National College of Ireland

Predictive Modelling to Forecast the Crop Yield and Classification of Plant Diseases

MSc Research Project
Data Analytics

Shreya Merkaje Ravi
Student ID: x18190910

School of Computing
National College of Ireland

Supervisor: Dr. Muhammad Iqbal

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Shreya Merkaje Ravi
Student ID:	x18190910
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Muhammad Iqbal
Submission Due Date:	30/09/2020
Project Title:	Predictive Modelling to Forecast the Crop Yield and Classification of Plant Diseases
Word Count:	4705
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	29th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predictive Modelling to Forecast the Crop Yield and Classification of Plant Diseases

Shreya Merkaje Ravi
x18190910

With the increase in population, there is a requirement to increase food production. As per the Food and Agriculture Organization(FAO)[1] *around 60% of the world's population practice agriculture. Production of crops is largely dependant on climate. Developing techniques to predict the yield of crops under various weather aspects will help the farmers in making better decisions regarding crops. Machine learning can be extensively used for predicting the yield of crops. In this paper, various machine learning algorithms are applied to predict crop yield. Along with this, the classification of different kinds of plant diseases is also carried out using a plant image dataset. Both regression and classification techniques are used for the first part and for the second part neural networks are used.*

1 Introduction

Agriculture is one of the main occupations in the world. An increase in population has put stress on food production and many modern technologies are now available to improve agricultural methods. Also, the change in climatic conditions it has become a necessity to improve and secure the food resources. Predicting crop yield will help the farmers to plan the type of crop. This also helps in improving food security and with the help of predicted values, import and export of crops can be improved. With the information such as temperature, rainfall, and production area it is possible to predict the yield accurately and then make suggestions to the farmer to make a better judgment on the type of crop that can be cultivated to maximize the yield.

In this paper, an approach is outlined to predict the yield and to classify the infected crop images. In the first part, historical data of rainfall and crops is used to forecast the yield by implementing regression techniques. The datasets are acquired from Kaggle and have data for every state in India. It consists of crops grown all around the country, their seasons, and production.

The second part of the project is the identification and classification of infected plant images and this dataset is acquired from Kaggle as well. The goal is to apply deep learning techniques for pest detection thereby removing any manual intervention to achieve an automated and real-time pest monitoring system.

1.1 Research Question

RQ: "How can machine learning be used to crop Yield and classify the pest attacks using agriculture, rainfall, and image data?"

[1] http://www.fao.org/home/en/

1.2 Research Objectives

1. The first objective is to implement a regressor to predict the yield of crops based on historical rainfall data.

2. The second objective is to classify the different diseases that occur in plants using neural networks.

Rest of this paper is structured as follows: Section 2 contains a detailed literature review, section 3 outlines the research the methodology used, section 4 explains the data procurement and cleaning, section 5 explains the implementation and evaluation of the applied machine learning algorithms and finally, section 6 concludes the research paper with the room for future work.

2 Literature Review

This section concentrates on the methods followed to forecast yield and suggest better crop selection, and also about the pest classification using different feature sets and data.

This section is further divided into three sub-sections. (1) Feature set selection, (2) Data mining methods, and (3) Other methods

2.1 Feature Set Selection

Many of the forecasting models make using data mining and machine learning algorithms. For better performance and accuracy it is important to select the best possible features.

(Gopal & Bhargavi 2019) used various feature selection algorithms before implementing the models to predict the yield. 30 years of data related to paddy crop was collected and had 16 features such as maximum temperature, minimum temperature, rainfall, precipitation, etc. After the cleaning of raw data various feature selection algorithms such as sequential forward feature selection, correlation-based selection, and random forest variable importance were applied to select different feature subsets. These selected features were then applied to the Multiple Linear Regression(MLR) model to identify the best possible feature set.

Machine learning models such as MLR, Artificial Neural Network (ANN) were used to predict the yield of crops. In ANN it was seen that the nodes and learning rate influenced the accuracy of the model (Ji et al. 2007). A hybrid of MLR-ANN has also been applied wherein the initial weights for the input layer are calculated by using the MLR equation. This hybrid model gave better prediction accuracy when compared to the other two applied models.

2.2 Data Mining Methods

(Gandge et al. 2017) looked into the various methodologies using which the crop yield can be predicted. Different soil factors and pesticides are considered for the prediction.

Different algorithms such as support vector machine, decision tree algorithm, neural networks, K-means algorithm, and Naive Bayes are explained with their efficiency.

(Ghadge et al. 2018) proposed a system to analyse the soil quality using data mining techniques. Various nutrients available in the soil are identified based on crop production and location. Both supervised and unsupervised learning algorithms such as Kohonen Self Organizing Map (Miljković 2017) and Back Propagation Network (Hecht-Nielsen 1992) were used in this study. This data can further be used to predict the type of crop suitable for a particular soil type and thus maximizing the yield. The proposed system took Ph value and location as input from the user and with the help of third-party APIs, different weather and soil aspects were acquired. Using these values different nutrients in the soil and its alkalinity is identified and compared with the predefined database.

(Gandhi et al. 2016) explained how different weather conditions influence the production of crops. The dataset used for this study was obtained from the open-source website which has variables like temperature, area, and precipitation. After cleaning the data, the Support Vector Machine (SVM) approach was used to generate functions from a labeled training data and Sequential minimal optimization(SMO) algorithm was used to analyse the performance of the SVM approach. The accuracy obtained was 0.69 which is less when compared to other data mining algorithms.

(Kumar et al. 2019) designed a model that is capable of suggesting suitable crops that can be grown based on various soil factors and predict pest attacks that the plants are prone to and provide measures to control them. Algorithms such as SVM, DTM, and Logistic Regression were applied. Data was collected from various online resources. Various factors of soil such as moisture, PH, temperature were considered for the study. Spyder IDE was used to design the model. All the algorithms that were applied gave good results and among them, SVM had the highest accuracy percentage of 89.66. This model is best suited where agriculture is is the primary source of income and even the smallest of investments can be taken care of. Using these farmers/agriculturists will get an idea of the kind of crops that can be grown and improve the yield.

(Satir & Berberoglu 2016) in this study, GIS techniques were used to predict the yield of potato. Two kinds of satellite images(Landsat-8 and Sentinel-2) were obtained during the growing stages of potato and two different sets of vegetation index were generated. Using these data maps were generated and based on vegetation index different zones were classified. The samples of potato yield were collected a couple of days before harvesting and were compared to the two different vegetation index. The results showed that the predicted values were better for Landsat-8 images. Using satellite images can help the farmers in understanding the variations in productivity but small and mid-scale farmer

(Oliveira et al. 2018) This paper focuses on predicting the yield of soybean and maize in two different locations in the US and Brazil. Different factors such as precipitation and temperature and different soil factors were used for the study. This predictive model made use of Deep Neural Network (DNN). As per the result obtained, the region is the US had better productivity when compared to the region in Brazil. It is also notable to see how both crops react to changing weather conditions. The yield of Maize was found to be more sensitive to changing weather and soil conditions compared to soybean. This knowledge will help farmers to obtain better yield in the future.

s cannot opt for this method since it is expensive.

(Wang et al. 2017) outlined the importance of food security and identifying plant diseases at the earliest. For this purpose, a deep learning model was built for assisting in plant disease control. Multiple neural network models were implemented and compared in the study. A convolution model with 8 layers was applied to the input data and accuracy of 79 was obtained but the accuracy seemed to reduce after 8 runs.

(Barbedo 2019) proposed the study on the diagnosis of plant diseases based on spots and lesions seen on a plant leaf or a part of the plant. The whole study took into consideration 14 species and 70 odd diseases. The study took into consideration the previous work in this field, their advantages, and disadvantages. The accuracy of each of these studies was noted and a new model was proposed which gave higher accuracy. GoogleNet architecture was used for the experiment owing to its better performance. The experiment had two different parts where the first deals with the classification and second with detecting plant diseases. CNN model was used for the experiment and it showed considerably better results but the experiment was conducted on a small dataset and may not show the same results when applied on larger datasets.

Another study by (Sabrol & Satish 2016) in this field was about the classification of diseases in the tomato plant. The study focused on five types of diseases found in tomato and their classification was carried out by extracting various features such as shape, color, and texture from both healthy and infected plant images. The supervised learning technique was used for the classification of the plant leaves and displayed an accuracy of 97.3%.

(Elangovan & Nalini 2017) experimented on plant disease image classification using image segmentation and SVM techniques. The authors discussed various techniques to segment the diseased part of the plant. Also, classification techniques to extract the features of infected leaf and the classification of plant diseases through the SVM classifier were explained and implemented.

(Mwebaze & Owomugisha 2016) proposed a smartphone-based diagnostic system for cassava diseases which deteriorates the crop health. Machine learning was used to solve the problem by identifying disease in the field by just analysing the image of a plant leaf. results obtained were different for various algorithms for identifying the severity of diseases. ORB features (Rublee et al. 2011) were selected and it showed considerable results. But the range of severities used in this study is not incomplete due to limited data at severity 5 but this does impact the practical approach of the study since at that severity a plant cannot be revived.

(Singh & Misra 2017) worked on the paper which presented an algorithm for image segmentation technique that is used for automatic detection and classification of plant leaf diseases. It also covers a survey on different disease classification techniques that can be used for plant leaf disease detection. Banana, beans, jackfruit, lemon, mango, potato, tomato, and sapota are some of those ten species on which the proposed algorithm is tested. With very little computational efforts the optimum results were obtained, which also shows the efficiency of the proposed algorithm in recognition and classification of

the leaf diseases. Another advantage of using this method is that plant diseases can be identified at an early stage or the initial stage. To further improve the classification process artificial neural networks can be used.

2.3 Alternative Methods

(Chanda & Biswas 2019) Acknowledged the different methods available for processing and classifying plant images and proposed a system to carry out this process more efficiently and effectively. The proposed method looked into classification using a back-propagation algorithm to obtain the weights with the help of Particle Swarm Optimization (PSO) (Trelea 2003) to overcome the issues of overfitting. The parameter values of PSO have to be set appropriately since a slight change in one value will a huge impact on performance. The obtained accuracy was 96.72% which is higher than the other referenced works in the paper. The use of PSO with backpropagation displays better accuracy but choosing the initial values for the PSO is a difficult task.

(Es-saady et al. 2016) proposed a system that is a combination of two SVM classifiers to recognize plant leaves diseases and is an enhancement of (El Massi et al. 2015) study. Also, the SVM was used in place of neural networks due to its ease of use and accuracy. The study was carried out in six different classes of plant diseases. Among the 2 classifiers, the first one makes use of color to classify the images. So, here all the diseases with similar colors are classified as one. The second classifier classifies the result from the first stage based on shape and texture. The proposed approach obtained 87.80% accuracy which was higher and efficient compared to (El Massi et al. 2015). This system can further be tested by using a larger dataset.

3 Research Methodology

The main intention of this research is to predict the crop yield to improve food security. Data mining methodologies can be used to discover and analyze different patterns hidden in the data and provide meaningful insights. In this paper, the prediction system is implemented using the Cross-industry process for data mining (CRISP-DM) methodology (Wirth & Hipp 2000).

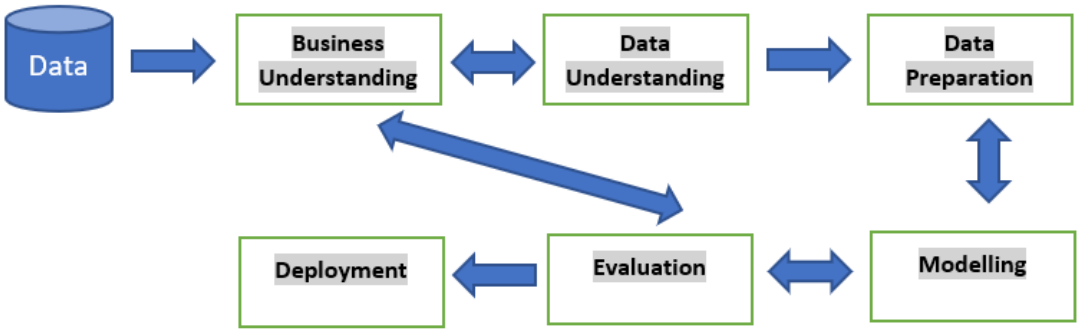

Figure 1: CRISP-DM Methodology

3.1 Business Understanding

In the present world, people are curious to know the outcome of a situation beforehand. It is the same case in the farming community. Since agriculture is an important resource for the economy and survival, it is important to improve the cultivation process to increase the yield. If a yield of a crop is known before cultivation, better decisions can be taken. Policymakers can make better judgments about import and export policies and improve food security.

This research aims at exploring the regression and classifiers in data mining and machine learning to find the best possible method to predict the crop yield and classify pests. For this purpose, a detailed study on the previous research papers was conducted to understand the advantages and disadvantages of available prediction systems for crop yield.

3.2 Data Understanding

1. Data Collection: Datasets are publicly available for research and were collected from Kaggle and both the datasets do not contain any confidential or personal information.
2. Data exploration: All the important variables available in the datasets are duly noted and accordingly data cleaning is carried out.

3.3 Data Preparation

1. Data pre-processing: The datasets consist of null values, misspelled words, abbreviations, etc. These need to take care before modelling.
2. Data Labelling: Data labeling and feature scaling have to be performed to improve the accuracy. Also, the images need to resized to a standard value for better classification results.

3.4 Modeling

1. Class Balancing: Check for any class imbalance issue since it will have a huge impact on the final result.
2. Modeling techniques: With the detailed literature survey, linear regression and decision tree classifier is implemented for the prediction of crop yield and CNN, VGG16, and Inception is implemented on the image dataset.

3.5 Evaluation

The results obtained from different machine learning models are evaluated and compared to find the one with the best accuracy.

3.6 Deployment

The proposed system is cost-free and can be implemented in small and medium-scale farmers where using satellite images or analyzing the nutrient contents of soil is not feasible.

4 Data Procurement

Since one of the objectives of this research is to predict the crop yield data related to agriculture(crops) and rainfall is selected. For the classification of plant diseases, an image dataset with plant leaf diseases is chosen. The data used for this research is outlined in this section

Datasets used:

1. Agriculture data [2]
2. Rainfall data[3]
2. Plant images data [4]

The agriculture(crops) and rainfall data is in comma-separated values (CSV) format and consists of data related to different States in India. The datasets consist of variables such as production, area, States, crops grown and rainfall received.

The plant leaf images dataset is again downloaded from Kaggle and consists of more than 70,295 images of plants belonging to 38 different classes.

Variable Name	Description
State_Name (String)	Name of the State in India
District_Name (String)	District names in the State
Crop_Year (Date)	Year when the crop was grown
Season (String)	Season in which the crop is grown
Crop (String)	Name of the crop grown
Area (Numeric)	Area used to cultivate the crop
Production (Numeric)	Crop produced in a year

Figure 2: Metadata of Agriculture dataset

Variable Name	Description
Subdivision (String)	Name of the State in India
Year (Date)	Year when the rainfall was recorded
Jan – Dec (Numeric)	Rainfall recorded in each month of the year
Annual (Numeric)	Total recorded rainfall in a year

Figure 3: Metadata of Rainfall dataset

4.1 Data Cleaning

4.1.1 Agriculture and Rainfall datasets

The first part of the research makes use of agriculture data and rainfall data to predict crop yield. There are discrepancies in words like 'and' and '', misspelled words, etc. The

[2]https://www.kaggle.com/abhinand05/crop-production-in-india/download
[3]https://www.kaggle.com/rstogi896/rainfall-in-india/download
[4]https://data.mendeley.com/datasets/tywbtsjrjv/1

consistency in data has to be maintained in both agriculture and rainfall dataset

The State names are different in both the datasets and this was taken care of by writing a function.

Karnataka = Coastal_karnataka + North karnataka + south Karnataka

In the above exampple, Costal Karnataka, North Karnataka and South South Karnataka are all the same state 'Karnataka'. Hence all these values needs to be combined and changed to 'Karnataka' alone.

'ANDAMAN NICOBAR ISLANDS' to be changed to Andaman and Nicobar Islands

In one above example, the case of names had to change. In a few cases, two or three States were merged as one. In this situation, individual state names were given to all the combined state values in the rainfall dataset and a function was written to perform the same. For example:

HARYANA DELHI CHANDIGARH was changed to Haryana

A new column (feature) "Rainfall" should be created in the agriculture dataset and the value of this will be the sum of rainfall in a particular season in which the crop is cultivated. For this purpose, a mapping is used for retrieving appropriate rainfall value from the rainfall dataset for each season available in agriculture data.

MONTHS	SEASON
OCTOBER to MARCH	RABI MONTHS
JULY to OCTOBER	KHARIF MONTHS
APRIL TO JUNE	SUMMER MONTHS
ALL	WHOLE YEAR
OCTOBER TO JANUARY	WINTER MONTHS
SEPTEMBER TO NOVEMBER	AUTUMN MONTHS

Figure 4: Months and their respective Seasons

Using the data defined in Figure: 4, new columns are added to the rainfall dataset.

In the next step, all the crops available are grouped into common categories such as pulses, fruits, cereals, etc. This also helps us to visualize the total percentage of crops grown around the country (Figure 5). From Figure 5, we can say that the most commonly grown category of crops is cereals.

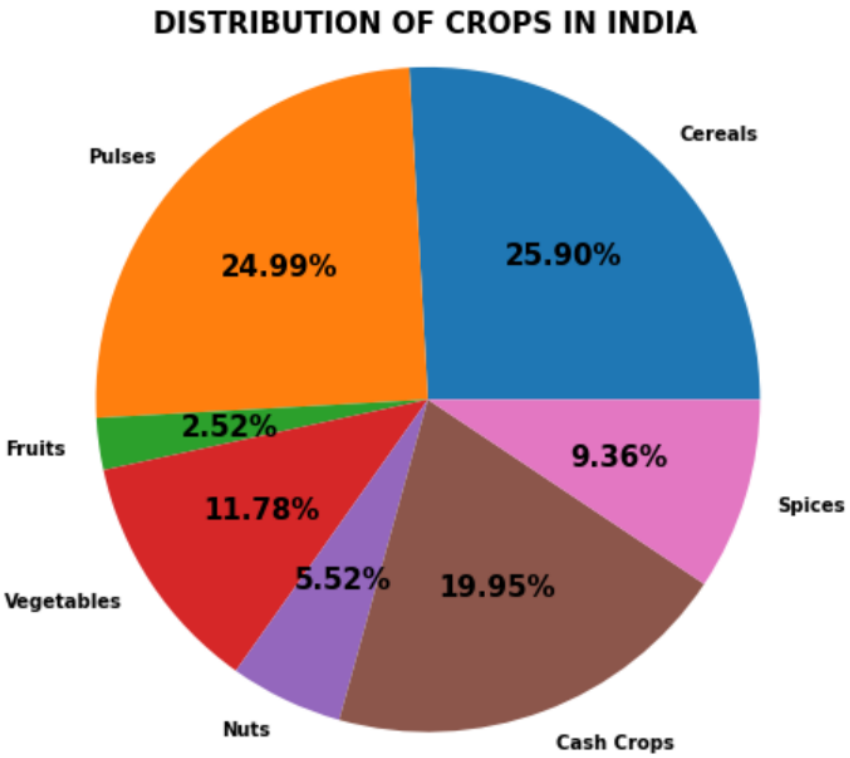

Figure 5: Distribution of crops in India

To convert the categorical values into numeric ones different encoding techniques were used:

1.One Hot Encoding

With the help of One Hot encoding process, it is possible to convert the Categorical variables into a form that can be used in Machine Learning algorithms for better prediction. An example of one-hot encoding is provided in Figure 6.

Gender
Male
Female
Not specified

Male	Female	Not specified
1	0	0
0	1	0
0	0	1

Figure 6: Example of One-Hot Encoding

2. Label Encoding

Using the label encoding process it is possible to convert the labels into numerical form. Reference for label encoding is provided in figure 7.

3. Binary Encoding

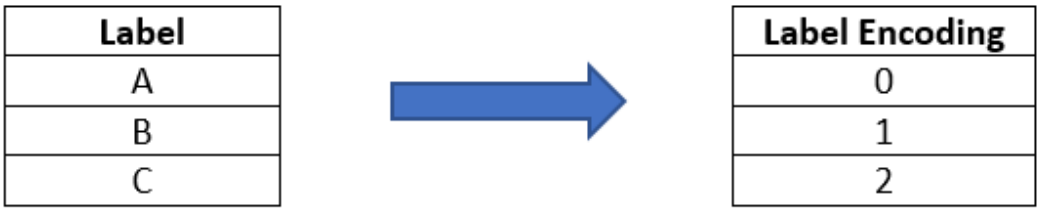

Figure 7: Example of Label Encoding

In binary coding the unit of information is a bit, that is either '0' or '1'.

All the above three encoders were applied on the merged dataset but label encoding was chosen due to its better performance in the research scenario. Along with label encoding, feature scaling is also applied to overcome the problem of number sequencing.

4.1.2 Plant Leaf Image dataset

This image dataset is 1.3GB in size and consists of over 70,295 images. Due to its volume, it is difficult to compute the algorithm in the system and hence the dataset was zipped and loaded into Google drive and then retrieved in Google Colaboratory (Colab) for computation.

Initially, while computation, the plant image dataset from Google drive was directly used for the computation. However, the time required for computation was high, and hence saving the dataset during the run time of Colab made more sense.

For the image dataset, each image is read from the from the directory and is converted into an array of size 256 X 256. The invalid images are removed since they do not contribute to the study.

The images are resized since there is a possibility of Some images varying in size, so to have a standard size all the images are resized.

4.2 Missing values

Missing values in the datasets were identified using msno.bar() function. Rainfall and production columns were found to have missing values and these were imputed using fillna() function with the mean.

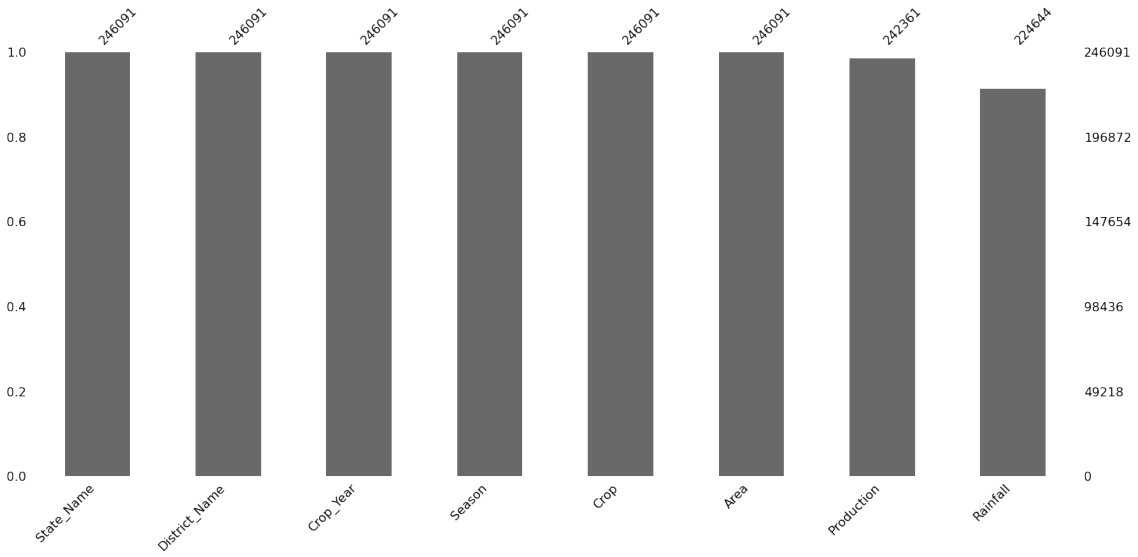

Figure 8: Missing values present in dataset

Figure 8 displays the missing values present in the dataset. As seen, each of the columns have 246091 except production and rainfall which have 242361 and 224644 values respectively and were updated using fillna() function.

5 Implementation

5.1 Prediction and Evaluation of crop yield

5.1.1 Random Forest Regressor

Random forest is one of the commonly used models. It is being used both as a regressor and classifier (Liaw et al. 2002). These are robust and considered to be efficient.

In the research, the target variable is 'production' which is numeric, also the predictors are of a numeric value.

Random Forest algorithm generates multiple decision trees from which each decision tree uses a part of the data sample and predicts the result.

Random Forest Regressor uses a bagging technique and all the trees run in parallel and have no interactions.

The accuracy of Random Forest Regressor is 88.6481 and R^2 value is 0.8864.

5.1.2 Decision Tree Regressor

Decision trees are constructed using an algorithmic approach that identifies ways to split the dataset based on different conditions and predicts the output. It is a non-parametric method.

The numeric variables from agriculture and rainfall dataset were extracted and used to predict the crop yield. The data was divided into train and test set in a ratio of 80:20 respectively.

The accuracy of the Decision Tree Regressor is 71.3110 and R^2 value is 0.7131.

5.1.3 Gradient Boosting Regressor

In Gradient Boosting Regressor weak learners are converted into strong learners. It trains the model in a gradual and sequential manner and thus increasing the accuracy.

The accuracy of Gradient Boosting Regressor is 82.7275 and R^2 value is 0.8272.

	Accuracy	R^2
Random Forest Regressor	88.6481	0.8864
Decision Tree Regressor	71.3110	0.7131
Gradient Boost Regressor	82.7275	0.8272

Figure 9: Result of all the Regressors

Figure 9 displays the accuracy and R^2 values of all the implemented models. Among all the applied models, we can see that Random Forest Regressor achieved the highest accuracy and Decision Tree Regressor had the lowest of accuracy. With this data we can conclude that Random Forest Regressor is efficient with test data in comparison to Decision Tree and Gradient Boost.

The obtained result was compared with the result of previous studies in this filed and was found to be better. Figure 10 displays the result obtained in (Nigam et al. 2019).

MODEL	Accuracy (In Percentage)
Random Forest Classifier	67.80
XGBoost Classifier	63.63
KNN Classifier	43.25
Logistic Regression	25.81

Figure 10: Result of Previous Study

5.2 Implementation, Evaluation and Result of Neural Networks

With the growing population, demand for crops has also increased drastically. Hence it is of most importance to improve the technology in farming. Identification and treatment of diseases in plants have always been a hurdle.

The most important task in pest management is identifying and treating the diseases appropriately. For the identification of pests, neural networks, and deep learning methods can be used effectively. In this paper, 70,295 images which belong to 38 diseases of various plants are classified using various models.

Along with the CNN model, transfer learning models such as VGG16 and InceptionV3 models were also implemented to achieve a model with the best accuracy.

Python syntax is used for this study. Keras which is a high-level API is used as a framework. But since Keras cannot operate by itself and requires a backend for operations, Google's TensorFlow is installed.

Before training any models, a basic connected neural network has to be built for the Plant Village dataset.

1. Image pixels are normalised
2. Encoding the categorical column.
3. Building the model architecture consisting of dense layers.
4. Split the dataset for training and testing purposes.
5. Finally, training the model.

5.2.1 Convolution Neural Network Model

Convolution Neural Networks (CNN) is one of the architectures of artificial neural networks which was proposed by Yann LeCun (LeCun et al. 1988). This architecture is popular for image classification. CNN is used for automatic tagging algorithms by Facebook and Amazon uses it for recommendation systems.

When the image is uploaded, the computer recognizes it as an array of pixels. In this paper, the images are resized to 256x256x3 and so the size of the array will the same size as the size of the image (256x256x3). Here 256x256 is the width and height, 3 is the RGB value.

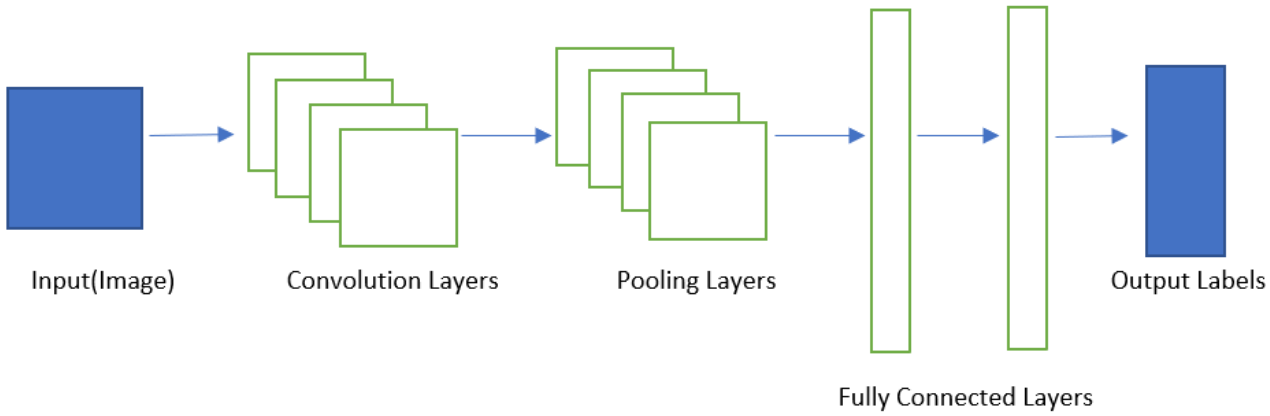

Figure 11: CNN Architecture

In the CNN model, there are multiple layers through which the input is passed before generating the output. Different layers found in this CNN model are convolutional layer, nonlinear layer, pooling layer, and fully connected layer.

1. The first layer found in the CNN model is the convolution layer where the features of an image are extracted. The filter available in this layer generates a matrix similar to the input matrix but smaller.

In general, a network will consist of multiple convolutional layers along with nonlinear and pooling layers. The output of one layer is always the input of the next layer.

2. To every convolution operation, a nonlinear layer is added since it has an activation function to generate a nonlinear property. Without this, the network will be unable to model the response variable.

3. The output of nonlinear layer servers as the input for the pooling layer. Here, the images are downsampled and volume is reduced. In short, it can be said that the images are compressed and are less detailed.

4. After all the above-mentioned layers, a fully connected layer is added. The output of the convolution network serves as the input for this layer. This layer creates an N-dimensional vector where N stands for the number of classes.

5. After the model is set up, it needs to be trained. For this purpose, the data is split into train and test datasets, and training data is used for this purpose.

A model can be tested after the training is complete and for this purpose, test dataset is used. Finally, the model can be used to evaluate new/different data.

The CNN model achieved an accuracy of 97% for the classification after 30 epochs. CNN models work really well while classifying image datasets and is most trusted among the available neural networks.

Figure 12 and Figure 13 displays the plot for loss and accuracy and it can be seen that loss has decreased and accuracy has gradually increased and is stable.

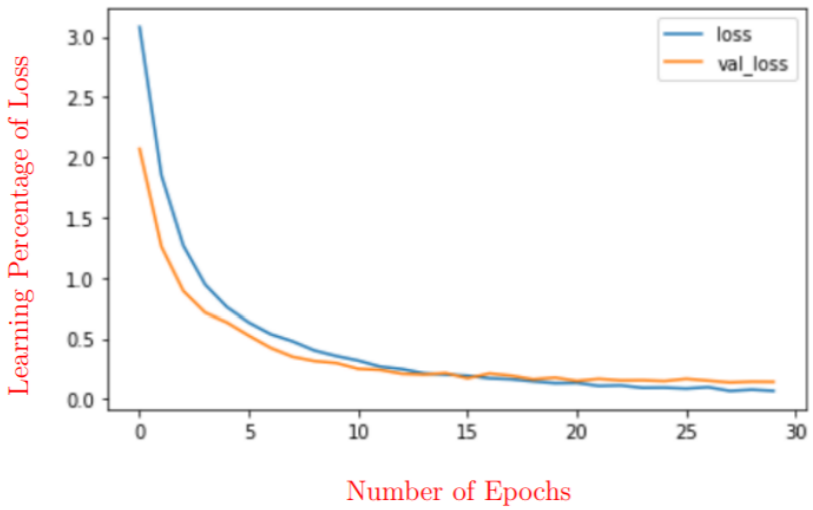

Figure 12: Plot for loss in CNN model

5.2.2 VGG16

VGG16 model was applied on plant image dataset as well. This model has 16 layers and goes deep. This model can penetrate deep.

The best result obtained via VGG16 is

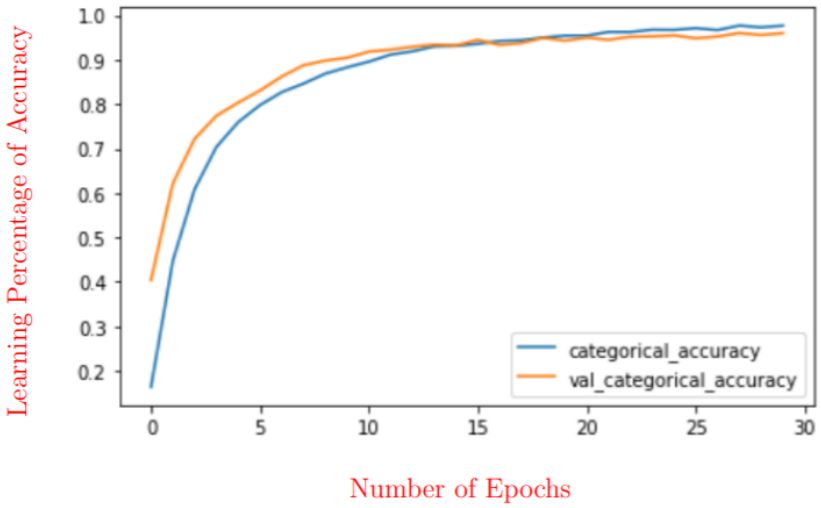

Figure 13: Plot for accuracy in CNN model

5.2.3 InceptionV3

Inception-v3 is popular because of its ability to recognize objects in images. This model is vastly used in medical filed, in Facebook for aiding photo tagging and etc.
Inception-v3 has two stages:

Stage 1: Deals with feature extraction

Stage 2: classification of data (Fully connected and softmax layer)

Inceptionv3 was found to be computationally expensive and hence was not taken into consideration for the research

6 Evaluation

The accuracy of the crop yield prediction model is quite good but since the regression techniques are used, the predicted values will be different from the actual values. The agriculture and rainfall dataset used for this research provides the generic details at the State and District level. For more accurate predictions, precise local data can be used. The accuracy obtained in plant disease classification is high with CNN achieving 97%.

7 Conclusion, Discussion and Future Work

Conclusion

The objective of this research was to predict the crop yield using rainfall data and classify the plant diseases by implementing machine learning algorithms. The business and ethical issues and the available solutions were understood with the help of a literature review.

Random Forest Regressor proved to be better at predicting crop yield and CNN was found to better at classifying plant disease images.

The major drawback here is the unavailability of the precise local data and missing values. The accuracy of the random forest classifier is 88% and CNN is 97%.

Discussion and Future work:

In this research, a new approach was used to predict crop yield using agriculture and rainfall data. However, with more variables, it is possible to understand the influence of various factors in crop yield.

Several methods are available to detect the plant diseases, but still, has room for improvement. In the market, there are several applications to identify the species of plant but not the plant diseases based on leaf images.

In this research, deep learning methodologies were used for the classification and identification of plant diseases from the images of leaves. The final model was successfully able to identify leaf and differentiate the plants with diseases from that of the healthy ones.

For future work, the datasets of vegetables can be added. Along with the classification of diseases, a feature to suggest possible treatment, fertilizers and pesticides can be implemented.

References

Barbedo, J. G. A. (2019), 'Plant disease identification from individual lesions and spots using deep learning', *Biosystems Engineering* **180**, 96–107.

Chanda, M. & Biswas, M. (2019), Plant disease identification and classification using back-propagation neural network with particle swarm optimization, *in* '2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)', IEEE, pp. 1029–1036.

El Massi, I., Saady, Y. E., El Yassa, M., Mammass, D. & Benazoun, A. (2015), Serial combination of two classifiers for automatic recognition of the damages and symptoms on plant leaves, *in* '2015 Third World Conference on Complex Systems (WCCS)', IEEE, pp. 1–6.

Elangovan, K. & Nalini, S. (2017), 'Plant disease classification using image segmentation and svm techniques', *International Journal of Computational Intelligence Research* **13**(7), 1821–1828.

Es-saady, Y., El Massi, I., El Yassa, M., Mammass, D. & Benazoun, A. (2016), Automatic recognition of plant leaves diseases based on serial combination of two svm classifiers, *in* '2016 International Conference on Electrical and Information Technologies (ICEIT)', IEEE, pp. 561–566.

Gandge, Y. et al. (2017), A study on various data mining techniques for crop yield prediction, *in* '2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)', IEEE, pp. 420–423.

Gandhi, N., Armstrong, L. J., Petkar, O. & Tripathy, A. K. (2016), Rice crop yield prediction in india using support vector machines, *in* '2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)', IEEE, pp. 1–5.

Ghadge, R., Kulkarni, J., More, P., Nene, S. & Priya, R. (2018), 'Prediction of crop yield using machine learning', *Int. Res. J. Eng. Technol.(IRJET)* **5**.

Gopal, P. M. & Bhargavi, R. (2019), 'A novel approach for efficient crop yield prediction', *Computers and Electronics in Agriculture* **165**, 104968.

Hecht-Nielsen, R. (1992), Theory of the backpropagation neural network, *in* 'Neural networks for perception', Elsevier, pp. 65–93.

Ji, B., Sun, Y., Yang, S. & Wan, J. (2007), 'Artificial neural networks for rice yield prediction in mountainous regions', *The Journal of Agricultural Science* **145**(3), 249.

Kumar, A., Sarkar, S. & Pradhan, C. (2019), Recommendation system for crop identification and pest control technique in agriculture, *in* '2019 International Conference on Communication and Signal Processing (ICCSP)', IEEE, pp. 0185–0189.

LeCun, Y., Touresky, D., Hinton, G. & Sejnowski, T. (1988), A theoretical framework for back-propagation, *in* 'Proceedings of the 1988 connectionist models summer school', Vol. 1, CMU, Pittsburgh, Pa: Morgan Kaufmann, pp. 21–28.

Liaw, A., Wiener, M. et al. (2002), 'Classification and regression by randomforest', *R news* **2**(3), 18–22.

Miljković, D. (2017), Brief review of self-organizing maps, *in* '2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)', IEEE, pp. 1061–1066.

Mwebaze, E. & Owomugisha, G. (2016), Machine learning for plant disease incidence and severity measurements from leaf images, *in* '2016 15th IEEE international conference on machine learning and applications (ICMLA)', IEEE, pp. 158–163.

Nigam, A., Garg, S., Agrawal, A. & Agrawal, P. (2019), Crop yield prediction using machine learning algorithms, *in* '2019 Fifth International Conference on Image Information Processing (ICIIP)', IEEE, pp. 125–130.

Oliveira, I., Cunha, R. L., Silva, B. & Netto, M. A. (2018), 'A scalable machine learning system for pre-season agriculture yield forecast', *arXiv preprint arXiv:1806.09244* .

Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. (2011), Orb: An efficient alternative to sift or surf, *in* '2011 International conference on computer vision', Ieee, pp. 2564–2571.

Sabrol, H. & Satish, K. (2016), Tomato plant disease classification in digital images using classification tree, *in* '2016 International Conference on Communication and Signal Processing (ICCSP)', IEEE, pp. 1242–1246.

Satir, O. & Berberoglu, S. (2016), 'Crop yield prediction under soil salinity using satellite derived vegetation indices', *Field crops research* **192**, 134–143.

Singh, V. & Misra, A. K. (2017), 'Detection of plant leaf diseases using image segmentation and soft computing techniques', *Information processing in Agriculture* **4**(1), 41–49.

Trelea, I. C. (2003), 'The particle swarm optimization algorithm: convergence analysis and parameter selection', *Information processing letters* **85**(6), 317–325.

Wang, G., Sun, Y. & Wang, J. (2017), 'Automatic image-based plant disease severity estimation using deep learning', *Computational intelligence and neuroscience* **2017**.

Wirth, R. & Hipp, J. (2000), Crisp-dm: Towards a standard process model for data mining, *in* 'Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining', Springer-Verlag London, UK, pp. 29–39.