International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Crop Selection and its Yield Prediction using Machine Learning Techniques

Aakanksha Attarde[1], Srushti Gangapurkar[2], Mansi Gandhi[3]

[1]*Student, Dept. of Computer Science and Technology, SNDT University, Maharashtra, India*
[2]*Student, Dept. of Computer Science and Technology, SNDT University, Maharashtra, India*
[3]*Student, Dept. of Information and Technology, SNDT University, Maharashtra, India*

Abstract - *There are various factors to be considered while selecting the crop such as crop that would give the maximum yield, crop that would give the maximum profit, etc. Hence it becomes a difficult task for the farmer to decide which crop would be the most suitable one considering all the factors that affect its production. Machine Learning techniques proves beneficial in this situation. Techniques such as SVM, KNN has been previously applied to predict crop. This paper proposes a predictive system using decision tree algorithm for selecting the suitable crop and also predict its yield by the Random Forest Algorithm. The prediction is made based on various climatic factors and historical data which also includes the pecuniary factors.*

Key Words: *Decision Tree, Random Forest, crop selection, yield prediction*

1. INTRODUCTION

Over half of the Indian population depends on the agriculture industry. Demand for food is growing on the contrary, the supply side faces constraints in land and farming inputs.

The farmer makes his decision regarding which crop to grow on his land, usually based on some intuition and factors such as making larger profits within a shorter period of time, with the lack of awareness about the requirement in the market. A wrong decision that is taken on the farmer's side could put a much bigger pressure on the financial condition of his family resulting in severe loss and in turn creating an imbalanced crop production all around the nation [7]. Selecting a wrong crop for cultivation may lead to loss in achieving a high yield rate and also simultaneously leads to shortage of food. These difficulties implies the need of smart farming which can be achieved with various machine learning algorithms [2]. Researchers in agriculture have been testing numerous forecasting methodologies to identify the most suitable crop for specific areas of land based on historical data [9].

In general, agro-climatic input parameters such as soil properties, rainfall, and temperature influence crop production. Predicting suitable crops for cultivation is an essential part of agriculture, with machine learning algorithms playing a major role in such prediction in recent years. In this scenario we have designed a recommendation system that predicts the type of crop that can be grown in a particular land and also predict its total yield.

2. LITERATURE REVIEW

Girish L. et al aims to increase the net yield rate of the crop, based on rainfall. In this paper an application is developed that suggests the crops for farmers in Tumakuru district based on predictive analysis so as to help the farmer in selecting the crops based on predicted rainfall values and crop price values. Methods such as Linear Regression, Support Vector Machine, K-Nearest Neighbors and Decision Tree were experimented for prediction purpose and it was observed that SVM gave highest efficiency. This project can be extended to predicted crop disease using deep learning technology [5].

Mayank Champaneri, Chaitanya Chandvidkar, Darpan Chachpara and Mansing Rathod built an interactive prediction system in order to predict the crop yield before cultivation based on the climatic data provided by the user. They used the Random Forest technique. The developed web page was user friendly and the predictions showed an accuracy of above 75 per cent [3].

Aksheya Suresh, K. Monisha, R. Pavithra and B. Marish Hariswamy aims to propose and implement a system that determines the suitable crop to be cultivated and its appropriate yield prediction. The model uses Decision Tree algorithm to classify crops based on location and season and further predicts the production of this selected crop based on available area and past data by Linear Regression algorithm.The above developed model gave an accuracy of 88.7 per cent when tested. The author also mentions that the model could be enhanced by using it with a mobile application which would make it even more flexible for farmer's use [4].

P.Priya , U.Muthaiah and M.Balamurugan focuses on building the model that predicts the yield of the crop based on the existing real data of Tamil Nadu using the Random Forest Algorithm. The dataset comprised of rainfall, perception, production, temperature to construct Random Forest. Their proposed work used tools such as RStudio due to its different advantages such as platform independence and ease of documenting and updating analysis.The proposer

here intended to make right decision for the farmer for the right crop [8].

In their work Keerthan Kumar T G , Shubha C and Sushma S A has proposed a machine learning system that would suggest a best suitable crop based on the land type i.e. various soil nutrients (both macro and micro). Various classification algorithms such as Linear Regression, Multi-Variate, Support Vector Machine, and Random Forest Classification and Decision tree were studied and Random Forest was chosen based on accuracy and error analysis. Soil grading was done in the first module of the system based on various soil nutrients as feature variables. The second module was the crop recommendation module dependent on the type of the soil [6].

3. PROPOSED SYSTEM FOR CROP SELECTION AND ITS YIELD PREDICTION

In this research, the developed is a system that is targeted towards people who are looking forward to starting farming but have no prior knowledge regarding farming. This system will also assist the existing farmers to make some correct decisions and do further planning regarding the cultivation of crop and market strategies. By using predictive machine learning algorithms, a system is built that comprises 2 component models. One is a crop selection model which is constructed using the Decision Tree technique. This model takes location, season, details of past crop produced, rainfall and the market prices of various crops as features. Based on these inputs the model suggests a crop to the user which would be the best suitable one to cultivate in that particular location. This crop is passed to the next model as input in order to predict its total yield. The second model is the yield prediction model which is built using the Random Forest algorithm. This model has an area of production and different soil nutrients deficiencies as its features. The algorithm then determines and predicts the total yield of the crop which was generated as an output in the previous model. This prediction is performed depending on the area that is available with the farmer and the soil properties of his field. Later the whole system is deployed using flask technology. An user friendly interface is made for the target users in order to interact and benefit from it. The proposed system minimizes the efforts taken to gather information about Fig. 1. Proposed Architecture cultivating crops. The following is a visual representation of the system and its flow of working.

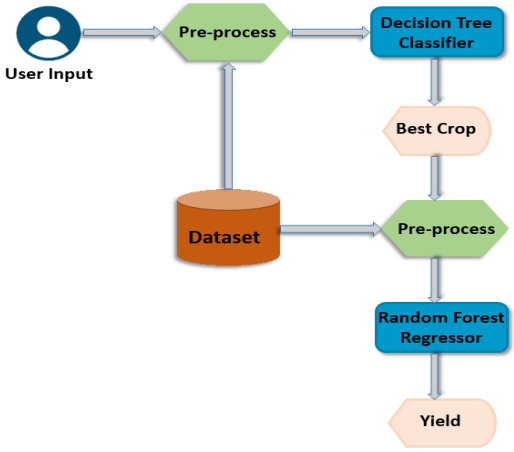

Figure 1: Proposed Architecture

3.1 DATASET

In order to perform the crop selection and its yield prediction which is being done for crops at district level, the main goal is to find and acquire a dataset. Since India is a country with a varied climate which changes at a frequent interval, data is considered at the district level for different states of India. Every location in the data is recognized by the respective state name and district name. Details of the data was collected from various sources such as ICRISAT-International Crops Research Institute for the Semi-Arid Tropics, Kaggle and National Portal of India. The dataset includes particulars about crop production of the past 4-5 years and it is categorized based on various parameters which are State Name, District Name, Crop Name, Area, Production, Rainfall, MSP and Soil Deficiencies.

	State	District	Year	Season	Crop	Area	Production	MSP	Rainfall	SD_N	SD_OC	SD_P	SD_K
0	Andhra Pradesh	ANANTAPUR	2010	Kharif	Arhar/Tur	66013.0	11156.0	3000	755.9	92.6200	54.2100	6.1300	15.8500
1	Andhra Pradesh	ANANTAPUR	2010	Kharif	Bajra	2010.0	1465.0	880	755.9	92.6200	54.2100	6.1300	15.8500
2	Andhra Pradesh	ANANTAPUR	2010	Kharif	Cotton(lint)	4338.0	2977.0	2500	755.9	92.6200	54.2100	6.1300	15.8500
3	Andhra Pradesh	ANANTAPUR	2010	Kharif	Groundnut	814077.0	446928.0	2300	755.9	92.6200	54.2100	6.1300	15.8500
4	Andhra Pradesh	ANANTAPUR	2010	Kharif	Jowar	2606.0	2069.0	900	755.9	92.6200	54.2100	6.1300	15.8500
...	...	...	...	...	...	...	...	...	...	...	...	...	...
12228	West Bengal	PURULIA	2014	Rabi	Gram	198.0	203.0	3175	1026.7	0.5027	0.5029	0.1125	0.0259
12229	West Bengal	PURULIA	2014	Rabi	Masoor	31.0	19.0	3075	1026.7	0.5027	0.5029	0.1125	0.0259
12230	West Bengal	PURULIA	2014	Rabi	Rapeseed &Mustard	1885.0	1508.0	3100	1026.7	0.5027	0.5029	0.1125	0.0259
12231	West Bengal	PURULIA	2014	Rabi	Safflower	54.0	37.0	3050	1026.7	0.5027	0.5029	0.1125	0.0259
12232	West Bengal	PURULIA	2014	Rabi	Wheat	1622.0	3663.0	1450	1026.7	0.5027	0.5029	0.1125	0.0259

Figure 2: Dataset

3.2 PREPROCESSING

Preprocessing is vital for data in any Machine learning method as the accuracy of the Machine Learning model highly depends on how clean data is fed to the model. Preprocessing step transforms the raw data into data that is suitable for the Machine Learning model.

Steps involved in the Preprocessing are:

- Removing rows if wherever any column contains NULL value
- OneHotEncoding: Since ML models do not accept categorical data and only understand numbers, thus OneHotEncoding is performed.
- StandardScaler method is used in order to get data with different ranges in the same scale.

• In order to make a prediction there must be attributes/features and target/label.
• In order to know how good the model works we need to keep aside part of data which is unseen to the model.

3.3. PREDICTION MODELS

1) Crop Selection Model:

Support Vector Machine: The support vector machine is a supervised machine learning algorithm which is used both for classification and regression but mostly for classification problems. The objective of the support vector machine algorithm is to find an optimal hyper plane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. The data points that are closest to the hyper-plane are called support vectors. The decision plane that divides the data points based on the target class is called a hyperplane. A margin is a gap between the two lines on the closest class points. This is calculated as the perpendicular distance from the line to support vectors or closest points. If the margin is larger in between the classes, then it is considered a good margin, a smaller margin is a bad margin. SVM uses a technique called the kernel trick, it converts non separable problems to separable problems by adding more dimension to it. Some of the kernel are linear kernel, polynomial kernel and Radial Basis Function Kernel.

Decision Tree: It is a tree-structured supervised learning algorithm used for classification and regression. Decision tree simply aims to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. In the decision tree structure, the internal nodes represent the features of a data set, branches represent the decision rules and each leaf node represents the outcome.

Values for the feature are preferred to be in categorical format, if not then they are converted to discrete before building the model. There are different feature selection measures to identify the feature which can be considered for the splitting at each levels. Entropy is calculated to measure disorder/impurity. Value of entropy ranges between 0-1. Impurity/ dis□order is lowest for extreme values and is highest for 0.5. Information gain or IG is a statistical property that measures how well a given attribute separates the training examples according to their label/target class. The feature with the highest information gain is picked as the splitting feature. Gini Index/Gini impurity, evaluates the probability of a specific feature that is classified incorrectly when selected randomly. Overfitting of the tree can be handled by tuning the hyper parameter. The max depth parameter can create a complex model if the tree grows deeper and captures more information about the training data which could lead to overfitting. On the other hand, very low depth could lead to underfitting. Min samples leaf parameter is the minimum number of samples required to be at a leaf node. The parameter min sample split lets specify 98the minimum number of samples required to split an internal node.

2) Yield Prediction Model:

After selecting the best suitable crop in the previous model, now this selected crop goes as an input to the next model that is the Yield Prediction model. Now depending upon the inputs about the area and soil nutrients this model predicts the total yield of that crop for that area.

Random Forest: Random Forest is an ensemble supervised machine learning technique. In ensemble it comes under bagging i.e. Bootstrap Aggregating. Random forest makes a prediction based on pre□dictions made by multiple decision trees in parallel form. And later combines them all to get a more precise and consistent result. The advantage of Random forest comes as it allows each individual tree to randomly sample from the dataset with replacement, resulting in different trees. Same can be done for the features which is known as feature bagging. And this introduces randomness to the algorithm. Number of trees to be created can be defined under the n estimators parameter of the algorithm. Fine tuning is done using the min samplee split, max depth, Min samples leaf parameters

3.4. DEPLOYMENT

A Web Interface is made in order for the targeted users to interact with the system. So the system is put to use by deploying it using the Flask framework. This is the interactive page where the user enters different inputs depending on which the system will suggest a crop and predict its total yield.

4. RESULT AND ANALYSIS

Among the SVM and DecisionTreeClassifier algorithm, the SVM gave the final accuracy of 44.67% and the latter gave that of 86.92. It was observed that the SVM did not improve on increasing the dataset, instead the model kept deteriorating. SVM does not perform very well, when the dataset has more noise, (i.e. the target classes are overlapping) [1] such as there could be situations where the resultant crop is same for 2 or more different sets of inputs. The RandomForestRegressor predicts the yield with an accuracy of 92.02%.

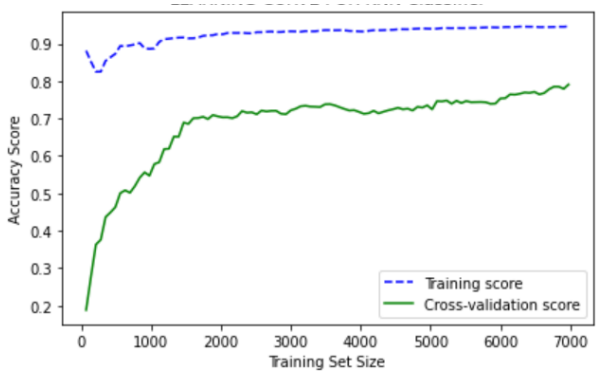

Figure 3: Learning curve for Crop Selection model

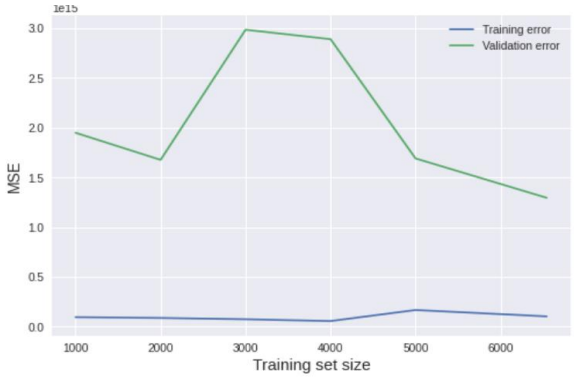

Figure 4: Learning curve for Yield Prediction model

5. CONCLUSION

In this paper the proposed is a Machine learning based Prediction of the most suitable crop to be cultivated and its yield. The prediction is based on factors like season, rainfall and land properties. Since the dataset is of non-parametric nature hence Decision Tree and Random Forest are used which are non-parametric effective machine learning modeling techniques for classification and regression problems. Decision tree algorithm have the capability of capturing descriptive decision making knowledge from the supplied data. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction. This adds additional randomness to the model. This system can further be raised up by developing its mobile application where in it can access the current location as the user input.

REFERENCES

[1] Understanding support vector machine (svm) algorithm. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/.
[2] R.Pavithra B.Marish Hariswamy Aksheya Suresh, K.Monisha. Crop selection and its yield prediction. International Journal of Recent Technology and Engineering, 8, 2020.
[3] Mayank Champaneri, Darpan Chachpara, Chaitanya Chandvidkar, and Mansing Rathod. Crop yield prediction using machine learning.
[4] Tatikonda Gayathri, Premamayudu Bulla, S Nyamathulla, and Naresh Alapati. Machine learning: Research approaches and opportunities. Journal of Critical Reviews, 7(12):2298–2306, 2020.
[5] L Girish. Crop yield and rainfall prediction in tumakuru district using machine learning.
[6] TG Keerthan Kumar, C Shubha, and SA Sushma. Random forest algorithm for soil fertility prediction and grading us□ing machine learning. International Journal of Innovative Technology and Exploring Engineering (IJITEE).
[7] Merin Mary Saji Lisha Varghese Er. Jinu Thomas Kevin Tom Thomas, Varsha S. Crop prediction using machine learning. International Journal of Future Generation Communication and Networking, 13(3):1896–1901, 2020.
[8] P Priya, U Muthaiah, and M Balamurugan. International journal of engineering sciences & research technology predicting yield of the crop using machine learning algorithm.
[9] A Suruliandi, G Mariammal, and SP Raja. Crop prediction based on soil and environmental characteristics using feature selection techniques. Mathematical and Computer Modelling of Dynamical Systems, 27(1):117–140, 2021