. Multi class SVM
5. Multiclass Neural Network

The 10-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithms is evaluated in precisely the same way

Input data: To test the efficiency of the system, a sample data is selected from the test dataset which had recorded large number of suicide count. As per the data, the actual output was "Cotton"

Table -4: Input data from dataset for Crop Prediction

STATES	MAHARASHTRA
DISTRICT	DHULE
Season	Kharif
YEAR	2014
Area	209900
Production	258200
Annual Rain	838
Avg_Temperature	24.73
Price	7800
Cost_per_Hectare	3500
Cost_per_quintal	3000
Suicides	0
Export	4000

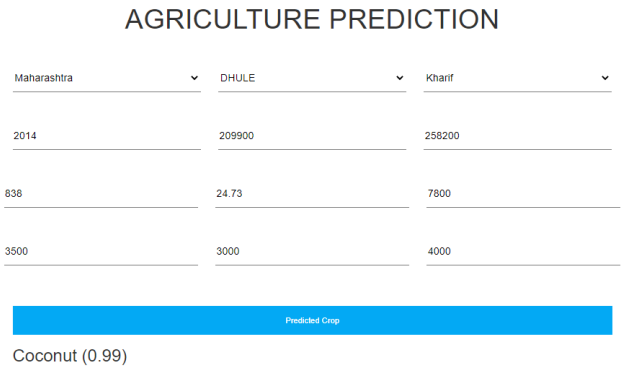


Fig. 10: Predicted Crop of the input data

Based on the above input from the test data, the crop recommended from the system is "Coconut", but in the test data, the crop grown that year was "Cotton". This states that if cotton was not grown, wrt to the given input parameters, the suicides count would have been less.

Thus, the system is reliable enough to provide the efficient crop prediction to properly utilize the agricultural amenities. Similarly the system is tested on the new dataset

Table -5: Input data for Crop Prediction

STATES	MAHARASHTRA
DISTRICT	DHULE
Season	Kharif
YEAR	2014
Area	209900
Production	258200
Annual Rain	838
Avg_Temperature	24.73
Price	7800
Cost_per_Hectare	3500
Cost_per_quintal	3000
Suicides	0
Export	4000

The data was tested on the above 5 algorithm. Each algorithm is tested for efficiency with the test data set. The comparative study of algorithm is given below-

Table -6: Comparative analysis of algorithm

Algorithm Name	Parameters	Predicted Crop	Accuracy
KNN Classifier	n_neighbors=10 metric='minkowski'	Maize	38.50%
Decision Tree	criterion='gini'	Maize	81.84%
Naïve Bayesian	GaussianNB	Moong	11.22%
Multiclass SVM	LinearSVC	Moong	8%
Neural Network	The first hidden layer 8 nodes and relu activation function. The second, third and fourth hidden layer 10 nodes and relu activation function. The output layer 123 node and softmax activation function. Optimiser – Adam	Maize	99.19%

4. CONCLUSION AND FUTURE WORK

Pre-processing of data is carried out using the training and testing sets of agricultural data to improve the performance of algorithm for determining the crop type. The agriculture depends on several factors like the area environmental conditions, crops prices, costs and its proper study is hugely needed. In this project, Knn, naïve Bayesian, Decision Tree, SVM and Neural network is used. The analysis and comparison of implemented algorithm has been carried out to show which algorithm is more efficient. Out of all these algorithms, Neural Network gave the better accuracy 99%. Neural network makes crop prediction more insightful to comprehend, weighing on how each of the factors affect.

With the use of this system, we can now recommend new crops that can be grown in places which have the suitable weather and economy conditions. If similar analytical study is done and followed, it can prevent wastage of land, increase in production, lesser number of farmer suicides which will help us to meet the demands of the people, harmony of farmers and also boost our economy.

The study can be extended to large data sets with large number of attributes. Other features like Soil data, Weather data, and fertilizers can also be taken into account for much better results. We can also improve the prediction and clustering by using some other algorithms. Apart from these, predictions of various other factors can be made, like crop production in the following year. If we will be able to predict the factors affecting the crop production, then we can plan the cultivation in a better way.

REFERENCES

1. Rupanjali D. Baruah, R.M. Bhagat, Sudipta Roy, L.N. Sethi "Use of Data Mining Technique for Prediction of tea yield in the face of Climate Change of Assam, India ", 2016 International Conference on Information Technology
2. Niketa Gandhi, Leisa J. Armstrong "Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India using data mining Techniques", 2016 IEEE International Conference on Advances in Computer Applications (ICACA)
3. Suvidha Jambekar, Shikha Nema , Zia Saquib , "Prediction of Crop Production in India Using Data Mining Techniques", 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).
4. Kuljit kaur, Kalnwalpreet Singh Attwal "Effect of Temperature and Rainfall on Paddy Yield using data mining", 2017 7th international conference on Cloud Computing, Data Science and Engineering- Confluence.
5. Mukesh kumar , Prof.Mayura nagar , "Big Data analytics in agriculture and distribution channel", Proceedings of the IEEE 2017 International Conference on Computing Methodologies and Communication (ICCMC)
6. Jharna Majumdar* , Sneha Naraseeyappa and Shilpa Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data", Springer Open 2017
7. https://data.gov.in/catalog/district-wise-season-wise-crop-productionstatistics?filters%5Bfield_catalog_reference%5D=87631&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc
8. https://towardsdatascience.com/6-ways-the-agricultural-industry-is-benefiting-from-datascientists-b778d83f61db
9. https://data.gov.in/resources/seasonal-and-annual-mean-temp-series-india-1901-2017
10. http://agricoop.nic.in/sites/default/files/MSP_2019-20%20%28English%29.pdf
11. http://agricoop.nic.in/recentinitiatives/minimum-support-pricesmsps-kharif-crops-2019- 20-season
12. https://vikaspedia.in/agriculture/market-information/minimum-support-price
13. https://data.gov.in/resources/state-ut-wise-cost-estimate-some-principal-crops-five-major-producing-states-2008-09-2013
14. https://data.gov.in/catalog/rainfall-india?filters%5Bfield_catalog_reference%5D=1090541&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc
15. https://data.gov.in/resources/subdivision-wise-rainfall-and-its-departure-1901-2015
16. https://data.gov.in/resources/details-farmers-suicides-2010-2014-ministry-agriculture-farmers-welfare
17. https://agriexchange.apeda.gov.in/
18. https://cacp.dacnet.nic.in/#