Analysis of Crop Yield Prediction by using Machine Learning Algorithms

Nebeesath Sunaina

CSE Dept, CCET, Valanchery, Kuttipuram.

Abstract**—** Agriculture is considered as one of the main and a very foremost culture that is practiced in India. But the very effect of climate changes and its unpredictability has caused a big impact on farming. Hence crop yield prediction has become a very important step towards a very effective crop production and management. Predicting the crops well ahead of its harvest time helps the farmers in the cultivation of the right crops. This paper hence proposes a brief analysis of crop yield prediction using machine learning algorithms like K-Nearest Neighbor (KNN) Algorithm and Support Vector Machine Algorithms. The Experimental results shows that the proposed work efficiently predicts the crop yield.

Keywords— Machine Learning,KNN,SVM,crop yield,prediction

I. INTRODUCTION

Agriculture plays an important role in the growth of the national economy. Some of the factors on which farming depends on are soil,climate,humidity,rainfall,temperature and so on. Most of the times, farmers fail to achieve the expected result due to many reasons. Yield prediction is carried out which involves predicting the yield of the crop based on the existing data. In the past, prediction were mainly based on the farmers previous experiences. Yield prediction helps in identifying the most feasible crops to be cultivated for a particular region according to given environmental conditions.

In this paper, machine learning algorithms , K-Nearest Neighbours and Support Vector Machines are used to predict the most suitable crop. The crop production depends on vaarious factors which changes with every square meters and mainly depends on the geography of the region,weather conditions ,soil type, humidity and so on. Huge data sets can be used for predicting their influence on the major crops of that particular district or state. Machine learning techniques have advanced considerably over the past several decades.

KNN is a supervised classification method. It is also called as lazy learner technique because it depends on learning by analogy. It does not have a specialized training phase and uses all the data for training while classification. It uses feature similarity approach to predict the values of new data points which means the new data points will be assigned a value based on how closely it matches the points in the training set.

Support Vector Machines is a linear model for classification and regression models. The algorithm here creates a line or a hyperplane that can separate the data into classes. SVM tries to make a decision boundary such that the separation between the classes are as wide as possible. SVM chooses the points or the vectors that helps in creating the hyperplane. These are called the support vectors.

The proposed system here analyses these supervised machine learning algorithms in forecasting yield prediction from 4 different regions like Mangalore, Kodagu, Hasana and Kasaragodu. The dataset is collected from different online sources like Kaggle.com and data.govt.in. It consists of features like location, area in square foots, temperature (in Celsius), humidity, yield, rainfall (in mm), type of the soil, yields, and price. The paper is organized as follows: Section II presents the related work and Section III discusses about the proposed system. The experimental results on agricultural data are discussed in section IV. Finally, Conclusion is given in the section V.

II. RELATED WORK

Yield forecasting is an important service in the field of agriculture. In this section we highlight a few works done in the field of agriculture by using machine learning.

Renuka, Sujata Terdal (2019) proposed a system where the supervised learning is used to form a model, which provides the predicted cost of the crop yield as their corresponding production order. They have partitioned this into four stages as Dataset collection, preprocessing steps, feature selection and finally applying the machine learning models. The machine learning models they had used here is KNN, SVM and decision tree models. The data sets used in this paper are rainfall dataset, soil dataset and the yield dataset which they have gathered from several online dataset sites.

Pankaj Bhambri, Inderjit Singh Dhanoa, Vijay Kumar Sinha and Jasime Khaur (January 2020) in their paper analyzed a paddy crop production by using predictive analysis to predict the future events by implanting machine learning algorithms such as KNN and SVM. Prediction analysis is performed using data mining techniques such as J48, LAD tree and LWL. They have also compared their performance in terms of the accuracy and computation times.

Judicael Geraud N.Zannou and Vinasetan Ratheil Houndji(June 2019) proposed a system based on machine learning techniques to estimate farm yields. The experiments were conducted on a sorghum field. They have used techniques like convolutional neural networks and linear regression models to implement their work in the TensorFlow platform. These algorithms aided them in detecting the different ears of sorghum on an image and to estimate their weight. They have obtained an average accuracy of 74.5% for the detection of sorghum.

Prof. D.S. Zingade, Omkar Buchade, Nilesh Mehta, Shubham Godekar, Chandan Mehta (December 2017) proposed a paper that applies machine learning and prediction algorithms like Multiple Linear Regression to identify the patterns among the data and process it per input conditions. It proposes the best feasible crops to be cultivated in the given conditions given the location of the user.

N Hemageetha proposed a paper where she discussed various data mining techniques like Market based analysis, Association rule mining, Decision Trees, Classification and clustering. Naïve Bayes, J48, K-mean algorithms are explained in this paper. This paper helped in understanding different data mining algorithms and classification mechanisms.

III PROPOSED SYSTEM

In the proposed system, supervised learning algorithms are used to form a model which will help us in providing choices of the most feasible crops that can be cultivated in that region along with its estimated yield. Two of the algorithms used here is K-Nearest Neighbor and Support Vector Machine. The main stages involved in the process are dataset collection, pre-processing the data, feature extraction and classification.

A. Dataset collection

The dataset used for this project is collected from various online sources like Kaggle.com and data.govt.in. We have taken the agricultural data of four regions namely, Mangalore, Hasana, Kodagu and Kasaragod. Some important features or the parameters which has the highest impact on the agricultural yield considered in the project are listed below.

- Rainfall (in mm)
- Humidity
- Temperature
- Area
- Yield
- Type of the soil
- Location
- Price

B. *Pre-processing the data*

After the selection of the dataset, it has to be pre-processed into a form that you can work with. Some of the steps are formatting, cleaning and sampling. Initially the data you have selected is converted into the format suitable for you to work with. Cleaning data is the removal or fixing of the mixed data. Sampling is taking a small representative sample of the selected data that may be much faster for exploring the solutions than electing the hole dataset.

C. Transforming the data

The final step is transforming the selected data. The preprocessed data here is then transformed into data that is ready for machine learning algorithms by using various engineering features like scaling, feature aggregation and so on. There may be several features that can be combined into a single feature which would be more meaningful to the problem you are trying to solve. Figure 1 below shows the final data to be used by the classifiers. Figures 2 and 3 shows the system design of the proposed system.

```
[illegible]
[illegible],Coffee
[illegible],Coffee
[illegible],Coffee
[illegible],Coffee
[illegible],Coffee
[illegible],Cardamom
[illegible],Cardamom
[illegible],Cardamom
[illegible],Cardamom
[illegible],Cardamom
[illegible],Pepper
[illegible],Pepper
[illegible],Pepper
[illegible],Pepper
```

Fig 1: Final dataset

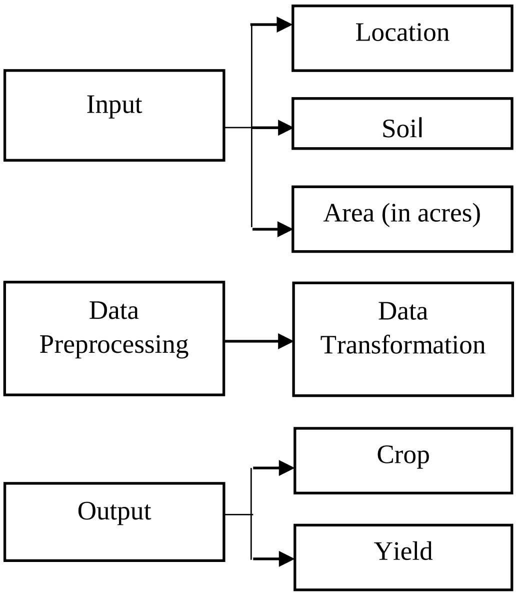


Fig 2: System design(a)

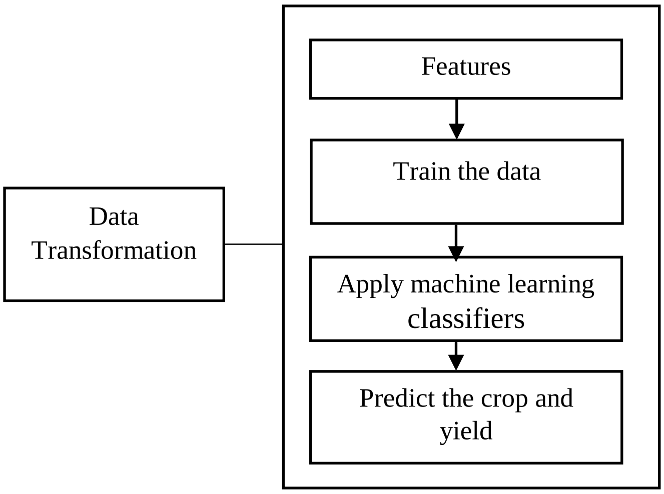

Data Transformation
Features
Train the data
Apply machine learning classifiers
Predict the crop and yield

Fig3: system design (b)

IV. ALGORITHMS USED FOR THE PROPOSED SYSTEM

The transformed data is then split into two sets namely, training sets and testing sets before applying the machine learning classifiers. The two machine learning classifiers used here are K-Nearest Neighbors and Support Vector Machine classifiers. Once the model is trained efficiently it is tested on the testing dataset which is different from the training data.

A. KNN algorithm

KNN is a supervised machine learning algorithm. It learns by analogy. It is a simple but a powerful approach for making predictions. In the project, according to the input given, the dataset is preprocessed to obtain the extracted dataset which is our training set. Test data is selected randomly from this training set. K-most similar records to the test record is calculated. Euclidian distance is calculated for finding the similarity between the records. Once k neighbours are discovered , a summarized predictions are made by returning the most common outcome.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x-y)^2} \quad (1)$$

Equation 1 shows the euclidian distance formula. It is used to calculate the distance between the two data points in a plane. KNN do not have a learning phase as such. It just calculates the distance between the test set and each row of the training set and returns the most similar ones as its neighbours. The lesser the distance, more similar the records are. Hence it is called as lazy learner technique or learning by analogy. But still it is considered as on of the most powerful agorithms.

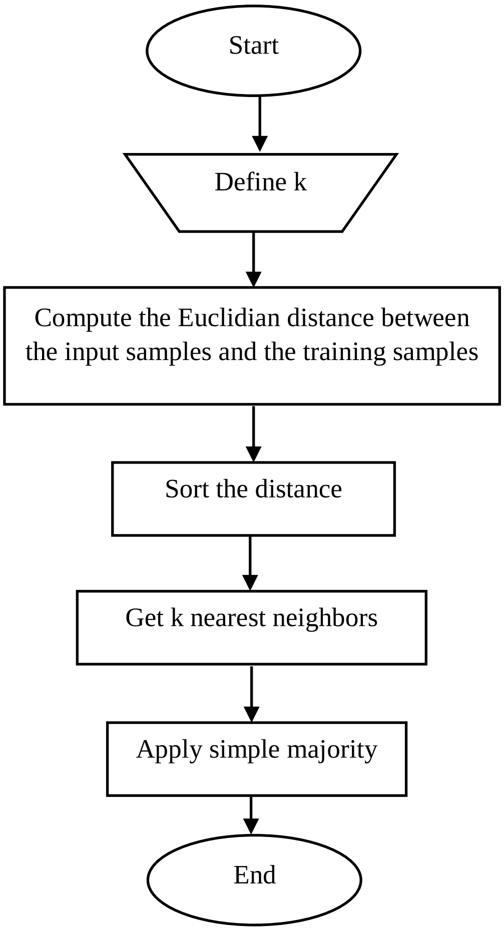

Start
Define k
Compute the Euclidian distance between the input samples and the training samples
Sort the distance
Get k nearest neighbors
Apply simple majority
End

Fig 4: KNN Classifier

Figure 4 shows the steps involved in the KNN Algorithm.

B) SVM algorithm

SVM algorithm is also a supervised machine learning algorithm which is used for both classification and regression problems. Hyperplane is the criteria that SVM uses to segregate the two classes. Finding the support vectors ie. the nearest data points to the hyperplane helps in giving the most optimal hyperplane. In the project, the first step is all the libraries are imported. Then the dataset is imported. It is then split into training samples and testing samples by using the sklearn library. A training model is built by importing the SVC classifier from sklearn SVM module.predict the values using the SVM algorithm. It has a higher acuuracy. It works well with all limited datasets. Kernel SVM contains a non-linear transformation function to convert the complicated non-linearly seperable data into linearly seperable data. Figure 5 shows the steps involved in the SVM algorithm.

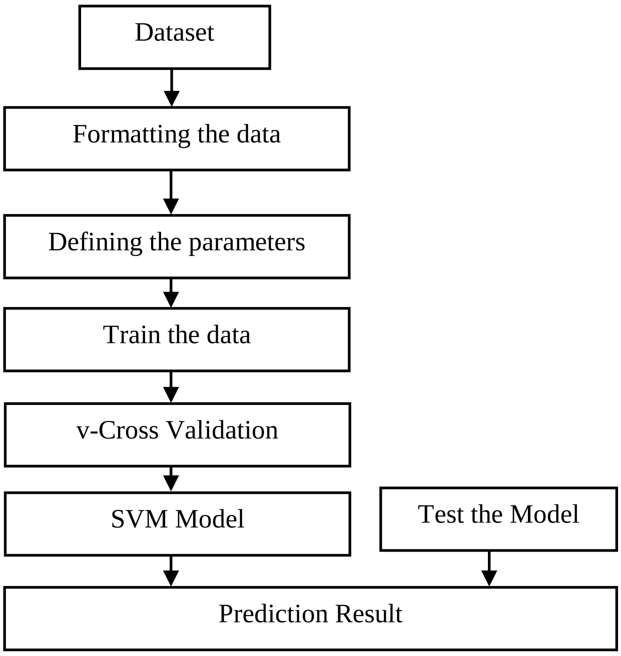
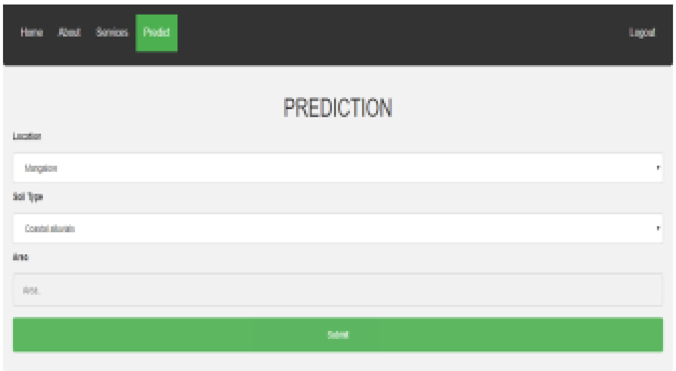


Fig 5: SVM Classifier

V. METHODOLOGY AND RESULTS

In the proposed system, initially it gathers basic user information like the Name, Email id, Password and phone number from the user. It is then stored into server's database. Once registered, Username and the password gets verified from the server side every time the user tries to log in to the website. The user can then select the option to predict the crop which is suitable for the particular location and the soil type as shown in the figure 6 below. He will also get the estimated yield of the predicted crop according to the number of acres of land he has given as the input.



PREDICTION

Location

Mangalore

Soil Type

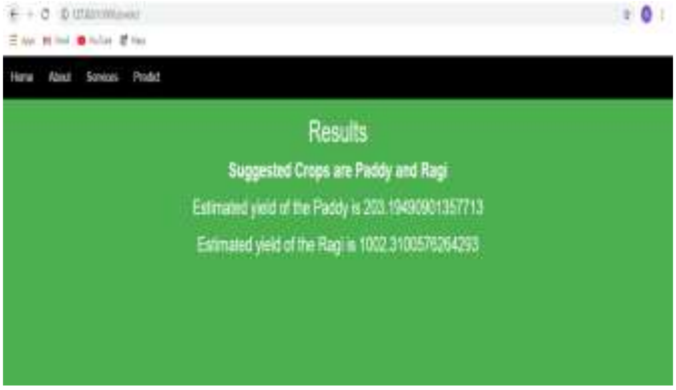
Coastal alluvials

Area

Area

Fig 6: Crop yield prediction

As shown in the figure 7 below, From the given inputs and by applying the machine learning classifiers, two choices of the crops are predicted along with its estimated yield for the given number of acres of land.



Results

Suggested Crops are Paddy and Ragi

Estimated yield of the Paddy is 203.19490901357713

Estimated yield of the Ragi is 1002.3100576264293

Fig 7: Prediction result

The implementation of KNN and SVM algorithm is done in Python and both the algorithms are compared in terms of their accuracy, Execution time in seconds from epoche, and in terms of their Precision and Recall scores. Classification Report is a report that is used to measure the quality of predictions from a classification algorithm. Precision, Recall, F1score are some of the metrics given in the Classification report. Precision is the ability of a classifier not to label an instance positive that is actually negative. Recall is the ability of a classifier to find all positive instances.

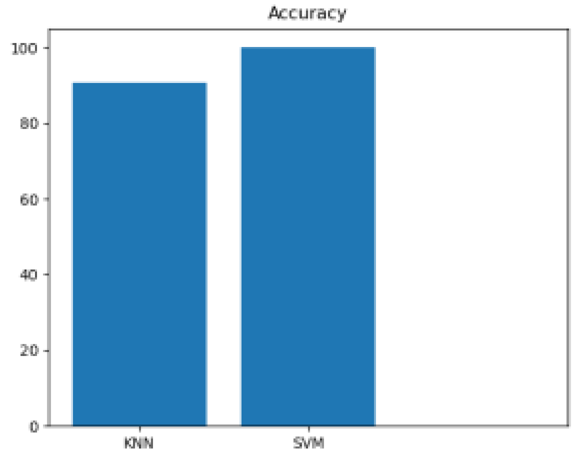


Fig 8: Accuracy of KNN and SVM

Figure 8 shows the accuracy of KNN and SVM classifiers. From the graph, it can be analysed that the accuracy of SVM is higher compared to that of the KNN classifier.

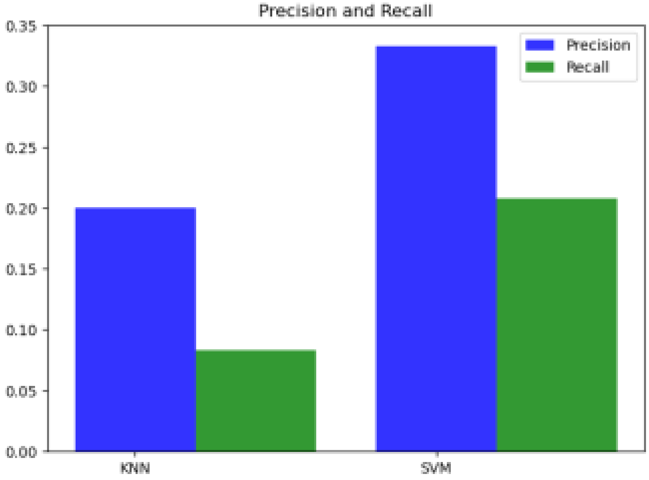


Fig 9: Precision and Recall

Figure 9 shows the Precision and Recall values of both KNN and SVM classifiers. The above graph shows that the precision and recall values of SVM is higher than that of KNN. The precision value of both the classifiers are higher than the recall value.

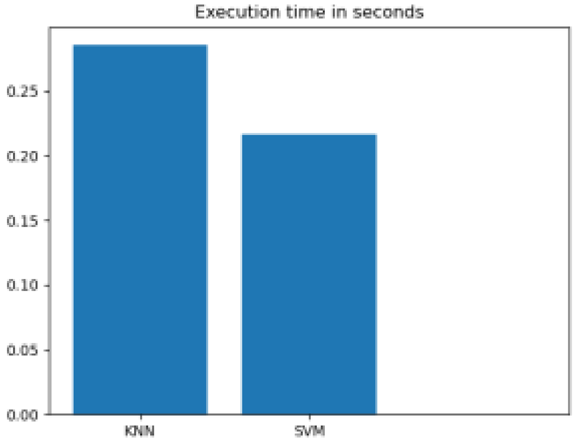


Fig 10: Comparison of Execution time(in seconds)

Figure 10 shows the time taken by the classifiers take for their execution. From the above figure, it can be analysed that SVM takes less time to execute compared to that of KNN.

VI. CONCLUSION

This project provides an easy to access and an efficient system to predict the crops and the estimation of its yield under the given conditions for a particular region. Farmers are still not connected with the modern technolgies. It efficiently bridges the gap between the rural farmers and the modern technologies. Machine learning algorithms have proved very effective in predicting the crops and its yield. From the comparison analysis of KNN and SVM, SVM works better than KNN for the dataset chosen. In the future, Farming can be taken to next levels by connecting all the farming devices to the internet using IOT.

REFERENCES

[1] Monali Paul, Santosh K. Vishwakarma, Ashok Verma, "Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining approach", 2015 International Conference on Computational Intelligence and Communication Networks.

[2] Abdullah Na, William Isaac, ShashankVarshney, Ekram Khan, "An IoT Based System for Remote Monitoring of Soil Characteristics", 2016 International Conference of Information Technology.

[3] Dr.N. Suma, Sandra Rhea Samson, S. Saranya, G. Shanmugapriya, R. Subhashri, "IOT Based Smart Agriculture Monitoring System", Feb 2017 IJRITCC.

[4] N. Heemageetha, "A survey on Application of Data Mining Techniques to Analyze the soil for agricultural purpose", 2016IEEE.

[5] DhivyaB, Manjula, Siva Bharathi, Madhumathi, "A Survey on Crop Yield Prediction based on Agricultural

Data", International Conference in Modern Science and Engineering, March 2017.

[6] GiritharanRavichandran, Koteeshwari **R S** "Agricultural Crop Predictor and Advisor using ANN for Smartphones", 2016 IEEE,

[7] R. Nagini, Dr. T.V. Rajnikanth, B.V. Kiranmayee, "Agriculture Yield Prediction Using Predictive Analytic Techniques, 2nd International Conference on Contemporary Computing and Informatics (ic3i),2016.

[8] Awanit Kumar, Shiv Kumar**,** "Prediction of production of crops using K-Means and Fuzzy Logic", IJCSMC, 2015

[9] JeetendraShenoy, YogeshPingle**,** "IOT in agriculture", 2016 IEEE.

[10] M.R. Bendre, R.C. Thool, V.R. Thool, "Big Data in Precision agriculture", Sept,2015 NGCT.

[11] Dr. Rakesh Poonia1, Sonia Bhargava "Prediction of Crops Methodology using Data Mining Techniques", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 6, Issue 10, October 2017.

[12] Mehta D R, Kalola A D, Saradava D A, Yusufzai A S, "Rainfall Variability Analysis and Its Impact on Crop Productivity - A Case Study", Indian Journal of Agricultural Research, Volume 36, Issue 1, 2002, pages: 29-33.