Crop Analysis

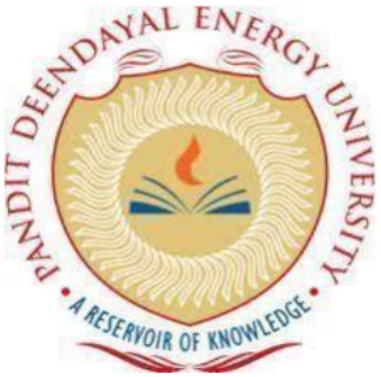
by Aniket Gupta

Submission date: 20-Nov-2022 03:30PM (UTC+0530)
Submission ID: 1959092953
File name: DM_Report_Crop_Analysis_Final_1.pdf (507.87K)
Word count: 2125
Character count: 11610

PANDIT DEENDAYAL ENERGY UNIVERSITY

SCHOOL OF TECHNOLOGY

Title : Crop Analysis and Prediction Using Machine Learning



Course: Data Mining

PROJECT REPORT

B.Tech. (Computer Science and Engineering)

Semester 5

Submitted To

Dr. Rajeev Gupta

Submitted By

Aniket Gupta (20BCP049)
Harsh Baheti (20BCP065)
Kunal Gupta (20BCP068)

Abstract:

This report aims to solve the problem of farmers cultivation using a public dataset “Crop Prediction and Analysis”. India is the second-most populous nation in the world, and the majority of its citizens work in agriculture as a primary employment. Farmers produce the same crops again and over again without attempting new varieties, and they apply fertiliser in arbitrary amounts without considering the level of deficiency. Therefore, in order to benefit farmers, we created the system utilising machine learning techniques. Based on soil characteristics and climatic conditions, our technology will recommend the optimum crop for a certain piece of land. Additionally, the system offers details on the type and amount of fertilisers that must be used for growing. In light of this, farmers can cultivate using our technique for crops that will increase their profit margin and can avoid soil pollution. We adopt a classification-based approach that uses three algorithms (Random Forest Classifier, Logistic Regression and KNN) to compare the metrics and choose the most refined model. So, therefore scaling them between 0 and 1 with MinMaxScaler. Secondly, we applied label encoding so that names of crops can be converted to number for the output. Lastly, we are going to apply different algorithms and the best accurate will be selected as to predict the best crop to grow in the conditions required.

Table Of Content:

List Of Figures

Chapter 1 – Introduction

To provide for the requirements of 1.3 billion people, more than 60% of the country's land is used for agriculture. Consequently, implementing new agricultural technologies is crucial . Machine Learning is well equipped when it comes to analyzing information on the characteristics of the soil, such as its temperature, moisture content, and chemical composition, all of which have an effect on crop growth. Today in agriculture, this can allow crops to be grown at much higher precision, which in turn significantly increases the effectiveness of farmers' decisions. The development of the models and evaluation of the one that makes the most accurate predictions were the main objectives. To recall the creation and application of the predictive model, a number of rule-based machine learning techniques were used. Several models were started using different machine learning (ML) algorithms that gathered raw data and then divided it into groups based on soil content or weather conditions. Next, the data set was processed using a variety of machine learning (ML) models, including random forest classifier, logistic regression and KNN. ML models indicated that the accuracy varied. Every model was given the same set of input parameters to process the data, and as a result, each model produced the best crop to be grown with a different level of accuracy. It has been decided to use the model with the maximum accuracy.

Chapter 2 - Literature Review

Paper - 1

This study uses agricultural and weather data to propose a machine learning framework for crop production prediction. In order to anticipate the yield of 80 crops in India using historical data, it also evaluates the effectiveness of potential machine learning techniques like regression, decision trees, random forests, support vector machines, and gradient boosting . Among the mentioned ML methods The decision tree regressor is a good regressor but is a weak learner and is more prone to overfitting, linear regression is a simple algorithm but cannot work well on complex data while Random Forest Classifier performed well on both training and testing datasets because of it's non-linear and ensemble nature. Hence the paper concluded that for the crop prediction dataset random forest classifier turns out to give the best accuracy among all others.

Paper - 2:

On a dataset from the Indian government, experiments by Aruvansh Nigam, Saksham Garg, and Archit Agrawal showed that the Random Forest machine learning method provides the best yield forecast accuracy. Simple Recurrent Neural Network, a sequential model, is more effective at predicting rainfall than LSTM is at predicting temperature. For the purpose of yield forecast, the article combines variables such as rainfall, temperature, season, area, etc.

When all parameters are considered, the results show that Random Forest is the best classifier.

Leo Brieman focuses on the reliability, power, and correlation of the random forest method. The random forest algorithm builds decision trees using several data samples, predicts the data from each subset, and then determines the best solution for the system through voting. The data was trained in Random Forest using the bagging method. The randomness added must maintain strength while reducing correlation in order to increase accuracy.

Crop yield prediction has been implemented by Balamurugan using simply the random forest classifier. To anticipate the agricultural output, various factors like rainfall, temperature, and season were considered. On the datasets, no further machine learning methods were used. Because alternative algorithms were lacking, comparison and quantification could not be done, making it impossible to provide the best algorithm.

Mishra, has theoretically described various machine learning techniques that can be applied in various forecasting areas. However, their work fails to implement any algorithms and thus cannot provide a clear insight into the practicality of the proposed work.

According to Dr. Y. Jeevan Nagendra Kumar, supervised learning allows machine learning algorithms to forecast an objective or outcome. This study focuses on supervised learning methods for predicting crop yields. It must create an acceptable function using a set of variables that may map the input variable to the desired output in order to obtain the outputs that are required. According to the paper, crop predictions may be made using the Random Forest ML method, which achieves the best accuracy value while taking into account the fewest number of models.

Chapter 3 - Proposed Methodology

This section explains the suggested methodology or method employed in this study to assist end users in crop prediction based on soil content and weather conditions Data gathering, pre-processing, splitting, construction of a classification model, and evaluation of the classification model. The following sections go into great depth on each phase.

3.1 Data Collection

The datasets used in this investigation were compiled from a publicly available dataset that linked measures to predict crop. The initial dataset was already pre-compiled into a number elements in nature or soil contents that were plotted alongside the crop to be grown. The dataset comprises 2200 instances ,7 attributes, a target feature with 22 different classes.

3.2Text Pre Processing

Null values has been checked by (isna()) function which concludes no missing values.

A quick check if the dataset is balanced or not using desc() which also concluded dataset is balanced and all the classes have equal or nearer value.

A very important plot to visualize the distribution between two features for all the combinations. It is great to visualize how classes differ from each other in a particular space using pair plot.

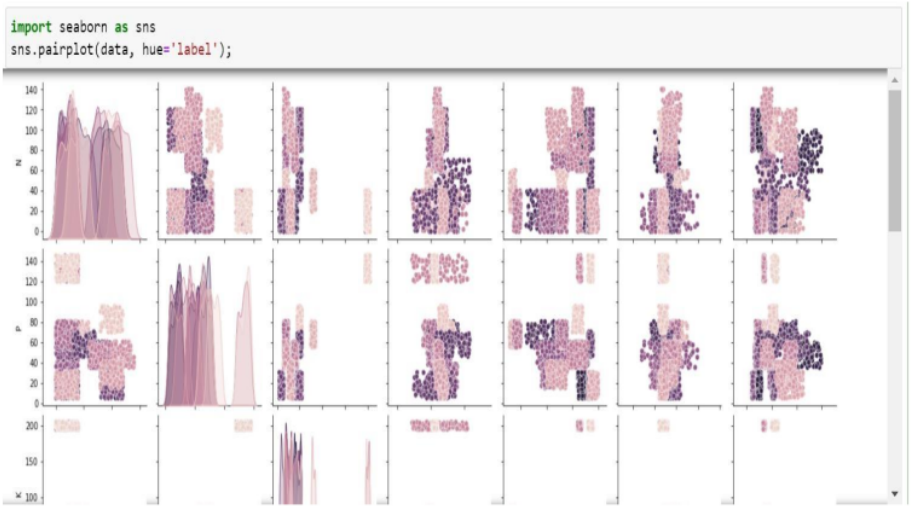


Figure 3.1 Pair Plot for data visualization.

- The average annual rainfall is high (120 mm on average) and the temperature is cool (less than 30°C) during the wet season..
- Rain affects soil moisture which affects pH of the soil. Here are the crops which are likely to be planted during this season.
- Rice requires a lot of rain (>200 mm) and humidity levels of 80%. It is understandable why East Coasts, which receive an average of 220 mm of rainfall annually, produce the majority of India's rice
- .Coconut is a tropical crop and needs high humidity therefore explaining massive exports from coastal areas around the country.

Correlation visualization between features. We can see how Phosphorous levels and Potassium levels are highly correlated using heatmap.

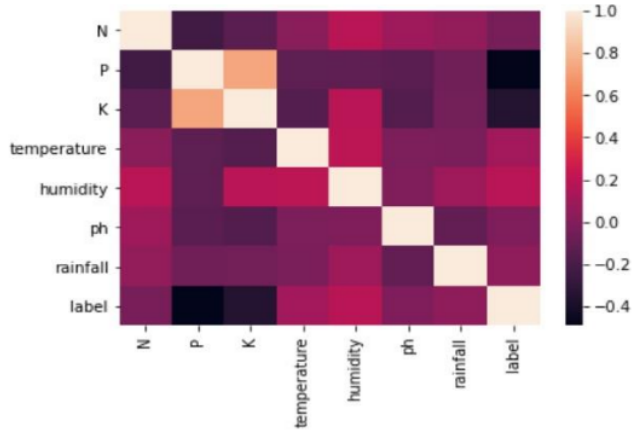


Figure 3.2 Heatmap between all the features.

3.3 Feature Engineering

The categorization criteria in the raw text are not understandable to the ML algorithm. For these algorithms to comprehend categorization rules, numerical characteristics are required. Consequently, feature engineering is a crucial stage in text categorization. The primary features from the raw text are retrieved in this stage, and the characteristics are then represented numerically. Feature scaling is required before creating training data and feeding it to the model.

As we saw earlier, two of our features (temperature and pH) are gaussian distributed, therefore scaling them between 0 and 1 with MinMaxScaler.

3.4 Data Splitting

The dataset is split into training and testing data inputs using the standard train-test-split methodology. The other 7 features serve as the input for estimating the condition, with the crop resulted serving as the value to be predicted. The dataset is randomised using the random state parameter in order to maintain the percentage of multi label data and train the model to produce an accurate model.

3.5 Machine Learning Model

Random Forest or Logistic Regression , according to studies, would produce the most accurate models. The dataset we used, however, differs from those previously employed in a number of ways. As a result, we first train the model using Random Forest with tuned hyperparameters, followed by Logistic Regression and K-Nearest Neighbors and it turned out that K-Nearest Neighbors gave more accurate results among others.

Chapter 4 - Implementation Details

Python is the programming language of choice for the current project for all implementations, from data extraction to model evaluation. Python is more suited for problem-solving in real-world scenarios because of its extensive library support for applications in the field of machine learning. The project's code was written in Jupyter notebooks. Python modules including NumPy, Pandas, Matplotlib, and Seaborn were used for all of the exploratory data analysis. Using the Scikit-Learn library and its classes, the model was chosen, trained, and evaluated . We employed a variety of multiclass classification algorithms to forecast the crop yield from various parameters.

Step 1: Importing dependencies

This step will download the libraries and resolves them so that they are available in our project.

Step 2: Analysing Data

Pandas library is used to represent data in tabular format in rows and columns.

Step 3: Checking null values and whether the data is balanced or not.

```
data.isna().any()
```

```
N              False
P              False
K              False
temperature    False
humidity       False
ph             False
rainfall       False
label          False
dtype: bool
```

Figure 4.1. Table for Null values

```
sns.countplot(y='label',data=data)
#This means the dataset is balanced.
```

```
<AxesSubplot:xlabel='count', ylabel='label'>
```

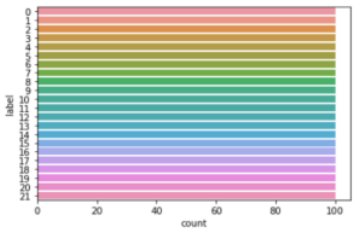


Figure 4.2. Graph for balanced data.

Step 4: Applying Label Encoding to convert the target variable from object to an integer.

Step 5: Exploratory Data Analysis(EDA):

Step 6: Feature Scaling is done by splitting dataset into training and testing model and then MinMaxScaler is applied.

Step 7: Now, evaluation of all the three model(Random Forest Classifier, Logistic Regression, K-Nearest Neighbors) is done and corresponding classification report is produced.

Step 8: Accuracies of all the models are compared and KNN produced highest accuracy among all.

```
plt.xlabel("Model")
plt.ylabel("Accuracy")
x = ['Logistic Regression','K-Nearest Neighbours','Random Forest Classifier']
y = [accuracy_score(ytest, pred), knn.score(X_test_scaled, ytest), accuracy_score(ytest, ypred)]
plt.scatter(x, y, c ="blue")
plt.show()
```

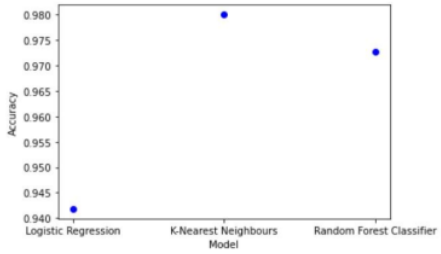


Figure 4.3 Comparison Graph of all the models.

Chapter 5 - Result Analysis

Metrics for evaluation are used to gauge the model's quality. How to assess your model is one of the most crucial machine learning concepts. Measuring how well our model forecasts the predicted outcome based on the supplied inputs is vital while developing it.

For various machine learning methods, we have multiple assessment criteria. We employ classification measures like Accuracy, Precision, Recall, and F1-score to assess classification models.

Classification Reports of all the model are a s follows:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	24
1	1.00	1.00	1.00	24
2	0.78	1.00	0.88	28
3	1.00	1.00	1.00	23
4	1.00	1.00	1.00	23
5	1.00	1.00	1.00	21
6	1.00	1.00	1.00	30
7	1.00	1.00	1.00	37
8	0.92	0.89	0.91	27
9	1.00	1.00	1.00	27
10	0.92	1.00	0.96	23
11	1.00	1.00	1.00	24
12	1.00	1.00	1.00	23
13	1.00	0.66	0.79	29
14	1.00	1.00	1.00	26
15	1.00	1.00	1.00	33
16	1.00	1.00	1.00	19
17	1.00	1.00	1.00	15
18	1.00	1.00	1.00	23
19	1.00	1.00	1.00	28
20	0.84	0.89	0.86	18
21	1.00	1.00	1.00	25
accuracy			0.97	550
macro avg	0.98	0.97	0.97	550
weighted avg	0.98	0.97	0.97	550

Figure 5.1 Random Forest Classifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	24
1	1.00	1.00	1.00	24
2	0.89	0.89	0.89	28
3	1.00	1.00	1.00	23
4	1.00	1.00	1.00	23
5	1.00	1.00	1.00	21
6	0.94	1.00	0.97	30
7	1.00	1.00	1.00	37
8	0.94	0.63	0.76	27
9	0.96	1.00	0.98	27
10	0.88	1.00	0.94	23
11	1.00	0.92	0.96	24
12	0.77	1.00	0.87	23
13	1.00	0.72	0.84	29
14	0.96	1.00	0.98	26
15	1.00	1.00	1.00	33
16	1.00	0.95	0.97	19
17	0.85	0.73	0.79	15
18	1.00	0.87	0.93	23
19	1.00	1.00	1.00	28
20	0.59	0.94	0.72	18
21	1.00	1.00	1.00	25
accuracy			0.94	550
macro avg	0.94	0.94	0.94	550
weighted avg	0.95	0.94	0.94	550

Figure 5.2 Logistic Regression

	precision	recall	f1-score	support
0	1.00	1.00	1.00	24
1	1.00	1.00	1.00	24
2	0.97	1.00	0.98	28
3	1.00	1.00	1.00	23
4	1.00	1.00	1.00	23
5	1.00	1.00	1.00	21
6	0.97	1.00	0.98	30
7	1.00	1.00	1.00	37
8	0.86	0.93	0.89	27
9	1.00	1.00	1.00	27
10	0.92	1.00	0.96	23
11	1.00	0.96	0.98	24
12	0.96	1.00	0.98	23
13	1.00	0.90	0.95	29
14	1.00	1.00	1.00	26
15	1.00	1.00	1.00	33
16	1.00	1.00	1.00	19
17	1.00	1.00	1.00	15
18	1.00	0.96	0.98	23
19	1.00	1.00	1.00	28
20	0.88	0.78	0.82	18
21	1.00	1.00	1.00	25
accuracy			0.98	550
macro avg	0.98	0.98	0.98	550
weighted avg	0.98	0.98	0.98	550

Figure 5.3 K-Nearest Neighbors

Chapter 6 - Conclusion & Future Work

Presently our farmers are now not employing technology and analysis properly, there is a danger that they will choose the wrong crop to cultivate, which will lower their income. We created this model to estimate the optimum crop for a specific piece of land and to provide information about the nutrients that must be added for better cultivation in order to prevent those types of losses. This way ML will help in the growth of agricultural sector.

References

1. https://www.ijert.org/research/comparative-analysis-of-machine-learning-algorithms-in-the-study-of-crop-and-crop-yield-prediction-IJERTCONV8IS14008.pdf

2. https://www.ijert.org/crop-yield-prediction-using-machine-learning-algorithms

3. https://www.researchgate.net/publication/343631997_Crop_Prediction_using_Machine_Learning_Approaches

4. https://www.ijert.org/prediction-and-analysis-of-crop-yield-using-machine-learning-techniques

Crop Analysis

ORIGINALITY REPORT

10%	9%	3%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ijert.org Internet Source	4%
2	www.irjmets.com Internet Source	2%
3	www.slideshare.net Internet Source	1%
4	ri.conicet.gov.ar Internet Source	1%
5	www.ncbi.nlm.nih.gov Internet Source	1%
6	"Recent Advances on Soft Computing and Data Mining", Springer Science and Business Media LLC, 2017 Publication	<1%
7	www.farnell.com Internet Source	<1%
8	medium.com Internet Source	<1%
	www.coursehero.com	

9	Internet Source	<1%
10	Aruvansh Nigam, Saksham Garg, Archit Agrawal, Parul Agrawal. "Crop Yield Prediction Using Machine Learning Algorithms", 2019 Fifth International Conference on Image Information Processing (ICIIP), 2019 Publication	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		