European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 06, 2020

Crop Yield Prediction Using Linear Support Vector Machine

N.Manjunathan[1], P.Rajesh[2], E. Thangadurai[3], A. Suresh[4]

[1,2]Assistant Professor, Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of Science and Technology, Chennai, Tamilnadu, India
[3]Assistant Professor, Department of Information Technology, Vivekanandha College of Engineering for Women (Autonomous), Namakkal, Elaiyampalayam, Tamil Nadu. kapildurai@gmail.com
[4]Associate Professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. prisu6esh@yahoo.com

**Corresponding author Email: nmanjunathan24@gmail.com*

***Abstract:** The main objective of this proposal is to build a Machine Learning model that can accurately predict the rice crop yield prediction. Over 97% of the population in India depends on rice for food and is the second-highest in overall agriculture productions. But during recent years the farmers had suffered a huge loss in productions due to unexpected weather change, no knowledge about soil, underground water, area supported crops. As crop production depends on a lot of these factors, it is important to follow these factors for successful crop yield. So we are proposing a model that can accurately predict the crop yield. The algorithms used were the Support Vector Machine (SVM). SVM is used to classify the crop based on the factors of the area, season. And we are also implementing a Web Application that enables the users to interact with the ML model and make their prediction with their given inputs. The proposed system uses the Weka tool for creating the machine learning algorithms and Html, CSS, JavaScript for developing the web application.*
***Keywords:** Support Vector Machine (SVM), Weka, Rice crop yield prediction, Web application.*

1. Introduction

According to the records of the previous year 2018 and 2019, there are approx. 145 million landholdings in India. We may assume that India has about 130 million farmers. In a country like India which has increased demand for food due to the increasing population in the country. The most disadvantaged situation is that farmers who have access to irrigation are better placed but those who are in rain-fed and drone-prone areas are most vulnerable. A single crop failure due to flood, lack of soil fertility, drought, climatic changes, lack of underground water and some other factors may destroy the crop and this affects the farmers. There is no commodity-

based farming in India till now. While in other countries the organizations advise farmers to grow specific crops according to the locality of the area and some other factors. Every farmer who produces the crop always tries to know how much yield will get from his expectations.

So we want to help farmers by creating a machine learning model that predicts the crop yield. Although there are models that help in yielding they are hardware-based which is expensive and difficult to maintain. We can also increase the yield of the crop by systematic study of different methods like planting, fertilization, irrigation and some other methods which help to optimize the production of crop but most of the farmers in India are illiterate to study them and follow the optimizing methods and follow them. So we came across an idea to implement a machine learning model which calculates the history of the field and suggest whether the crop should be planted in that area or not, basically it benefits the farmers and saving them lot of trouble.

Although our proposed system is limited only to a certain crop which is Rice, cause most of the farmers in India rely on farming Rice. The proposed system predicts the crop yield accurately using SVM to produce accurate results and helps farmer to choose the right crop according to the area and climatic conditions because in prediction process of the system we include the data of soil nitrogen, underground water, temperature, rainfall which may produce the accurate results in recommending farmer to invest in farming that crop or not.

2. Literature Survey

[1] Proposed an approach using Markov Chain Theory, this model focused on corn and cotton yield and the prediction of diseases and pests related to these crops. This method gives better results than the regression model. Predicting diseases and pests is done using the sensor-based approach. The Naive Bayes Kernel algorithm is used to compare the patterns from the crop data. A sample dataset of crops is given as a training set to Naive Bayes Kernel algorithm and the raw dataset of soil sample and temperature. The model gives the pattern comparison of both datasets. If the pattern is consistent then there is no disease and crop growth will be good, If it is inconsistent then the disease is predicted.

In [13], a wireless sensor is implemented for sensing soil moisture and data related to soil. This wireless sensor node can be used in applications like in-field soil moisture collection and other kinds of site data collection. A machine learning model is built to predict moisture for n number of days, this model also predicts temperature, humidity, wind speed, solar radiation, precipitation and soil temperature. SVM and RVM Machine Learning techniques are used in this model. These parameters were used in crop yield predicting models.

[12] proposed a classification and clustering are used which the two main techniques of data are mining. The future data is classified and predicted using classification and prediction techniques which are two ways of analyzing data. The goal is to get more accuracy to test training of a classification algorithm. Here supervised and unsupervised both were used. The classification techniques used were Naive Bayes, Random Forest, Artificial Neural Network, and Decision Tree. Among these classifiers the one with highest accuracy is used to develop the model. In this model ANN is used for better classification and higher accuracy. The soil and weather are the two parameters in the dataset using which the model is trained and the

yield is predicted. This model provides accurate predictions of crop suitable for particular region.

In [14] the model used soil dataset, rainfall dataset, and previous year's prediction data to create the model which is designed using SVM, KNN and Decision Tree algorithms. But only Sugarcane crop data is taken and which only of Karnataka state data this model not is useful in other places or for other crops.

3. Proposed Framework

The proposed model will mainly focus on crop production based on four factors and one Machine Learning algorithm called SVM (support vector machine). SVM is used to classify whether rice can grow in that area based on the data from soil, temperature, underground water and rainfall. And also implementing a web application with HTML, CSS, and JavaScript. The web application can be used to interact with the Machine Learning model and by providing inputs we can get the prediction output. The application also uses the weather API from yahoo to get the current temp at that location, which will be one of the factors.

Support Vector Machine (SVM)

SVM (Support Vector Machine) is a machine learning algorithm that comes under the supervised category and is used for binary classifications problems. The objective of this algorithm is to plot a hyper plane in an N-dimensional space, where N is the number of features that are going to be in a dataset, that distinctly classify the data points.

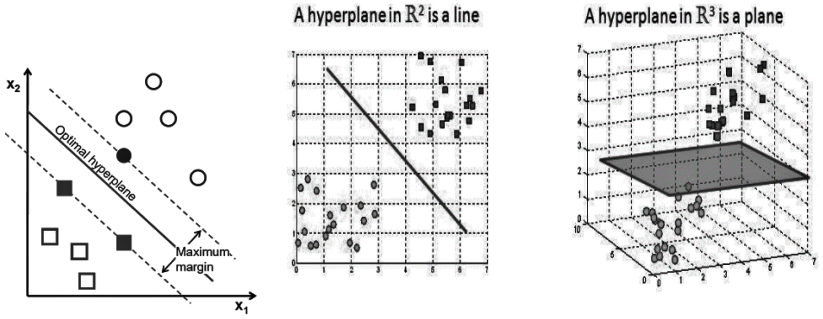

Figure 1: Maximum Margin and Hyper planes

There can be any number of hyper planes plotted, but the algorithm's main target is finding the plane that has a Maximum Margin i.e the maximum distance between data points of the features being plotted. The more the distance more accurate will be the classification. As shown in fig 1, that the data points are far from all the other points from the Optimal hyper plane, making it as a Maximum margin.

Cost Function

The main objective is to maximize the margin. So hinge loss is used to do that, the cost is zero if the predicted value and actual value are of the same sign, if they are not we can calculate hinge loss value. And adding a regularization parameter will balance the margin maximization and loss.

$$J(\theta) = C[\sum_{i=1}^{m} y^{(i)}\, Cost_1(\theta^T\,(x^{(i)}) + (1 - y^{(i)})Cost_0(\theta^T(x^{(i)})] + \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

In this proposed model we used Linear SVM which suited our kind of prediction. In Linear SVM the loss function is as similar to that of Logistic Regression. The x-axis here is the output i.e $\theta^T x$. Just like the Sigmoid function, the hypothesis used here is when $\theta^T x >= 0$, predicts 1, otherwise, predicts 0.

Data Set

A proper dataset is required for proper training of a model. There are four data factors used in this proposal which are Nitrogen percentage in soil, the annual rainfall in mm, the annual underground water recharge and the annual temperature. The dataset is focused on the Indian State Tamil Nadu, which has over 32 districts. So we have collected the data from the year 1997 to 2013 which contains all the annual values of the data of the factors mentioned above. The rainfall data is obtained from the Indian Gov website, the underground data and the soil nitrogen percentage data is obtained from self-research and the remaining data is obtained from Kaggle which includes the annual production value of the wheat crop from the year 1997 to 2013 and the area in sq. ft. The target variable in this dataset is "output" which has two values which are 0 and 1. 0 represents whether rice farming in that area is suitable or not and vice versa. This dataset contains a total of 12 features and 500 observations as shown in Fig 2

State_Name	District_Name	Crop_Year	Season	Crop	Area	Production	Rainfall	Soil (N)	Underground_Water	Temp(C)	Output
Tamil Nadu	ARIYALUR	2008	Kharif	Rice	24574	70854	691.3	223	41306	23	{0}
Tamil Nadu	ARIYALUR	2009	Kharif	Rice	25978	80462	404.7	233	44737	20	{0}
Tamil Nadu	ARIYALUR	2010	Kharif	Rice	25211	97869	557.6	401	42802	38	{1}
Tamil Nadu	ARIYALUR	2011	Kharif	Rice	24097	97257	980.1	192	43035	43	{1}

Figure 2: Data set

4. Experimental Setup

Exploratory Data Analysis

The dataset is obtained by combining three datasets as shown in Fig 2. The important part before building a model is to analyze the data first and extract the features which are causing the output target variable. The steps include are filling null values, dropping the features which are not necessary, visualizing the data, normalization.

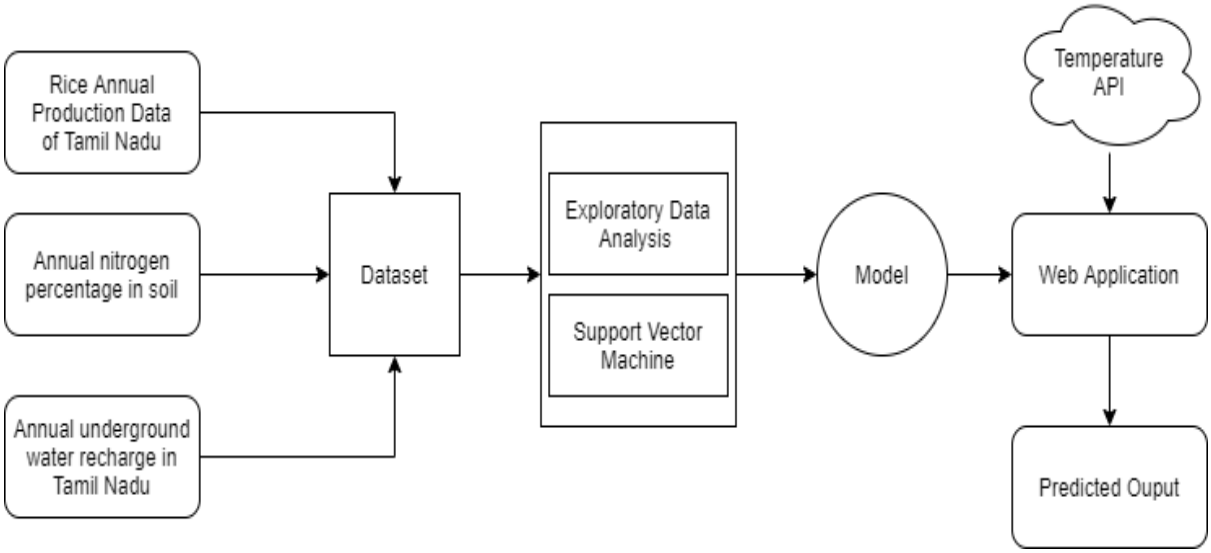

Figure 3: Architecture Diagram

Splitting data into Training and Test

The dataset is split into two Training and test. We can also select the proposition of their division metric. In this model, the training set is 70% of the dataset and 30% is the test set. The training of a model also depends on this proportion as more training of data more chances of better accuracy.

Training the Model

The next step is to train the model using our preferred algorithm. We choose trial and error method and wanted to select the best algorithm which gives us better accuracy, first we selected Non-Linear SVM which got us the precision of 0.62 i.e 62% accuracy, which is very low and not suitable for our dataset. Then we choose Linear-SVM which got us the precision value of 0.93 i.e 93% accuracy.

Parameter Tuning

To increase the accuracy or the precision score we can tune the parameters which define the training model. We got 93% of accuracy with Linear-SVM with a cross-validation score of 10 (k=10). So by changing the value of the cross-validation score to 5 (k=5) got us better results with the accuracy percentage of 96.3.

K-fold cross-validation

It is a type resampling technique used to evaluate or estimate the Machine Learning model skill or accuracy on unseen data. It divides the data into k groups and trains each group separately and the precision score from all the groups are averaged to the final precision or accuracy value.

Prediction

From the finalized model we can start predicting our values. We need to provide five parameters, rainfall in mm, average underground water recharge, nitrogen in soil, area in sq.ft and temperature. The result will be of two types one is 0 and the other is 1. If 0 is the output then it is not recommended to grow rice in that area. If the output is 1 then it is recommended to grow rice in that area.

5. Results and Discussion

From the EDA (Exploratory Data Analysis) we found that average temperature, nitrogen in soil and season had the most effect on rice production. And the only season recommended is Kharif. Our Machine Learning model which is trained with Support Vector Machine has managed to obtain an accuracy percentage of 96.5% as shown in Fig 4. And a root means square error (RMSE) of 0.18. Our web application takes the input and displays the result whether at the current location growing rice is recommended or not. This can be used by many people and it is still under development and we are planning to support many other languages as possible for other state people to understand and also implement other state

```
Correctly Classified Instances         481               96.5863 %
Incorrectly Classified Instances        17                3.4137 %
Kappa statistic                          0.9317
Mean absolute error                      0.0341
Root mean squared error                  0.1848
Relative absolute error                  6.8312 %
Root relative squared error             36.9626 %
Total Number of Instances              498
Ignored Class Unknown Instances                 1

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.967     0.035     0.963      0.967     0.965      0.932    0.964     0.944     {0}
               0.965     0.033     0.969      0.965     0.967      0.932    0.966     0.952     {1}
Weighted Avg.  0.966     0.034     0.966      0.966     0.966      0.932    0.965     0.948
```

Figure 4: Validation Results

6. CONCLUSION

There is so much to explore in machine learning yet, as there can be new algorithms, new techniques in the future. Our paper is a simple crop prediction recommendation system

which is only limited to one state Tamil Nadu, as we hope to do more papers on other Indian states and encourage other fellow researchers to also persue research in the agriculture field, as this is our main source of food all over India. It alone contributes 60% of the entire GDP. But since 2018 it is gradually decreasing, the per capita water availability is also decreasing, which will result in a lot of crop production failures. And also there are multiple numbers of suicides of farmers all over India, who just work very hard and don't get the expected results due to many factors. This paper is a small contribution to the agriculture field and dedicated to all the farmers, to help them in their farming, so that they can get profits and benefits of the new technologies which they don't have any idea of. So finally we want to conclude that as an Engineer we should take responsibility and contribute our knowledge to the betterment of our society or country.

References

1. Xiao Z., Song W., Chen Q, **2018**,"Predicting the Yield of Crop Using Machine Learning Algorithms", *International Journal of Engineering Sciences & Research Technology*.
2. Tseng F., Chao H **2018**"Predictive Analysis to Improve Crop Yield Using a Neural Network Model", *IEEE, 2018.*
3. Fan Q. and Ansari N., **2019,"**Crop Yield Prediction Using Data Analytics and Hybrid Approach", *IEEE transaction*.
4. Satyanarayanan M., Bahl P., Caceres R., Davies N., **2009."**Machine Learning Methodologies for Paddy Yield estimation in India: a Case Study", *IEEE transaction.*.
5. Boukerche A. and Pimenta Sorghum Yield Prediction using Machine Learning, IEEE, 2019.
6. A. G. Howard, **2017** "Efficient Convolutional Neural Networks for Mobile Vision Applications". *Cornell University Library.17.*
7. Mark Sandler, **2018** "MobileNetV2: Inverted Residuals and Linear Bottle necks", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*
8. Shaoqing Ren, **2017** "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.39, Issue: 6, June 2017, pp. 1137-1149.*
9. Ali Farhadi, **2016** "You Only Look Once: Unified, Real-Time Object Detection", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.779-788.*
10. Tobi Delbruck, **2017**"Color temporal contrast sensitivity in dynamic vision sensors", *IEEE International Symposium on Circuits and Systems (ISCAS), pp. 638.*