How to Cite:
Saxena, R., Peter, J. S. P., Saxena, S., & Sapra, A. (2022). Crop price and yield prediction using data science technique. *International Journal of Health Sciences*, *6*(S3), 4777–4793. https://doi.org/10.53730/ijhs.v6nS3.6953

Crop price and yield prediction using data science technique

Ritwik Saxena
Dept of Computer Science and Engineering, SRM Insitute of Science and Technology, Kattankulathur
Email: rs3624@srmist.edu.in

J. Selvin Paul Peter
Associate Professor, SRM Insitute of Science and Technology, Kattankulathur
Email: selvinpj@srmist.edu.in

Shwetha Saxena
Assistant Professor, Amity Business School, MP
Email: ssaxena8@gwl.amity.edu

Anubhav Sapra
Dept of Computer Science and Engineering, SRM Insitute of Science and Technology, Kattankulathur
Email: as2549@srmist.edu.in

***Abstract*---**Agriculture is primarily responsible for increasing the state's economic contribution around the world. The most significant agricultural fields, however, remain underdeveloped because of the absence of ecosystem control technology adoption. Crop output is not improving as a result of these issues, which has an impact on the farm economy. As a result, the plant yield prediction helps to support the growth of agricultural productivity. To address this issue, agricultural industries must use machine learning algorithms to forecast crop yield from a given dataset. In order to capture various pieces of information, the supervised machine learning technique must be employed to analyse the dataset. Some examples include variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments, and so on. A comparison of machine learning algorithms was performed to see which one was the most accurate at forecasting the most basic crop. The results reveal that the effectiveness of the proposed machine learning algorithm technique is commonly compared to the best accuracy using entropy calculation, precision, recall, F1 Score, sensitivity, and specificity.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 18 Jan 2021, Manuscript revised: 09 May 2022, Accepted for publication: 27 April 2022

Keywords---crop price, yield prediction, data science technique, univariate analysis.

Introduction

Agricultural study has bolstered the global economy. It is a neighbourhood that provides incalculable advantages to mankind as a whole. Agricultural crop prediction continues to be difficult, despite recent advancements that includes the use of a variety of high-tech resources, techniques, and processes. Agricultural technologies and site-specific farming are two new scientific fields that use data demanding ways to improve agricultural productivity while reducing environmental effect. Accurately identifying crops for cultivation based on soil and environmental conditions is crucial for agricultural output and has long been a hot research topic. Machine learning is used in most current crop yield estimating systems, However, there has been little or no research into predicting certain crops for a given area based on soil and environmental conditions. Crop cultivation is dominated by factors such as type of soil, nutrients, micronutrients, temperature and rainfall. Because the parameters differ for each zone, resulting in a large crop prediction data set, it's necessary to choose critical elements that aid in identifying acceptable crops for certain geographical locations. Feature selection approaches are used to administer the procedure.

In the field of prediction, machine learning algorithms play a significant role. Feature selection approaches are used to reduce overfitting and choose significant features from the data set for the prediction process, which has improved machine learning performance. Filter, wrapper, and embedding are three types of feature selection strategies. Filter strategies are unaffected by the classifier's performance, whereas wrapper methods choose characteristics that aid the classifier's performance. This research focuses on wrapper feature selection strategies in particular. To predict an appropriate crop to be cultivated and appraise the effectiveness of the feature selection proceeding, the features selected are supplied to the naive bayes, decision tree and random forest classifiers. The goal of this research is to choose critical elements from a knowledge collection in order to increase crop prediction accuracy. The ultimate benefaction of this research is to advance a completely new modified technique for recursive feature elimination for selecting the apt key features using permeation data of crop supported by soil type and environment. Because the algorithm does not need an update with data set at any iteration when using permutation data sets, it takes less time to compute than existing recursive feature elimination methods.

Review of literature survey

Data from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) was combined with county-level data from the US Department of Agriculture to construct empirical models forecasting maize and soybean yield in the Central United States (USDA). As part of our research, we looked into the ability of MODIS to capture inter-annual yield fluctuations. Our findings show that the MODIS

two-band Enhanced Vegetation Index (EVI2) provides a significantly stronger basis for predicting maize yields than the commonly used Normalized Difference Vegetation Index (NDVI). The incorporation of crop phenology data from MODIS improved model performance both within and between years. Surprisingly, employing MODIS Land Cover Type data with intermediate spatial resolution to identify agricultural areas had little effect on model performance when compared to USDA crop-type maps with higher geographic resolution. 65 to 75 days after maize greenup and 80 days after soybean greenup, the highest associations between vegetative indices and yield were observed. The EVI2 was the simplest 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), owing to the fact that the normalised difference water index is sensitive to irrigation in semi-arid areas with low density agriculture. The enhanced vegetation index and the normalised differential water index both predicted soybean yield as well (R2 = 0.69 and 0.70, respectively.

Overall, our findings show that using crop phenology and a combination of EVI2 and NDWI in maize and soybean yield models based on remote sensing has a significant impact. Crop yields can also be assessed via remote sensing. Crop condition and output projections at the state and county levels are of great importance to the US Department of Agriculture. The National Agricultural Statistical Service of the United States Department of Agriculture conducts field interviews with sampled farm operators and collects crop cuttings to estimate crop yields at the regional and state levels. Additional geographic data is needed by the National Agricultural Statistics Service (NASS) to provide timely information on crop status and potential yields. Throughout this work, the crop model EPIC (Erosion Productivity Impact Calculator) was extended for simulations at regional scales. This study looks into using remotely sensed data from satellites to analyse the size and variance of crop condition parameters in real time, and it also looks into using those characteristics as an intake to a crop production model. This disquisition was carried out in North Dakota's desert region and in the state's southeast portion. The first goal was to figure out how to include attributes from satellite photos into a crop growth model to simulate spring wheat yields at the sub-county and county levels. During the season, the input parameters gathered from remotely sensed data provided geographic integrity as well as real-time calibration of model simulation parameters to guarantee that projected and observed circumstances were in agreement. The link between the satellite data and the crop model was provided by a radiative transfer model.

A geographic data system grid was used to verify the model parameters, and it functioned as a platform for aggregating yields at the local and regional levels. The model parameters were set up using a model calibration. Landsat data from over three southeast counties in North Dakota were used for this calibration. With inputs collected from NOAA AVHRR data, the model was then used to simulate crop yields for the state of North Dakota. However, due to a lack of installed ecosystem control systems, agricultural fields remain underdeveloped. Crop production isn't improving as a result of issue 6, which has an impact on the farm economy. Hence during this paper, a development of agricultural productivity is enhanced supported the plant yield prediction. In the initial step Firefly optimization algorithm is proposed for Feature Selection (FFFS) on the

gathered data of different features like plant images, soil characteristics and weather factors. Then, the foremost selected optimal features are classified supported the Modified Fuzzy Cognitive Map (MFCM) algorithm for predicting the expansion of plant yield. the anticipated outcome is transmitted to the farmer's through smart phones which helps for identifying the expansion of plant and improving the harvesting. The experimental results show that the effectiveness of the proposed technique is often compared with the opposite prediction techniques.

Existing system

Prediction of crop cultivation is an important component of agricultural process, and it depends on characteristics example soil, rainfall and temperature, and thus the amount of fertiliser applied, notably phosphorus and nitrogen. Cultivators are incapable to plant akin crops in every place due to differences in these elements. Agriculture relies heavily on the ability to predict the best crop to grow. This paper proposes the MRFE, an unique technique for electing alienates that employs a permutation crop data set and a ranking mechanism to determine the best crop for a given location.

- Drawbacks of the Existing System
- In the current crop yield and agricultural production cost forecast system, no machine learning or deep learning concepts are applied.
- In the existing system for the prediction of crop yield and cost for the crop production any metrics reports were not mentioned.

Proposed system

To increase the accuracy and for the optimization of the existing system of crop yield and crop production cost prediction, a better and optimized system is proposed which includes the concepts of Data Science technique and implement different ML and DL concepts for providing the better accuracy.

Exploratory data analysis

In this section initially data is loaded then it is governed for cleanliness then dataset is beautified for analysis. Ascertain that all actions are meticulously documented and that cleaning decisions are justified.

Training the dataset

- To run this application, we will need the sklearn train test split class and the numpy module.
- The load data () method is thus encapsulated in the data dataset variable.
- The dataset is further divided into training data and testing data using the train test split approach.
- The y prefix in variable denotes the goal values, while the y prefix denotes the feature values.
- The dataset is randomly divided into training and testing data at a 67:33 ratio in this technique. Then any algorithm can be encapsulated.

- The training data is then fed into this method in the next line, so that the machine can be trained using this data. Thus, the training phase is now completed.

Testing of dataset

- We've got the measurements of a replacement flower in a numpy array named 'n,' and we want to figure out what flower family it belongs to.
- This is accomplished by utilising the prediction method, which accepts an array as input and returns a forthcoming aimed value as an output.
- As a result, the estimated goal value is zero.
- Finally, the test scores are found to be the ratio of right guesses to total predictions.
- This is accomplished by the score method, which compares the test set's specific values to the expected values.

Advantages of the proposed system

- The idea is to use the machine's estimates to help farmers and the government. These magazines claim to have surpassed their contenders, yet there is not a single article or public acknowledgement that their work has helped farmers. If there are any major obstacles to progressing that task to the next phase, identify them and find a solution.
- It is aimed at cultivators who want to handle their farm professionally by monitoring, planning and analysing all agricultural ex.
- The major advantage of using the proposed system for crop yield and crop production cost over the existing system is that it is more accurate and optimized as it consists of the concepts of Data Science techniques and implements various Machine learning and deep learning concepts to provide better accuracy.

System study
Objectives

The goal is to develop a crop yield prediction machine learning model which can eventually replace the supervised machine learning classification models by predicting the most accurate shape using supervised algorithms.

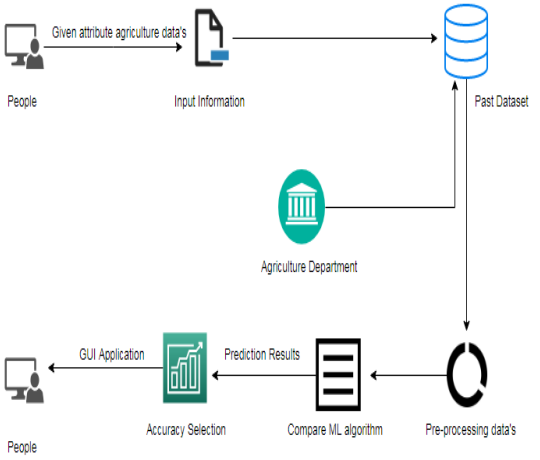
Project goals

- Variable identification data exploration and analysis.
- Analyzing data in a single variable
- Data exploration analysis, both bi-variate and multi-variate.
- Outlier detection method based on feature engineering.
- To forecast the outcome, a comparison algorithm is used.

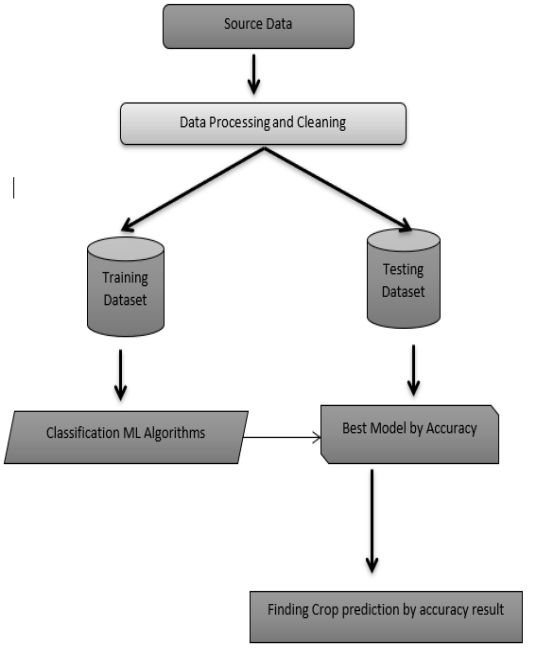
Scope of the project

The goal of the study is to apply machine learning techniques to examine a dataset of crop record for the agriculture sector. Farmers find it more difficult to forecast agricultural yields. In crop selection, we're seeking to reduce the risk factor.

System architecture



Work flow diagram



Implementation
Data recoding

Machine learning validation procedures are used to determine the Machine Learning (ML) model's error rate, which can be accounted for as the dataset's on-the-edge-of-truth error rate. Validation procedures are unnecessary if the data volume is large enough to be assigned to a population. In real-world circumstances, However, when dealing with knowledge selections that aren't genuinely reflective of a dataset's copulation. To locate missing values, duplicate values, and determine whether the variable is a float or an integer. Before altering the hyper parameters, the data sample enables for an unbiased evaluation of a model that has been trained on the training dataset. The evaluation is more skewed since competence on the validation dataset is incorporated into the model design. Although it is often changed on a regular basis, the validation set is moved to analyse a model. ML experts use this information to fine-tune the model's excitable parameters. Knowing your information and its properties can aid you in deciding the algorithm to utilise to develop your model throughout the data identification method.

Using Python's Pandas library, a variety of various data beautifying activities are performed, with a focus on the most important data beautifying work, missing values, and it is ready to clean data more quickly. It would rather spend less time beautifying information and more time understanding and modifying it. Some of them are simply inadvertent errors. Other times, there is a more serious reason for data missing. Understanding the various sorts of missing data is crucial from a statistical approach. The kind of mislaid information determines how it is stuffed in, how mislaid values are identified, and how basic inference and statistical methods are wont to deal with lost information. Before incorporating missing data into code, it's critical to understand where it came from.

Data cleaning process

Importing library packages and loading the specified dataset. Analyze variable identification using data shapes, data types, and missing and duplicate value evaluation. A substantial dataset is a trial of data from your model's training that is used to assess model skill while developing models and procedures. Validation and test datasets can be used to evaluate your models in the most straightforward way feasible. Data cleansing, preparation by renaming the given dataset and eliminating the column, and other steps are required to examine the multi-variate, bi-variate and uni-variate processes. Depending on the dataset, different cleaning procedures and approaches will be used. Data cleaning's first purpose is to find and correct mistakes and abnormalities in data so that it can be used for 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

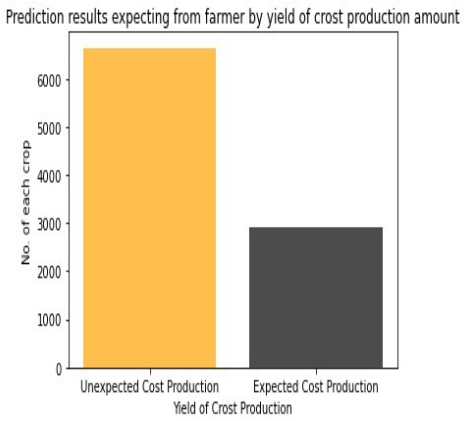
	Crop_Year	Area	rainfall	Average Humidity	Mean Temp	Cost of Cultivation (`/Hectare) C2	Cost of Production (`/Quintal) C2	Yield (Quintal/ Hectare)	cost of production per yield
Crop_Year	1.000000	-0.033998	0.001955	-0.012323	0.053708	-0.008896	0.004706	0.017974	0.006590
Area	-0.033998	1.000000	-0.020967	0.007052	0.002213	0.001313	0.004999	-0.004052	-0.001414
rainfall	0.001955	-0.020967	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

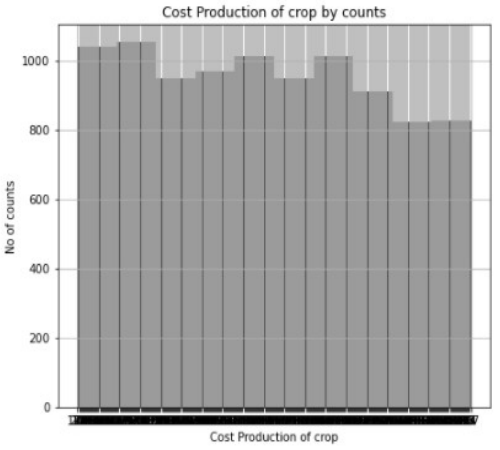
Visualization data exploration and analysis

Data visualisation is a crucial skill in applied statistics and machine learning. Data visualisation is an important tool for acquiring a qualitative understanding of data. This will be useful for exploring and learning about a dataset, and it may aid in the detection of patterns, corrupt data, outliers, and other issues. Data visualisations, rather than correlation or denotation, can be utilised to describe and demonstrate significant relationships in more intuitive plots and charts for stakeholders. Data visualisation and preparatory data analysis are two separate topics, and some of the books listed above are well worth reading in depth.

It's possible that data won't add up until it's shown in a visual format, such as charts and plots. Both applied statistics and applied machine learning require the ability to quickly visualise knowledge samples and other data. It will show you how to use the many sorts of plots we'll need when visualising data in Python, as well as how to use them to better understand your own data.

- How to use line plots to visualise statistical data and bar charts to visualise categorical data.
- How to use histograms and box plots to summarise data distributions.





Result prediction by accuracy of different algorithms

It will be learned how to use scikit-learn to build a test harness in Python to match various different machine learning algorithms. It's vital to match the performance of numerous different machine learning algorithms consistently. This test harness can be used as a framework for your own machine learning issues, with more and alternative algorithms added as needed. The performance characteristics of each model will vary. You may gain an approximation of how accurate each model could be on unseen data by using resampling methods like cross validation. It must be prepared to use these estimations to select one or two of the best models from the batch you've built. When working with a new dataset, it is a good notion to look at the data using several methodologies in order to examine it from many angles. A similar concept can be applied to model selection. To choose one or two to complete, you should employ a range of different approaches to monitor the predicted accuracy of your machine learning algorithms. Multiple visualisation approaches are widely used to highlight the typical accuracy, variance, and other elements of the model accuracies distribution. Ensure that each technique is tested in the same way on comparable data, which can be done by requiring that each algorithm be evaluated on a uniform test harness. Three different algorithms which are compared:

- Random Forest
- Decision Tree Classifier
- Naive Bayes

To test each algorithm, the K-fold cross validation technique is utilized, which uses a comparable random seed to ensure that identical divides to the training data are formed and any algorithm is valuated in the similar manner. Install Scikit-Learn libraries and build a Machine Learning Model before comparing the methods. This library package requires preprocessing, a linear model using the logistic regression method, cross validation using the KFold method, an ensemble using the random forest method, and a tree using the decision tree classifier. Separating the plaything and the test set is also a good idea. One can forecast the outcome by comparing accuracy.

Prediction of result by accuracy

To forecast a value, the logistic regression technique employs an equation with independent predictors. The expected value is frequently in the negative infinity to positive infinity range. In order to classify variable data, the algorithm's output is required. By comparing the simplest accuracy, the logistic regression model has a higher accuracy in predicting the outcome. False Positives (FP): In order to classify variable data, it requires the algorithm's output. When compared to the simplest accuracy model, the logistic regression model has a better accuracy in predicting the outcome.

- False Negatives (FN): a person who is expected to default as a payer When the actual value of a class is yes but the anticipated value is no. For example, if the traveler's real class value suggests that he or she will live whereas the anticipated class implies that the passenger will die.
- True Positives (TP): a person who refuses to pay is known as a defaulter. These are the correctly predicted positive values, meaning that the true class value is yes, and so the predicted class value is yes as well. For example, if the actual class value indicates that this passenger survived and the expected class value also suggests that this passenger survived.
- True Negatives (TN): a person who is expected to default as a payer These are the accurately predicted negative values, implying that the actual class is worthless and the projected class is also worthless. For example, if the actual class indicates that this passenger did not survive and the anticipated class indicates the same.
- 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)

The most straightforward performance metric is accuracy, which is just the ratio of correctly predicted observations to all observations. If our model is accurate, one might conclude that it is the best. Yes, accuracy is a useful statistic, but only if the datasets are symmetric and the false positive and false negative outcomes are virtually comparable.

Precision

The percentage of optimistic predictions that are accurate.

Precision = TP / (TP + FP)

Precision is defined as the ratio of precisely predicted positive observations to all expected positive observations. This metric answers the query of what percentage of all commuter who were classed as surviving actually did. Precision is

associated with a low percentage of false positives. Our accuracy is 0.788, which is excellent.

Recall

The percentage of correctly projected positive adhered values is known as the proportion of correctly anticipated positive observed values. (The percentage of real defaulters predicted properly by the model)
Recall = TP / (TP + FN)

Yes, recall is defined as the proportion of accurately predicted positive observations to the total number of observations in the class.
F1 Score: Precision and Recall's weighted mean is this. As an outcome, both false negatives and false positives are considered. Although it is less straightforward than calculating accuracy, F1 is often more valuable than accuracy, particularly when the class distribution is unequal. When the cost of false positives and false negatives is equal, accuracy improves. If the value of false negatives and false positives is significantly varied, it is preferable to focus at both Recall and Precision.

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

General Formula

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

A. Algorithms and techniques

Computers use what they've learned from previous data to categorise new data. This is called "supervised learning". Speech recognition, handwriting recognition, biometric identity, document classification, and other classification problems are only a few examples. Algorithms that use supervised learning learn from labelled data. When the algorithm analyses the data, it looks for patterns and links them to new data that hasn't been labelled yet. Then, the algorithm decides which label should be given to the new data based on the patterns.
Python packages used

sKlearn

- sklearn is a Python machine learning library that incorporates a number of different machine learning techniques.
- Some of its modules are employed here, including train test split, Decision Tree Classifier or Logistic Regression and accuracy score.

NumPy

- You can use it to make quick math calculations in Python.
- When you use it, you can read data from numpy arrays and do things with it.

Pandas

- Reads and writes a variety of files.
- Data frames make it simple to manipulate data.

Matplotlib

- Data visualisation is an effective tool for identifying trends in a dataset.
- Data frames make it simple to manipulate data.

Decision tree classifier

It's a well-known and powerful algorithm. The decision-tree algorithm is classified as a supervised learning algorithm. It works with both categorical and continuous output variables. What we think about when we build a decision tree. At the start, we think of the whole training set as the root. Although attributes are intended to be categorical for data collection, they are recognised as continuous. Records are disseminated recursively based on attribute values. As a root or internal node, we rate qualities using statistical approaches. A decision tree creates classification or regression models in the form of a tree structure. It breaks down a large data set into smaller and smaller parts while simultaneously building a decision tree to go along with it. A leaf node represents a classification or decision, while a decision node contains two or more branches. The best predictor is represented by the root node, which is the highest decision node in a tree. A leaf node represents a classification or decision, while a decision node contains two or more branches. The best predictor is represented by the root node, which is the highest decision node in a tree. Decision trees can handle both categorical and numerical data. In the shape of a tree structure, a decision tree generates classification or regression models. For classification, it employs a set of mutually exclusive and exhaustive if-then rules. Using the training data, the rules are learned one by one, one by one.

The tuples covered by a rule are eliminated each time it is learned. On the training set, this approach is repeated until a termination condition is met. It's developed using a recursive divide-and-conquer strategy from the top down. All of the characteristics should be categorical in nature. If not, they should be separated ahead of time. The information gain concept is used to find attributes at the top of the tree that have a stronger impact on classification. An excessive number of branches might emerge from overfitting a decision tree, indicating anomalies due to noise or outliers.

```
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.]
```

Random forest classifier

Random forests, alternatively called random decision forests, are a type of ensemble learning technique for classification, regression, and other tasks that involves training a large number of decision trees and then producing the class that represents the mean prediction (regression) of the individual trees. Decision trees have a propensity to overfit their training set, which is corrected by random choice forests. Random forest is a supervised machine learning technique based on ensemble learning. Ensemble learning is a kind of machine learning in which multiple versions of the same algorithm are merged to create a more accurate prediction model. The random forest algorithm combines various similar algorithms, such as multiple decision trees, to form a forest of trees, hence the name "Random Forest." Both regression and classification problems can be solved using the random forest method. The random forest algorithm is performed in the following steps:

- Choose N records at random from the collection and use them to build a decision tree.
- Rep steps 1 and 2 as needed for the number of trees in your method.

Each tree in the forest predicts a value for Y for a new record in the event of a regression issue (output). Averaging all of the predicted values from all of the trees in the forest yields the end result. In the event of a classification challenge, each tree in the forest anticipates the category to which the new record belongs. Finally, the category with the most votes receives the new record.

```
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.]
```

Naive bayes algorithm

The Naive Bayes algorithm is a simple methodology for making predictions based on the probabilities of each attribute belonging to each class. If you wanted to model a predictive modelling problem probabilistically, this is the supervised learning strategy you'd use. Naive Bayes simplifies the calculation of probabilities by assuming that the probability of any attribute belonging to a specific class value is independent of all other features. This is a bold assumption, yet it yields a quick and effective solution. The conditional probability is the probability of a class value given an attribute value. We can calculate the likelihood of a data instance belonging to a class by multiplying the conditional probabilities for each attribute for a particular class value. To make a forecast, we can calculate the probabilities of each class's instance and choose the class value with the highest likelihood. The Bayes Theorem is used to build the Naive Bayes statistical categorization approach. On the market, it's one of the most simple supervised learning algorithms that you can use. The Naive Bayes classifier is a reliable, fast, and accurate algorithm. On large datasets, Naive Bayes classifiers have great accuracy and speed.

The Naive Bayes classifier posits that the influence of one characteristic in a class is not dependent on the effect of other features. For example, a loan applicant's value is determined by his or her income, prior loan and transaction history, age, and geographic area. Even though these traits are interrelated, they are still evaluated separately. As this assumption makes calculation easier, it is regarded as naive. Class conditional independence is also the term for this assumption.

```
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
```

Conclusion

The analytical approach included data cleaning and processing, missing value analysis, exploratory analysis, and model construction and evaluation. Finally, we anticipate the crop using a machine learning algorithm, with different results. As a result, the following crop forecast insights emerge. Farmers will be able to learn about crops that have never been cultivated before and see a list of all available crops, which will benefit them in picking which crop to produce, because this system will cover the most types of crops. Furthermore, this strategy incorporates previously collected data, allowing the farmer to obtain insight into market demand and costs for specific crops.

References

1. A. Mark Hall, “Feature selection for discrete and numeric class machine learning,” Comput. Sci., Univ. Waikato, pp. 359–366, Dec. 1999.
2. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” J. Mach. Learn. Res., vol. 3, pp. 1157–1182, Jan. 2003.
3. R. Gilad-Bachrach, A. Navot, and N. Tishby, “Margin based feature selection–theory and algorithms,” in Proc. 21st Int. Conf. Mach. Learn. (ICML), 2004, p. 43.
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. Ranjini, A. Suruliandi, and S. P. Raja, “An ensemble of heterogeneous incremental classifiers for assisted reproductive technology outcome prediction,” IEEE Trans. Comput. Social Syst.early access, Nov. 3, 2020, doi: 10.1109/TCSS.2020.3032640. [7] H. Liu and R. Setiono, “A probabilistic approach to feature selection-a filter solution,” in Proc. 13th Int. Conf. Int. Conf. Mach. Learn., vol. 96, 1996, pp. 319–327.
7. R. Kohavi and G. H. John, “Wrappers for feature subset selection,” Artif. Intell., vol. 97, nos. 1–2, pp. 273–324, Dec. 1997.
8. 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.
9. 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.
10. M. A. Hall and G. Holmes, “Benchmarking attribute selection techniques for discrete class data mining,” IEEE Trans. Knowl. Data Eng., vol. 15, no. 6, pp. 1437–1447, Nov. 2003.
11. H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502, Apr. 2005. [13] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” Chemometric Intell. Lab. Syst., vol. 83, no. 2, pp. 83–90, Sep. 2006.
12. Araåzo-Azofra and J. M. Benítez, “Empirical study of feature selection methods in classification,” in Proc. 8th Int. Conf. Hybrid Intell. Syst., Barcelona, Spain, Sep. 2008, pp. 584–589.
13. Altmann, L. Toloåi, O. Sander, and T. Lengauer, “Permutation importance: A corrected feature importance measure,” Bioinformatics, vol. 26, no. 10, pp. 1340–1347, May 2010.
14. M. B. Kursa and W. R. Rudnicki, “Feature selection with the Boruta package,” J. Stat. Softw., vol. 36, no. 11, pp. 1–13, 2010.
15. G. Ruß and R. Kruse, “Feature selection for wheat yield prediction,” in Research and Development in Intelligent Systems. London, U.K.: Springer, 2010, pp. 465–478.
16. J. Camargo and A. Young, “Feature selection and non-linear classifiers: Effects on simultaneous motion recognition in upper limb,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 27, no. 4, pp. 743–750, Apr. 2019.
17. M. B. Kursa, A. Jankowski, and W. R. Rudnicki, “Boruta—A system for feature selection,” Fundam. Inf., vol. 101, no. 4, pp. 271–285, 2010.
18. R. Rajasheker Pullanagari, G. Kereszturi, and I. Yule, “Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression,” Remote Sens., vol. 10, no. 7, pp. 1117–1130, 2018.
19. Choudhary, S. Kolhe, and H. Rajkamal, “Performance Evaluation of feature selection methods for Mobile devices,” Int. J. Eng. Res. Appl., vol. 3, no. 6, pp. 587–594, 2013.

20. F. Balducci, D. Impedovo, and G. Pirlo, “Machine learning applications on agricultural datasets for smart farm enhancement,” Machine, vol. 6, no. 3, pp. 38–59, 2018.
21. J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, “Feature selection for SVMs,” in Proc. Adv. Neural Inf. Process. Syst., vol. 13, 2001, pp. 668–674.
22. Bahl et al., “Recursive feature elimination in random forest classification supports nanomaterial grouping,” NanoImpact, vol. 15, Mar. 2019, Art. no. 100179.