International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

AGRICULTURAL PRODUCT PRICE AND CROP CULTIVATION PREDICTION BASED ON MACHINE LEARNING TECHNIQUE

Manikandan.U,Kadheer Khan.R,Kumaresan .A

Final year CSE Department - Dhaanish ahmed college of engineering.

ABSTRACT:

Among worldwide, agriculture has the major responsibility for improving the economic contribution of the nation.To prevent this problem agriculture sectors have predict the crop from given data set. Farmers are showing various signs of accepting the modern ways of agriculture. However, still the most agricultural fields are under developed due to the lack of deployment of ecosystem control technologies. Due to these problems, the crop production is not improved which affects the agriculture economy. Hence a development of agricultural productivity prediction. To prevent this problem, Agricultural sectors have to predict the crop from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the best crop. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with entropy calculation, precision, Recall, F1 Score, Sensitivity, Specificity and Entropy.

INTRODUCTION

DATA SCIENCE:

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structuredand unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux. The term "data science" was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market. Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.

Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us

in finding out the hidden insights or patterns from raw data which can be of major use in the formation of big business decisions.

ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving.

Learning processes. This aspect of AI programming focuses on acquiring data and creating rules for how to turn the data into actionable information. The rules, which are called algorithms,provide computing devices with step-by-step instructions for how to complete a specific task.

Reasoning processes. This aspect of AI programming focuses on choosing the right algorithm to reach a desired outcome.

Self-correction processes. This aspect of AI programming is designed to continually fine-tune algorithms and ensure they provide the most accurate results possible. AI is important because it can give enterprises insights into their operations that they may not have been aware of previously and because, in some cases, AI can perform tasks better than humans. Particularly when it comes to repetitive, detail-oriented tasks like analyzing large numbers of legal documents to ensure relevant fields are filled in properly.

Natural Language Processing (NLP):

Natural language processing (NLP) allows machines to read and understand human language. A sufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. strategies use the occurrence of words such as "accident" to assess the sentiment of a document. Modern statistical NLP approaches can combine all these strategies as well as others, and often achieve acceptable accuracy at the page or paragraph level. Beyond semantic NLP, the ultimate goal of "narrative" NLP is to embody a full understanding of commonsense reasoning. By 2019, transformer-based deep learning architectures could generate coherent text.

OBJECTIVES

The goal is to develop a machine learning model for Crop yield Prediction, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm.

LITERATURE SURVEY

General

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an

organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them

Review of Literature Survey Title : Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics

Author: Douglas K. Bolton∗ , Mark A. Friedl

Year : 2013

We used data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) in association with county-level data from the United States Department of Agriculture (USDA) to develop empirical models predicting maize and soybean yield in the Central United States. As part of our analysis we also tested the ability of MODIS to capture inter-annual variability in yields. Our results show that the MODIS two-band EnhancedVegetation Index (EVI2) provides a better basis for predicting maize yields relative to the widely used Normalized Difference Vegetation Index (NDVI). Surprisingly, using moderate spatial resolution data from the MODIS Land Cover Type product to identify agricultural areas did not degrade model results relative to using higher-spatial resolution crop-type maps developed by the USDA. Correlations between vegetation indices and yield were highest 65–75 days after greenup for maize and 80 days after greenup for soybeans. EVI2 was the best index for predicting maize yield in non-semi-arid counties (R2 = 0.67), but the Normalized Difference Water Index (NDWI) performed better in semi-arid counties (R2 = 0.69), probably because the NDWI is sensitive to irrigation in semi-arid areas with low-density agriculture. NDVI and EVI2 performed equally well predicting soybean yield (R2 = 0.69 and 0.70, respectively). In addition, EVI2 was best able to capture large negative anomalies in maize yield in 2005 (R2 = 0.73). Overall, our results show that using crop phenology and a combination of EVI2 and NDWI have significant benefit for remote sensing-based maize and soybean yield models.

Title : Crop Yield Assessment from Remote Sensing

Author: Paul C. Doraiswamy, Sophie Moulin, Paul W. Cook, and Alan Stern

A model calibration was performed to initialize the model parameters. This calibration was performed using Landsat data over three southeast counties in North Dakota. The model was then used to simulate crop yields for the state of North Dakota with inputs derived from NOAA AVHRR data. The calibration and the state level simulations are compared with spring wheat yields reported by NASS objective yield surveys.

Title : A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data

Author: . Becker-Reshef a, ∗, E. Vermote a , M. Lindeman b , C. Justice

Year : 2010

Wheat is one of the key cereal crops grown worldwide, providing the primary caloric and nutritional source for millions of people around the world. In order to ensure food security and sound, actionable mitigation strategies and policies for management of food shortages, timely and accurate estimates of global crop production are essential.

This study combines a new BRDF-corrected, daily surface reflectance dataset developed from NASA's Moderate resolution Imaging Spectro-radiometer (MODIS) with detailed official crop statistics to develop an empirical, generalized approach to forecast wheat yields. The first step of this study was to develop and evaluate a regression-based model for forecasting winter wheat production in Kansas. This regression-based model was then directly applied to forecast winter wheat production in Ukraine. The forecasts of production in Kansas closely matched the USDA/NASS reported numbers with a 7% error. The same regression model forecast winter wheat production in Ukraine within 10% of the official reported production numbers six weeks prior to harvest. Using new data from MODIS, this method is simple, has limited data requirements, and can provide an indication of winter wheat production shortfalls and surplus prior to harvest in regions where minimal ground data is available.

Title : Plant Yield Prediction Model Using Firefly based Feature Selection with Modified Fuzzy Cognitive Maps
Author: 1D. Sabareeswaran and 2R. Gunasundari

Among worldwide, agriculture has the major responsibility for improving the economic contribution of the nation. However, still the most agricultural fields are under developed due to the lack of deployment of ecosystem control technologies. Hence in this paper, a development of agricultural productivity is enhanced based on the plant yield prediction. Initially, different features such as plant images, soil characteristics, and weather factors are gathered and Firefly (FF) optimization algorithm is proposed for Feature Selection (FFFS).

Title : Crop yield forecasting on the Canadian Prairies using MODIS NDVI data
Author: M.S. Mkhabelaa,∗ , P. Bullocka, S. Raj b, S. Wangc, Y. Yang

Year : 2010

Although Normalised Difference Vegetation Index (NDVI) data derived from the advanced very high resolution radiometer (AVHRR) sensor have been extensively used to assess crop condition and yield on the Canadian Prairies and elsewhere, NDVI data derived from the new moderate resolution imaging spectroradiometer (MODIS) sensor have so far not been used for crop yield prediction on the Canadian Prairies. Therefore, the objective of this study was to evaluate the possibility of using MODIS-NDVI to forecast crop yield on the Canadian Prairies and also to identify the best time for making a reliable crop yield forecast.

EXISTING SYSTEM:

Crop cultivation prediction is an integral part of agriculture and is primarily based on factors such as soil, environmental features like rainfall and temperature, and the quantum of fertilizer used, particularly nitrogen and phosphorus. These factors, however, vary from region to region: consequently, farmers are unable to cultivate similar crops in every region. Predicting a suitable crop for cultivation is critical to agriculture. In this work, the MRFE, a novel approach, has been proposed for selecting salient features using a permutation crop data set and a ranking method to identify the most suitable crop for a particular region.

PROPOSED SYSTEM

Exploratory Data Analysis

In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis. Make sure that you document your steps carefully and justify your cleaning decisions.

TRAINING THE DATASET

- For example, to import any algorithm and train_test_split class from sklearn and numpy module for use in this program.
- Then we encapsulate load_data() method in data_dataset variable. Further we divide the dataset into training data and test data using train_test_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.

TESTING THE DATASET

- Now we have dimensions of a new flower in a numpy array called 'n' and we want to predict the species of this flower. We do this using the predict method which takes this array as input and spits out predicted target value as output.
- So the predicted target value comes out to be 0. Finally we find the test score which is the ratio of no. of predictions found correct and total predictions made. We do this using the score method which basically compares the actual values of the test set with the predicted values.

ENVIRONMENTAL REQUIREMENTS

1. Software Requirements:

Operating System : Windows

Tool :Anaconda with Jupyter Notebook

2. Hardware requirements:

Processor : Pentium IV/III

Hard disk : minimum 80 GB

RAM : minimum 2 GB

MODULE DESCRIPTION:

DATA PRE-PROCESSING

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer.

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers use this data to fine-tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model.

- User forgot to fill in a field.

- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be used or interpreted.

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- import libraries for access and functional purpose and read the given dataset
- General Properties of Analyzing the given dataset
- Display the given dataset in the form of data frame
- show columns
- shape of the data frame
- To describe the data frame
- Checking data type and information about dataset
- Checking for duplicate data
- Checking Missing values of data frame
- Checking unique values of data frame
- Checking count values of data frame
- Rename and drop the given data frame
- To specify the type of values
- To create extra columns

Data Validation/ Cleaning/Preparing Process

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

	State_Name	District_Name	Crop_Year	Season	Crop	Area	rainfall	Average Humidity	Mean Temp	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)	cost of production per yield
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	0.012360	57	62	23076.74	1941.55	9.83	19085.4365
1	Andaman and Nicobar Islands	NICOBARS	2001	Kharif	Arecanut	1254.0	0.084119	56	58	12610.85	1691.66	6.83	11554.0378
2	Andaman and Nicobar Islands	NICOBARS	2002	Whole Year	Arecanut	1258.0	0.080064	58	53	32683.46	3207.35	9.33	29924.5755
3	Andaman and Nicobar Islands	NICOBARS	2003	Whole Year	Arecanut	1261.0	0.181051	57	58	13209.32	2228.97	5.90	13150.9230
4	Andaman and Nicobar Islands	NICOBARS	2004	Whole Year	Arecanut	1264.7	0.035446	63	67	22560.30	1595.56	13.57	21651.7492

	Crop_Year	Area	rainfall	Average Humidity	Mean Temp	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)	cost of production per yield
Crop_Year	1.000000	-0.033996	0.001955	-0.012323	0.053708	-0.008896	0.004706	0.017974	0.006590
Area	-0.033996	1.000000	-0.020667	0.007052	0.002213	0.001313	0.004999	-0.004052	-0.001414
rainfall	0.001955	-0.020667	1.000000	-0.009805	-0.311162	-0.030486	-0.014171	0.069267	-0.046942
Average Humidity	-0.012323	0.007052	-0.009805	1.000000	-0.252588	-0.009753	0.030869	-0.131063	-0.090237
Mean Temp	0.053708	0.002213	-0.311162	-0.252588	1.000000	0.098394	-0.041967	0.057664	0.015582
Cost of Cultivation (`/Hectare) C2	-0.008896	0.001313	-0.030486	-0.009753	0.098394	1.000000	0.025679	0.028143	-0.055532
Cost of Production (`/Quintal) C2	0.004706	0.004999	-0.014171	0.030869	-0.041967	0.025679	1.000000	-0.308414	-0.056326
Yield (Quintal/ Hectare)	0.017974	-0.004052	0.069267	-0.131063	0.057664	0.028143	-0.308414	1.000000	0.743026
cost of production per yield	0.006590	-0.001414	-0.046942	-0.090237	0.015582	-0.055532	-0.056326	0.743026	1.000000

MODULE DIAGRAM

GIVEN INPUT EXPECTED OUTPUT

input : data

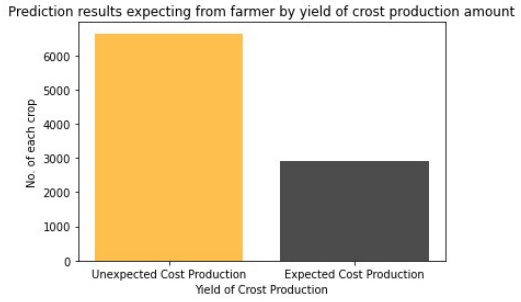
output : removing noisy data

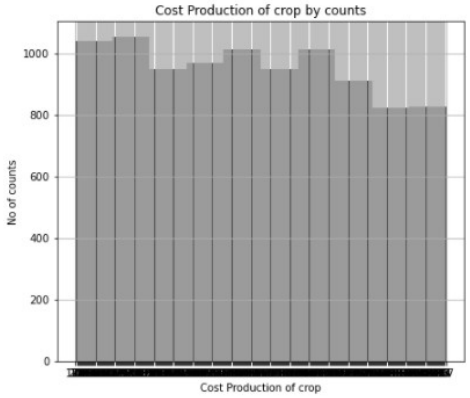
EXPLORATION DATA ANALYSIS OF VISUALIZATION

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative understanding. This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little

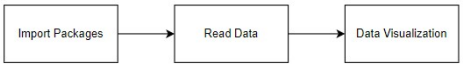
domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots.





MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data

output : visualized data

COMPARING ALGORITHM WITH PREDICTION IN THE FORM OF BEST ACCURACY RESULT

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. In the next section you will discover exactly how you can do that in Python with scikit-learn. The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data and it can achieve this by forcing each algorithm to be evaluated on a consistent test harness.

In the example below 4 different algorithms are compared:

- Random Forest
- Decision Tree Classifier
- Naive Bayes

The K-fold cross validation procedure is used to evaluate each algorithm, importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithm is evaluated in precisely the same way. Before that comparing algorithm, Building a Machine Learning Model using install Scikit-Learn libraries. In this library package have to done preprocessing, linear model with logistic regression method, cross validating by KFold method, ensemble with random forest method and tree with decision tree classifier. Additionally, splitting the

train set and test set. To predicting the result by comparing accuracy.

Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. It need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no.

True Positive Rate(TPR) = TP / (TP + FN)

False Positive rate(FPR) = FP / (FP + TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct.

Precision = TP / (TP + FP)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

Recall = TP / (TP + FN)

Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are

very different, it's better to look at both Precision and Recall.

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

ALGORITHM AND TECHNIQUES

Algorithm Explanation

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

Used Python Packages:

sklearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.
- Here, we are using some of its modules like train_test_split, DecisionTreeClassifier or Logistic Regression and accuracy_score.

NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

Decision Tree Classifier:

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

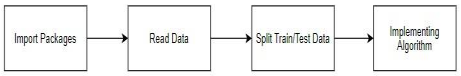
- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision

node has two or more branches and a leaf node represents a classification or decision

```
Classification report of Decision Tree Classifier Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1776
           1       1.00      1.00      1.00      1087

    accuracy                           1.00      2863
   macro avg       1.00      1.00      1.00      2863
weighted avg       1.00      1.00      1.00      2863

Accuracy result of Decision Tree Classifier is 100.0

Confusion Matrix result of Decision Tree Classifier is:
 [[1776    0]
 [   0 1087]]

Sensitivity :  1.0

Specificity :  1.0
Cross validation test results of accuracy:
[1. 1. 1. 1. 1.]
```

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data

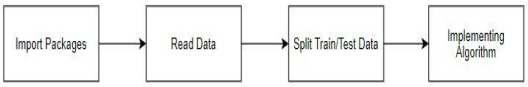
output : getting accuracy

Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

```
Classification report of Random Forest Results:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1776
           1       1.00      1.00      1.00      1087

    accuracy                           1.00      2863
   macro avg       1.00      1.00      1.00      2863
weighted avg       1.00      1.00      1.00      2863

Accuracy result of Random Forest is: 100.0

Confusion Matrix result of Random Forest is:
 [[1776    0]
 [   0 1087]]

Sensitivity :  1.0

Specificity :  1.0

Cross validation test results of accuracy:
[1. 1. 1. 1. 1.]
```

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data

output : getting accuracy

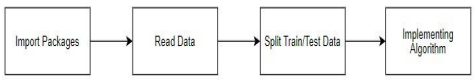
Naive Bayes algorithm:

- The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically.
- Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

- The probability of a class value given a value of an attribute is called the conditional probability. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.
- Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

```
Classification report of Naive Bayes Results:

              precision    recall  f1-score   support

           0       0.95      1.00      0.97      1776
           1       1.00      0.91      0.95      1087

    accuracy                           0.97      2863
   macro avg       0.97      0.95      0.96      2863
weighted avg       0.97      0.97      0.96      2863

Accuracy result of Naive Bayes is: 96.50716032134125

Confusion Matrix result of Naive Bayes is:
 [[1776    0]
 [ 100  987]]

Sensitivity :  1.0

Specificity :  0.9080036798528058
```

MODULE DIAGRAM

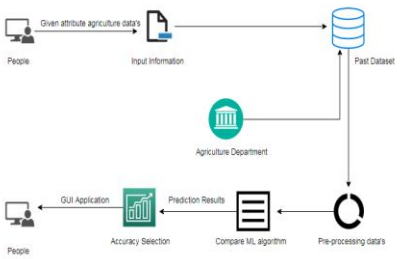


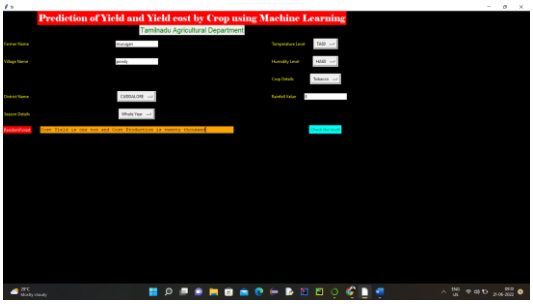
GIVEN INPUT EXPECTED OUTPUT

input : data

output : getting accuracy

SYSTEM ARCHITECTURE:





CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finally we predict the crop using machine learning algorithm with different results. This brings some of the following insights about crop prediction. As maximum types of crops will be covered under this system, farmer may get to know about the crop which may never have been cultivated and lists out all possible crops, it helps the farmer in decision making of which crop to cultivate. Also, this system takes into consideration the past production of data which will help the farmer get insight into the demand and the cost of various crops in market.

REFERENCES

[1] A. Mark Hall, "Feature selection for discrete and numeric class machine learning," *Comput. Sci., Univ. Waikato*, pp. 359–366, Dec.

[4] P. S. Maya Gopal and R. Bhargavi, "Feature selection for yield prediction in boruta algorithm," *Int. J. Pure Appl. Math.*, vol. 118, no. 22, pp. 139–144, 2018.

[5] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Trans. Comput. Social Syst.*, vol. 8, no. 1, pp. 214–226, Feb. 2021.

[6] K.

[7] H.

[8] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.

[9] H. Wang, M. Taghi Khoshgoftaar, and K. Gao, "Ensemble feature selection technique for software quality classification," in *Proc. 22nd Int. Conf. Softw. Eng. Knowl. Eng.*, 2010, pp. 215–220.

[10] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statist. Comput.*, vol. 27, no. 3, pp. 659–678, May 2017.