Comparative Study of Machine Learning models for Crop Recommendation System

Riddhi Narkar
Computer Engineering
A.P. Shah Institute of Technology
Thane, India.
19102003@apsit.edu.in

Avishkar Dalvi
Computer Engineering
A.P. Shah Institute of Technology
Thane, India.
20202002@apsit.edu.in

Aarya Totey
Computer Engineering
A.P. Shah Institute of Technology
Thane, India.
19102070 @apsit.edu.in

Radha Rakshe
Computer Engineering
A.P. Shah Institute of Technology
Thane, India.
19102067@apsit.edu.in

Prof. D.S. Khachane
Computer Engineering
A.P. Shah Institute of Technology
Thane, India.
dskhachane@apsit.edu.in

***Abstract*—Farming involves a lot of processes and steps to be done for a good harvest. Many of these processes involve decision-making, for example, the type of fertilizer, irrigation, crop, insecticide, etc. to be used. This work explores one such sector-crop recommendation. Crop recommendation. We did a comparative study of different machine learning models on a crop dataset to find out which model is the best for this task.**

Keywords—component, formatting, style, styling, insert (*key words****)***

I. Introduction

India has a primary sector economy, which means that the majority of our population is involved in agriculturally based activities and are dependent on crops related business for their livelihood. As per the census in 2011; in India, approximately 118 million people are farmers and 144 billion people are labourers working in agricultural fields [5].

Crop cultivation anywhere in the world depends on various factors and a good yield is directly dependent on studying and knowing these factors prior to harvesting. Precision agriculture is the science of improving crop yields and assisting management decisions using high-technology sensor and analysis tools [4]. Crop recommendation systems are an integral part of precision agriculture [2]. A crop recommendation system is a technology that suggests suitable crop options for farmers based on various factors such as soil quality, climate, soil pH, temperature, rainfall, and other relevant parameters and historical records to provide personalized recommendations for a particular farm or region. This can help farmers increase their yields, optimize resource use, and maximize profits while reducing the risk of crop failure.

In this work, we used a public crop dataset to train different Machine Learning models and found the best model for this task.

II. Dataset

The dataset we used contains a total of 2200 entries. The attributes are N, P, K, temperature, humidity, pH, rainfall, and crop.

Here N, P, K are the values of macro-nutrients nitrogen, phosphorous, and potassium respectively and are measured in ppm (parts per million).

TABLE I. Dataset attributes' details

Attributes	Details
N	Value of Nitrogen in ppm (parts per million)
P	Value of Phosphorous in ppm (parts per million)
K	Value of Potassium in ppm (part per million)
Temperature	Measured in ºC (degree celsius)
Humidity	Measured in gm^{-3} (grams of water vapour per cubic metre of air)
pH	Measured on a scale of 0 to 14
Rainfall	Measured in mm (millimetres)
Label	Target variable (crop name)

Temperature is measured in °C (degree celsius). Humidity is measured in gm^{-3} (grams of water vapour per cubic metre of air).

pH is the acidity/basicity of the soil. Rainfall is measured in mm (millimetres). The target variable is label, which has crop names. The data is collected from Indian regions and soil.

TABLE I contains a detailed summarization of all the attributes in the dataset we considered for the study.

There are a total of 22 different crops in the dataset namely: muskmelon, kidneybeans, papaya, pigeonpeas, blackgram, cotton, mothbeans, mungbean, watermelon, orange, mango, banana, rice, pomegranate, chickpea, apple, jute, grapes, lentil, coffee, maize, and coconut.

III. Methodology

The dataset was split into 3 parts - 80% for training, 10% for testing and 10% for validation.

Since this is a categorical type dataset, the first choice was naturally, logistic regression. Linear regression for classification, however, won't be a good choice, as its cost function is not optimal for classification problems [6]. Along

with this, we used a decision tree, another well-known and powerful model for classification models.

The dataset we had had many attributes, and since this problem is a classification, SVM was another good option. This work was initiated with having higher hopes from SVM due to its powerful boon of saving from the 'Curse of Dimensionality'.

The dataset had many factors and as crops need the perfect conditions of rainfall and soil nutrients, it was concluded that these attributes are related to each other in some or the other way [3]. To test that out, Naive Bayes was also implemented, as it assumes that no attribute is dependent or related to any other attribute and all are completely independent.

Lastly, we implemented a few ensemble techniques for more effective training [1]. Of the ensemble techniques, we used random forest, which uses a forest of a myriad of decision trees; and XGBoost.

A. Logistic Regression

Logistic regression is a statistical machine learning model which estimates the probability of an event (in this case, classification) and needs a dataset of independent variables. The output here is a mere probability, and hence, is bounded in the interval [0, 1]. The logistic function is as follows:

$$f(x) = \frac{1}{1 + e^{-x}}$$

$$ln\left(\frac{f(x)}{1 - f(x)}\right) = \beta_0 + \beta_1 * x_1 + \cdots + \beta_k * x_k$$

B. Naïve Bayes

Naïve Bayes, just like Logistic Regression 'naively' assumes complete conditional independence. Naive Bayes calculates the probability of a result assuming strong conditional independence between all attributes. It uses Bayes' theorem for its calculation. Bayes' theorem states that:

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

In plain English, Bayes' theorem can be demonstrated as:

$$posterior = \frac{prior * likelihood}{evidence}$$

A Naïve Bayes classifier uses this formula to calculate the probability of classification.

C. Decision Tree

A decision tree is a non-parametric classification tool and can be used for both regression and classification tasks. It has a tree structure, consisting of a root, internal nodes and leaves. It helps to evaluate different choices between several courses of action. Every internal node is a decision and on the basis of the decision, you travel from the root to the leaves. The leaves are any of the possible outcomes of the classification (target variable values).

D. SVM

SVM, an abbreviation of Support Vector Machine is an extremely powerful model for classification, outlier detection and regression. Its performance is independent of the total attributes in a dataset; it even works in cases where the total number of attributes (or dimensions) is greater than the total number of samples. SVM calculates the most optimal boundary between the classes in the dataset to classify.

E. Random Forest

It is an ensemble learning technique which creates a forest of 1-level decision trees. It can be used for both regression and classification tasks. A decision tree classifier becomes more vulnerable to overfitting when the tree grows. Hence, to avoid that, only trees of height one are taken in a random forest classifier. This method helps maintain the accuracy of decision trees. The individual trees need to be as uncorrelated to each other as possible to further increase accuracy.

F. XGBoost

XGBoost is an open-source optimized gradient-boosting library. It is an ensemble technique, and it makes a stronger prediction using multiple weak models and tweaking their weights and improving their performance. XGBoost stands for 'Extreme Gradient Boosting'. It is efficient even in large dataset sizes, It can be used for both classification an regression tasks.

In XGBoost, decision trees are created in sequentially. Weights, which are very integral and important in XGBoost, are assigned to all the independent variables. This is then fed into a decision tree and predictions are made. After this first iteration, the weights of wrongly predicted variables are increased and a 2nd iteration is made in a similar fashion. Such ensembling makes for a more precise model.

IV. Observations

Ensemble techniques proved to be more powerful than regular techniques. Random Forest and Naïve Bayes have the second highest accuracy, out of which Random Forest is an ensemble technique. The other ensemble technique implemented was XGBoost which has the highest accuracy.

Naïve Bayes, even being a regular model stood alongside Random Forest. Thus, opposing our previous assumption that the attributes are fairly dependent on each other, Naïve Bayes got the second-highest accuracy. Hence, the assumption thus is false, and the attributes, mathematically are quite independent.

V. RESULTS

The resulting accuracies of all models are given below:

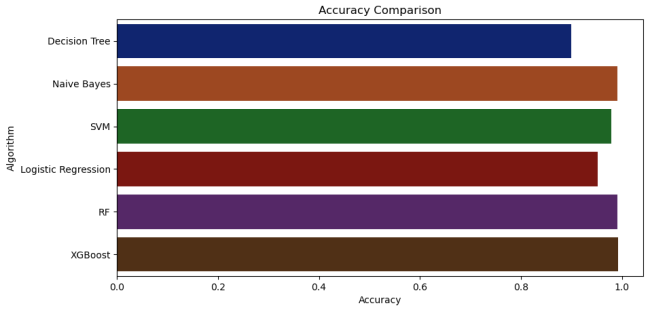

Fig. 1. A Bar Graph comparing accuracies

Fig. 1 illustrates a comparative bar graph with accuracy on the x-axis and model names on the y-axis. Accuracy is depicted as a value between 0 and 1.

The decision tree model yields the least accuracy of all, probably due to inaccuracies cause by a large number of attributes.

Logistic Regression stands ahead of the decision tree. SVM ranks next, doing significantly better as the number of attributes is independent to its performance.

Both Naïve Bayes and Random Forest, two powerful ensembling techniques yield the same accuracy and better than SVM.

Lastly, XGBoost stands first at 99.3182% accuracy, beating all other models. Hence the weighting technique for weaker models clearly helps in this case. Weaker decision trees, when been weighted appropriately and then recalculated for results yielded a better learning mechanism for this dataset.

TABLE II. ACCURACY COMPARISION

Models	**Accuracies (in %)**
Decision Tree	90.0000 %
Naïve Bayes	99.0909 %
SVM	97.9545 %
Logistic Regression	95.22727 %
Random Forest	99.09091 %
XGBoost	99.3182 %

TABLE II. showcases the accuracy comparison in a table format. Here accuracy is depicted as a percentage.

Hence, the best classification machine learning model is XGBoost, for crop recommendation using the dataset we used to implement this classification problem.

REFERENCES

[1] Nidhi H Kulkarni, Dr. G N Srinivasan, Dr. B M Sagar, Dr.N K Cauvery, "Improving Crop Productivity Through A Crop Recommendation System Using Ensembling Technique", 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018.

[2] Pradeepa Bandara, Thilini Weerasooriya, Ruchirawya T.H., W.J.M. Nanayakkara, Dimantha M.A.C, Pabasara M.G.P., "Crop Recommendation System", International Journal of Computer Applications (0975 – 8887) Volume 175– No. 22, October 2020.26.

[3] Sayed Mazhar Ali, Bhagwan Das, Dileep Kumar, "Machine Learning based Crop Recommendation System for Local Farmers of Pakistan", A Research Gate Preprint Online Publication 2021, Pakistan.

[4] Dhruv Piyush Parikh, Jugal Jain, Tanishq Gupta, Rishit Hemant Dabhade, "Machine Learning Based Crop Recommendation System", International Journal of Advanced Research in Science, Communication and Technology, 2021, Chennai, India..

[5] "Sectorwise GDP of India", https://statisticstimes.com/economy/country/india gdpsectorwise.php, sourced from Ministry of Statistics and Programme Implementation, 17th June, 2021.

[6] Younes OMMANE, Mohamed Amine RHANBOURI, Hicham CHOUIKH, Mourad JBENE, Ikram CHAIRI, "Machine Learning based Recommender Systems for Crop Selection: A Systematic Literature Review", A Research Gate Preprint Online Publication, 2022.