Capstone Project Proposal Document

Project Guide : Prof. Raghu B A, Associate Professor, PES University
Project Team : PES1201800051 Srish Srinivasan
PES1201800089 Akash Kumar Rao
PES1201800102 Vishruth P Reddy
PES1201800291 Ishan Agarwal

In collaboration with RRSC South, ISRO Bangalore under the guidance of Dr. Ramasubramoniam, Soil and Agri Scientist.

Crop Prediction

1. Introduction

Our project comes under the domain of Precision Agriculture. Therefore, it is very important to understand what Precision Agriculture means at its core. Precision Agriculture is a system to manage farms that is based on the use of advanced technologies at every step of the agriculture process. It is based on observation, measurement and response to the variability in crops. The aim of precision farming is to develop a decision support system for the management of a farm with the goal being maximization of returns with the efficient and judicious use of resources. In simpler terms, the goal is to increase the harvest through the efficient usage of seeds, fertilizers and pesticides. With the involvement of advanced technology, decisions are made purely based on data thereby reducing the risk of failure to a large extent as decisions are no longer made based on intuition and luck. And by making use of resources in an efficient way, we are saving the environment from the rapid depletion of its natural resources.

Precision agriculture originated in the 1980s in the United States of America. Precision agriculture was the most important development of the third wave of modern agricultural revolutions. In the first agricultural revolution that took place between the 1900 to 1930, people saw a large increase in mechanized agriculture. This resulted in each farmer producing food that was more than sufficient to satisfy the hunger of 26 people. Later in the 1960s, people witnessed the second agricultural revolution in the form of the Green Revolution which gave birth to genetic modification. At this stage, farmers were able to produce food that was more than sufficient to satisfy the hunger of 156 people. And finally, with rapid technological advancements and its integration in the field of agriculture gave birth to the third agricultural revolution which was widely known as Precision Agriculture. The other synonyms for Precision Agriculture include “Precision Farming”, “Satellite Farming” and “Site Specific Crop Management”. Input recommendation map for fertilizers was the first outcome in this domain and this was based on the grid soil sampling that was done. However, it was in its very early stages and was not practiced much. But with the advent of smartphones, high speed networks and enormous amounts of satellite data, precision agriculture has become very popular and has seen a steep growth in the last 5 years.

As we all know, a field has heterogeneous zones and with the aid of technology, we can identify these zones and manage their variability. Therefore, it is important to have some knowledge about the technologies being used. The GPS has been one of the most important enabling factors for the practice of precision farming. The farmer's ability to precisely locate his/her position in the field led to the development of spatial variability maps for variables such as crop yield, nutrient levels, humus content and soil moisture content. Data similar to the ones mentioned in the previous sentence are extracted using sensor arrays mounted on GPS-equipped combine harvesters.

These sensors work in real time thereby collecting a wide spectrum of information ranging from pigment levels to water status, along with multispectral images. This data is used in combination with satellite images through variable rate technology which include seeders, sprayers, etc. in order to achieve optimal distribution of resources. In order to avoid the possibility of an overlap or an underlap during the process of sowing seeds or the application of fertilizers and pesticides, onboard computers and GPS-navigators are used in vehicles. Digital maps and variable rate applications are being used for fields based on variable characteristics and the calculation of fertilizer dosage for every individual zone respectively. With the help of recent technological advancements, real time sensors placed in soil can wirelessly transmit data without the need of human intervention post sensor installment in the soil. For the remote monitoring of fields, drones and satellites have been put to use. Unmanned aerial vehicles are not very expensive and can be controlled by pilots without the need of any prior experience. These drones are equipped with multispectral cameras that can capture multiple images of the field.

A lot of sensors have been predominantly wireless ones that have been used in order to numerically capture the influence of field indicators such as moisture, pressure, rainfall, temperature, etc. And finally, with the help of applications that are hosted either as a mobile application or a web application, all the data that has been captured will be analyzed in order to obtain some meaningful insights that can be used to manage farms in an efficient manner.

Coming to the complexity involved in practicing precision farming, it is a little complex because most of the technologies that are being used are new and therefore requires a skilled workforce in order to make good use of this technology. For instance, a person who lacks the required skills will find it difficult to analyze a satellite image or repair an onboard computer. But at the same time, technological solutions that are simple do exist and can be accessed by every farmer. Some of these simple solutions include weather sensors, wireless modems, etc.

The next very important thing to look into is the cost involved in the incorporation of precision farming in your farms or agricultural fields. At present, some of the sophisticated equipment and software are quite expensive and therefore the precision farming technologies are mostly seen only in large farms owned by affluent farmers.

But it is a well-known fact that with increasing developments in technology, it only becomes more affordable and easier to use. With this natural dynamic, the focus is on enabling easy access to these technologies with very little cost involved.

Stepping into the future, precision agriculture is something that will be unavoidable and it makes absolutely no sense in not making use of it because it is extremely profitable. In one of

the blogs by the name onesoil.ai, we learnt that the American farmers are able to save on a large amount of money that they spend on agriculture through the incorporation of precision farming in their fields.

With precision farming up and running, farmers will be able to:

1. Make improved decisions
2. Improve the inherent quality of the farm products
3. Enhance marketing of farm products
4. Improve relationships with landlords and local money lenders

So, our focus in this project is to improve the decision making involved in the process of crop selection with the help of machine learning algorithms by taking into account the soil properties and the surrounding atmospheric conditions.

2. Project Scope

The aim of this project is to build a predictive model to recommend the most suitable crop to grow based on the various parameters that influence the fertility of the soil.

This project enables the farmers to grow the most suitable crop by factoring in various soil characteristics like N, P, K contents and pH and atmospheric conditions like temperature, humidity and rainfall. This results in greater yield of crop and therefore, stabilizing their financial status.

In this project, the focus is on analysing the existing data and employing suitable models in order to give the best recommendations possible to the farmers. On the other hand, we will not be diving too deep into the implementation of how the data will be extracted but we will be researching about the methods used to collect the same. One of our data sources is only limited to 22 crops but we will make an effort to find more data in order to make this product more robust with regard to its recommendation power.

3. Literature Survey

a. Paper 1

The authors of [1] have proposed a machine learning based solution for the analysis of imperative soil parameters and their influence on the kind of crops that could be suitably grown in a given soil. The various soil nutrients are treated as the independent variables and the grade of the soil is the target variable. The regression algorithm along with RMSE values were employed to predict the rank of a soil and on applying a few classification algorithms for the purpose of crop recommendation, they found that Random Forest was the most accurate model.

In order to have a good yield, it is important that the soil is rich in the required nutrients. So, the main goal in this project was to rank a soil sample by examining its nutrient contents (Macronutrients and Micronutrients) and then recommend the most suitable crop that could be grown in this soil.

In the first part of the project, the contents of various soil nutrients such as EC, K, pH, Mn, Zn, S, P, and B are considered as the independent variables and the grade of the soil is considered as the dependent variable. So, a Multivariate Linear Regression model was built to predict the fertility of soil on a scale of 1-5.

A linear combination of the independent variables was chosen as the hypothesis function. The cost function chosen was:

1. Xi = vector of independent variables
2. h_θ = hypothesis function
3. Yi = True value of the response variables
4. m = normalization parameter

The Gradient Descent Algorithm was adopted to minimize the cost function. Then hypothesis testing was carried out on the test dataset in order to check for the model's correctness and efficiency and the RMSE value was used to determine accuracy of the model.

In the second component of the project, the authors attempted to recommend crops using machine learning algorithms such as Support Vector Machines, Random Forest Classification and Decision Tree and based on the RMSE value the best model was chosen. The true accuracy of the model will be obtained when real-time data would be passed to this model.

The most important feature is used to split a node and then recursively the next most important feature is looked for from the subset of the remaining features thereby generating a highly accurate classifier with wide diversity.

In order to split a node, only a selective group of features are selected among all the features. An element of randomness is introduced through the use of random thresholds for the feature set. A Random Forest Classifier applies a technique known as bootstrap aggregation or bagging to the tree learners. From the training set, random sampling with replacement was performed and to each of these samples, trees were fit.

Then a voting is performed among all the predictions output by all the trees in order to arrive at the final result. To ensure that the variance is low and at the same time the bias is also kept low, the bootstrapping procedure was applied. If the trees are not related to each other, then the average of the outputs produced by these trees are more robust to noise but in the case of a single tree, the prediction made can be very easily influenced by the noise.

Therefore, the idea behind using different samples from the training sets was to develop trees that are highly uncorrelated. The number of trees used in a random forest classifier is usually in the range of a few hundreds to several thousands and this number is heavily dependent on characteristics of the training data set.

Learnings from [1] are

1. The Random Forest Algorithm is based on ensemble learning and proved to be a very effective algorithm for classification.
2. The basic idea is to build multiple decision trees from randomly selected subsets of the data. And then when a new data instance comes in, it is put through all these decision trees and a majority vote is taken in order to give the instance its final classification.
3. Each tree as individual entities might not be ideal, but as a group they can perform really well.
4. Since there are numerous trees, the existence of any errors or uncertainties associated with any of the trees are taken care off by this algorithm.

b. Paper 2

Yield rate of crops is dependent on two broad factors, the first being genetic development of seeds which lies more on the fields of Bio-Technology and the second is crop selection management. The latter factor is more algorithmic and thus an algorithm can be developed for this job specifically. This concept drives this paper towards building an algorithm for just this purpose. It is provided with the necessary inputs and information, and a selection pattern is expected in return. Some factors that go as input to this algorithm include the predicted yield of the respective crop. This prediction is done by various different machine learning models over various different factors like soil properties such as Nitrogen, Phosphate and Potassium contents, pH properties of the soil, weather conditions, rainfall predictions or forecasts. The machine learning models used to perform this task are the most prominent ones that have proved their value in various other fields of research. The newer models, some that use Boosting techniques had not been tried into this field of research until this paper was published, and the paper aimed at trying them out as well and place a detailed comparative analysis for the readers. These techniques included GDBT (Gradient Boosted Decision Tree) and RGF (Regularized Greedy Forest).

The method is composed of 2 significantly differentiable parts that work together. The first part uses machine learning models like Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree Learning, Random Forest, Gradient Boosted Decision Tree (GBDT), Regularized Greedy Forest (RGF) to predict yield rate of crop before the season arrives. This is possible due to the fact that the season of the entire year in India depends on summer rainfall data of said year. Therefore, it is possible to predict a season in advance, provided that the rainfall data is available.

Crops can be classified as:

- Seasonal crops
- Whole year crops
- Short time plantation crops
- Long time plantation crops

Second major part of the research work includes appropriate crop selection. Once the yield of all the types of crops is calculated, it is to be followed by selection among these crops that maximize the yield and also keep seasonal rotations in consideration along

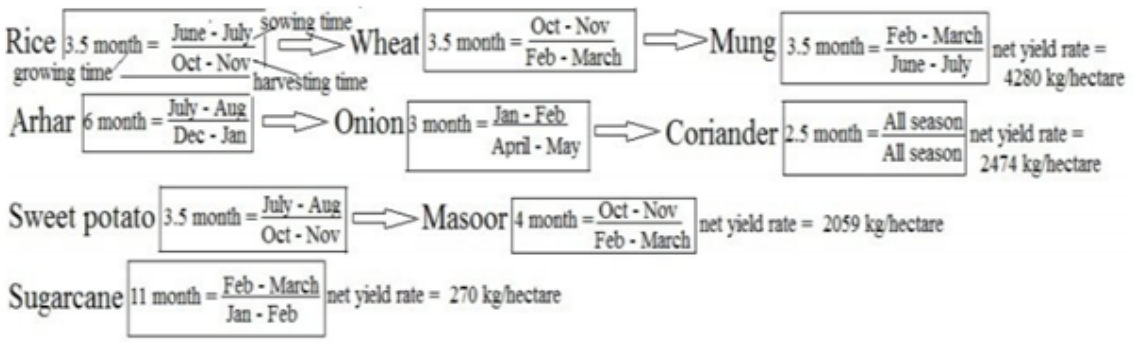
with minimizing crop-less days that are waste of farmers' resources. Below is the algorithm used in this part of the research work.

Algorithm: Crop Selection Method(CSM)

```
cropSelector(prese
ntTime)
   if presentTime ≥ End of
      Season then return 0
   end if else
      if presentTime= sowingTime then
         return cropSelector(presentTime + 1)
      end if else
         cropSowingTable←cropInputTable(presentT
         ime) L: crop ← max{crop →prRate /
                        crop → plantationDay}
                  crop ∈ cropSowingTable

        if (presentTime + crop → platationDay) ≥
              End of Season then
           //remove crop from cropSowingTable
           cropSowingTable ← cropSowingTable -
           crop if cropSowingTable is NULL then
              return cropSowingTable(presentTime +
              1)
           end if else
              go to L
           end else end if
        else
           update(OutputcropTable, crop)
           npr ← (crop → prRate +
           cropSelector(presentTime + crop
           →plantationDay))
           return npr
        end else end else
  end else
end cropSelector
```

Taking an example to further explain the work done by the authors makes things easier to understand. Consider the below image where four different crop rotation options are available for the farmer to choose. The algorithm is provided with all of this data, that is, each crop's seasonal information, as well as yield rate for the year provided by the machine learning model from the first part of the work.



The algorithm chooses the first option, that is, Rice followed by Wheat and Mung in respective order. The net yield rate for this option is the highest among all the options and can be grown in the said order without any seasonal conflicts.

In Conclusion, the final sequence of crops is the end result of the Crop Selection Method (CSM) with inputs such as rainfall of the year and all the crops' seasonal information. The work done is remarkable and has been cited multiple times for further research work and enhancements. One of many important points made in the paper is the yield in the upcoming season can be predicted using historical rainfall data. The machine learning models used were never tried before for this particular application and the results turned out to be very satisfactory.

There are some cons to the work as well which should be mentioned with details. The first one being, the limitation of the geographical diversity of the data collected. The data is collected from a single farmer residing in Patna district of Bihar, India. This data and results might be appropriate for the said district, but cannot be extrapolated to entire country's Agriculture Sector. The yield of the crops uses rainfall data alone, whereas in reality it depends on various other factors like soil parameters, climatic and weather conditions. Therefore, including these factors as well into a future study holds a lot of potential.

c. Paper 3

In our country, due to the lack of accessibility to technology, farmers grow crops purely based on the history of crops grown in that region and based on intuition. It may work out sometimes but they go under heavy losses mostly. Putting in months of effort to see their crop not prosper is very sad and is a huge loss to the farmers. Due to lack of awareness and knowledge, they might perform certain practices extensively or may not be doing enough which results in poor yield. Overuse of fertilizers, insecticides and pesticides can also lead to poor production. Even if the farmer doesn't factor into the atmospheric conditions, he will go under loss due to poor planning. Educating the farmer may be helpful but it is a tedious job. To make their work simpler and more profitable, we can use machine learning techniques to predict the best crop a farmer or horticulturist must grow by factoring in all the parameters to increase his profits.

This early prediction can help farmers plan for either annual or seasonal crops. Use of precision farming can give more accurate results due to the high inspection and efforts on a small area. Since we can get almost accurate values of soil parameters, if weather conditions are known more accurately before-hand, farmers can adapt accordingly and try growing a suitable crop.

Employing machine learning can give us insights about the soil fertility, the elements in soil and atmospheric conditions which can be used to precisely predict what crop can be grown in that particular field. In this paper, the authors have employed a variety of machine learning methods such as supervised, reinforcement and unsupervised learning. Techniques such as regression, clustering, classification etc are used to predict the perfect crop for the provided conditions.

Linear regression is a technique used when we have 2 parameters of interest. When one parameter is dependent on the other, linear regression comes into play. The only drawback is that it works only for linear data and not for non-linear or complex data.

Artificial Neural Network and especially Back Propagation Neural Network is used when we have multiple parameters of interest which decide the crop yield. Here, the 3 layers namely, the input, the hidden and the output layer decide the crop yield values. Weights can be adjusted to get a better recommendation. ANN not only works for linear data but complex data as well.

Support Vector Machines is another technique that gives very accurate results. The problem of overfitting doesn't affect SVM. SVM is used when we have lots of parameters to consider and especially atmospheric conditions. This removes the issues of changing the weights to obtain a desirable value as that was the case in ANN.

The authors have used several metrics to validate the output of the predictions. These metrics give us the accuracy of the predictions. Mean Squared Error, Mean Absolute Error and Root Mean Squared Error are the metrics employed by the authors to validate the outcomes.

Metrics	**Formula**
Root Mean Squared Error	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y_i})^2}$
Mean Squared Error	$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y_i})^2$
Mean Absolute Error	$MAE = \frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \overline{y_i}\rvert$

The paper also throws light on various techniques used for predicting various crops. Multiple Linear Regression gives the best results for tea crop yield as it is only dependent on the soil conditions such as being acidic, well drained and light soil. Pepper, potato and tomato are predicted using MLR as well.

ANN is used for predicting Maize and wheat crops due to non-availability of data and presence of non-linear data.

SVM is also used to predict the yield of Maize crops as they are unaffected by overfitting. It also distinguished between a crop and weed growing in the surrounding areas.

The inputs to the models will be broadly categorized into weather and non-weather inputs.

Weather Parameters	**Non-weather Parameters**

Temperature	Soil Moisture
Rainfall	pH
Humidity	Crop Type
	Seed Variety
	Salts such as N, P, K, C, Ca, Mg, Mn, S etc.

The pros of this paper is that they have incorporated different machine learning techniques for prediction and validated them using different performance metrics. The cons of the paper are that they didn't work on any big data models for prediction but mentioned that further work can be done along big data lines. So, they should have used an appropriate title for their research work as it was a little misleading.

d. Paper 4

In this paper, the authors have taken into account the different ML algorithms which are used in crop prediction over various other studies and have tried to add more attributes to the system in order to improve the results. They have compared the prediction of the ideal crop from using different models to get a better understanding of how to use ML techniques in the future.

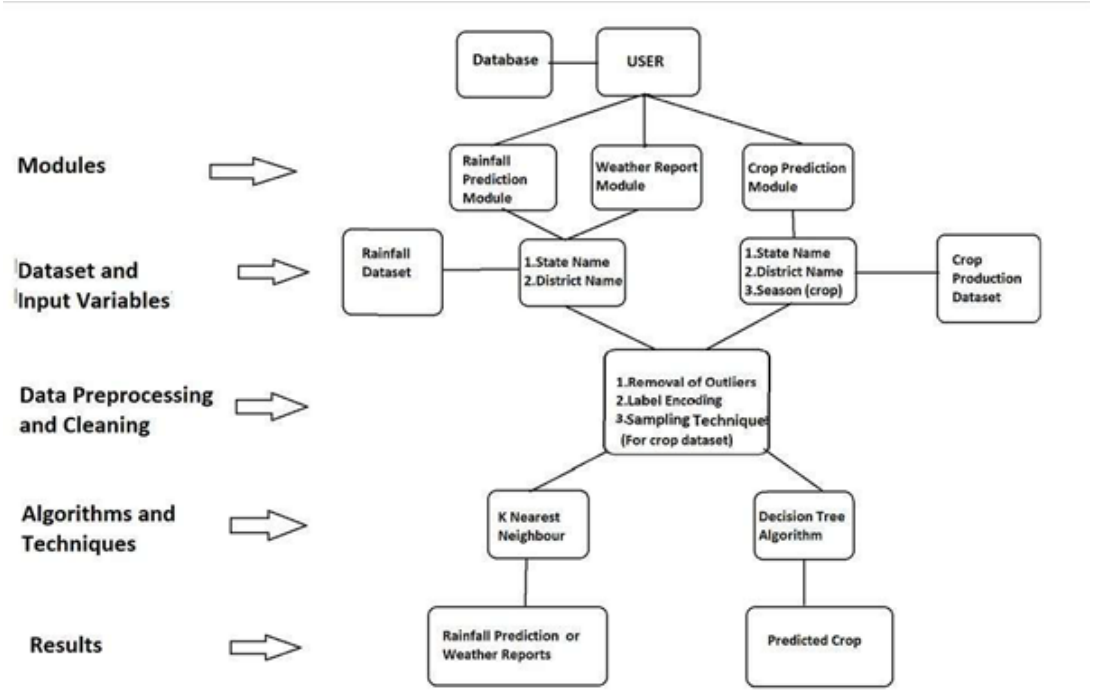
In order to increase yield rate of the crops, several biological and chemical approaches have been implemented over the years like better quality seeds, proper use of insecticides and pesticides, use of fertilisers, etc. The method of crop prediction identified by the authors based on previous work done i.e. the crop selection method (CSM) distributes crops into:

1. Seasonal
2. Whole Year

3. Short plantation period
4. Long-time plantation

The data of these were then taken for a particular selected region (as agriculture depends on the type of place and various factors like climate, soil, etc.) and then the farmers could be given a list of crops they would choose from along with the desired sequence in which the crops could be planted so as to increase the total yield throughout the season. This may also improve land reusability and hence the resources available thus further improving the farmers' profit. Thus, the already existing systems can give the suitable crop keeping in mind the yield over a particular selected region.

The previous work the authors surveyed made use of ML algorithms with one attribute and thus they made a system to add more attributes to it so that along with crop, the time of the year and the weather prediction is also taken into account. This is shown in their work flowchart below.



The data must include:

- Soil Parameters
- Soil Type
- Soil pH
- Humidity
- Temperature
- Wind
- Rainfall
- Production
- Cost of cultivation
- Previous year yield results

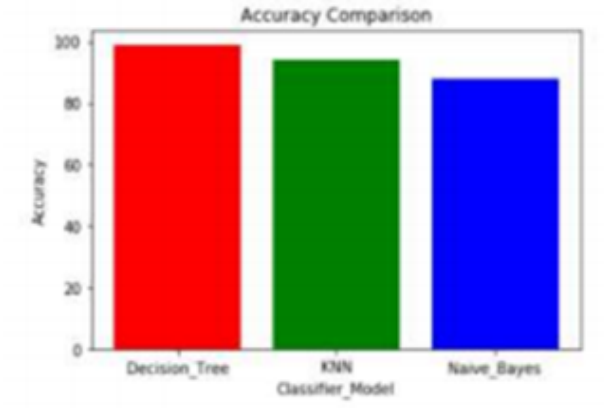
The data is pre-processed and fed into KNN, Decision tree and Naive Bayes classifier and

the results from all of these are compared.

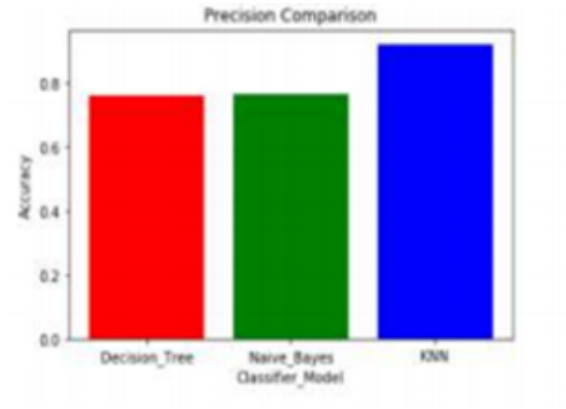
(The selection attributes of the Decision tree being Gini index, entropy and information gain.)

The results were as follows:

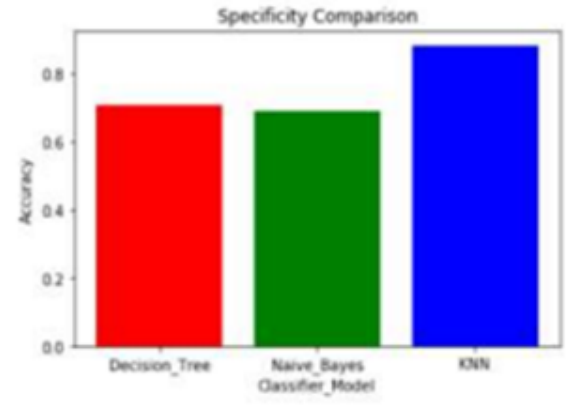
- Accuracy:



- Precision:



- Specificity



After studying the results from the three models and their comparisons, the conclusion obtained is that Decision tree shows poor performance when the

dataset is having more variations but naïve bayes provide better results than decision trees for such datasets. The combination classification algorithms like naïve bayes and decision tree classifiers are better performing than use of a single classifier model.

Thus, we can make use of this study and also cross check the findings when using different models.

Literature Survey Comparison

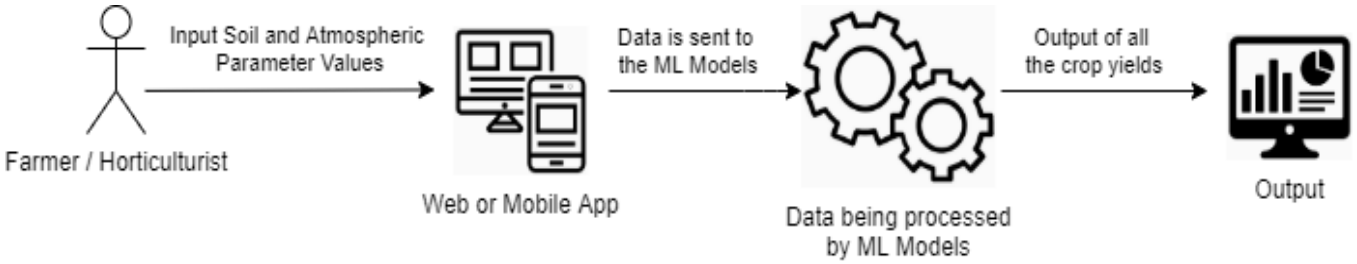
Paper1: Random Forest Algorithm for Soil Fertility Prediction and Grading Using Machine Learning by Keerthan Kumar, T.G., Shubha, C. and Sushma, S.A, 2019. **Algorithm used:** Multivariate Linear Regression, Random Forest Classifier	**Paper 2: Crop Selection Method to maximize crop yield rate using machine learning technique,** Kumar, R., Singh, M.P., Kumar, P. and Singh, J.P., 2015 **Algorithms used:** ANN, SVM, KNN, Decision Trees, Random Forest, Gradient Boosted Decision Tree (GBDT), Regularized Greedy Forest (RGF).
Paper 3: AN APPROACH FOR PREDICTION OF CROP YIELD USING MACHINE LEARNING AND BIG DATA TECHNIQUES, Palanivel, K. and Surianarayanan, C., 2019. **Algorithm used:** Multiple Linear Regression, ANN, SVM.	**Paper 4: Crop Prediction using Machine Learning Algorithms,** Patil, A., Kokate, S., Patil, P., Panpatil, V. and Sapkal, R., 2020. **Algorithms used:** KNN , Decision tree and Naive Bayes classifier.

4. Methodology

The design approach we have planned on using is as follows:

1. Choose a specific geographic location like a district or state, for eg: Karnataka.
2. Gather past data for the chosen area.
3. Allow the users to enter real-time data into the application.
4. The application will predict and display a list of crops along with their yield percentage which will be ranked in hierarchy.

In the backend, the machine learning models such as SVM, Decision Trees, ANN etc will process the data provided and will predict the range of crops.



- Currently, we have implemented 4 machine learning algorithms namely **Decision Trees, K-Nearest Neighbors, Naïve Bayes Classifier and Random Forest Classifier**. Among these models, the **Naïve Bayes Classifier** gave the maximum accuracy of **99.47%**.
- We will make an attempt to implement some of the **ensemble machine learning algorithms such as AdaBoost and XGBoost** and also an **Artificial Neural Networks** model in order to achieve the recommendation of crops with a good accuracy score.

5. Implementation

Decision Tree Algorithm Classifier

Decision Tree

Importing the libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Importing the dataset

In [2]:
```
dataset = pd.read_csv('Crop_recommendation_dataset.csv')
```

In [3]:
```
dataset.head()
```

Out[3]:

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

Extracting Target Variable and independent variables

In [5]:
```
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
```

Splitting the dataset into training set and test set

In [6]:
```
crop_labels = list(dataset.label)
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.4, random_state = 0, stratify = crop_labels)
```

Feature Scaling

In [7]:
```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Training the Decision Tree model on the Training set

In [8]:
```
from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
classifier.fit(X_train, y_train)
```

Out[8]:
```
DecisionTreeClassifier(criterion='entropy', random_state=0)
```

Predicting the Test set results

In [9]:
```
y_pred = classifier.predict(X_test)
```

Accuracy Score

In [10]:
```
from sklearn.metrics import accuracy_score, confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print(cm, '\n')
ac = accuracy_score(y_test, y_pred)
print("Accuracy Score: ", ac)
```

```
[[40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0 36  0  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0]
 [ 0  0  0 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 40  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  1  0  0 37  0  0  0  0  0  0  0  0  0  0  2  0]
 [ 0  0  0  0  0  0  0  0  0 40  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0 40  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0 40  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0 40  0  0  0  0  0  0  0  0]
 [ 0  0  2  0  0  0  0  0  0  0  1  3  0 34  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0 40  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 40  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 40  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 40  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 40  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 40  0]
 [ 0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0 38]]

Accuracy Score:  0.9806818181818182
```

Applying k-Fold Cross Validation

In [11]:
```
from sklearn.model_selection import cross_val_score
accuracies = cross_val_score(estimator = classifier, X = X_train, y = y_train, cv = 10)
print("Accuracy: {:.2f} %".format(accuracies.mean()*100))
print("Standard Deviation: {:.2f} %".format(accuracies.std()*100))
```

```
Accuracy: 98.03 %
Standard Deviation: 1.28 %
```

Satellite Crop Prediction

1. Introduction

Our project comes under the domain of Precision Agriculture. Our aim in this project is to build a recommendation system that recommends the most suitable crop to be grown given the properties of the soil, the surrounding atmospheric conditions of a specific location and Satellite image data (LANDSAT) of the location. We will be narrowing down our focus onto a specific part of India to obtain greater accuracy.

2. Project Scope

The aim of this project is to build a predictive model to recommend the most suitable crop. Upon studying through previously done work it was found that the most common form of crop prediction was CSM (Crop Selection Method). CSM algorithm works on prediction of crop yield rate based on favorable condition in advance and gives a sequence of crops with highest net yield rate. Although this technique proves to be effective it requires a lot of data including soil profile , rainfall, wind, previous year datasets .On top of this the data training is complex and time consuming and in developing countries or places not used in agricultural activities before even the data collection appears to be costly and time consuming .This is where remote sensing(RS) has an edge as all it needs is the data from remote sensing satellites which can be further analyzed to obtain the different spectral indexes [eg. NDVI,EVI,NDWI]. Then simple regression models can be used to get the desired yield health results.

3. Literature Survey

1. **Yield prediction with machine learning algorithms and satellite images [2020]**

   The author has used given satellite images over Iran in this paper and used it along with some simple regression models to obtain the yield prediction over a particular region of Iran along with the optimal time. The author has chosen a specific crop i.e barley and has predicted the yield for the same.

   The author has used the remote sensing and climate data from the Google Earth Engine (GEE) platform these were integrated with four machine learning algorithms i.e backpropagation neural network (BPNN), decision tree (DT), gaussian process regression (GPR) and K-nearest neighbor regression (KNN) algorithms.

   It is further seen that the two indexes used by the author are NDVI and EVI which are calculated as follows :-

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

$$EVI = 2.5 * \frac{(NIR - RED)}{(NIR + C_1 * RED - C_2 * BLUE + L)}$$

Compared to NDVI, EVI does not saturate at high canopy densities, can reduce canopy background signal and atmospheric influence, and improve vegetation dynamics in high biomass areas.A combination of NDVI and EVI can provide more information about the crop yield, which helps yield prediction.

Considering three evaluation indicators (r2,RMSE and MAE), GPR, DT, BPNN and KNN models showed higher accuracy, with r2 (0.84–0.69) and RMSE (< 737 kg ha −1) and MAE (< 650 kg ha −1), respectively . Therefore, these models are very suitable for predicting atmospheric performance in southern Iran.

Results showed that the GPR method had the highest prediction accuracy and showed the best generalization ability compared to the others. The GPR model was able to accurately estimate barley yield approximately 1 month before harvest. The results also showed that the accuracy of the prediction depends on location and the time interval. Also, among the variables used in the GEE, three variables, including EVI, minimum temperature and precipitation, had the most significant effect on the modeling process.

2. **Crop yield estimation using satellite images: comparison of linear and non-linear models [2020]**

The goal of this research was to develop and evaluate linear and non-linear models to estimate crop yield from satellite data.The authors used images from the Landsat and SPOT satellites to obtain soybean and corn yield in the central region of Córdoba (Argentina) .This was also compared with field data collected over the places to check the validity of the model.

Since in the cultivated regions where the chosen lands are small the data from Landsat and SPOT satellites were sufficient due to their spatial resolution. Four images , two from each, were examined which were taken on days with clear weather and after the required image processing ,the subset of all the said were taken in order to cover the entire area. This was further used for training and testing the data on various ML models.

MLR models were used where the yield was calculated as:

$$\text{Yield} = \sum_{i=1}^{k} a_i v_i + b$$

where the regression variables are vi = surface reflectance of i band of SPOT/Landsat satellite and model constants are ai and b.

Neural networks were also used with an input layer whose number of neurons was equal to the number of bands considered in each model; a hidden layer was designed with the same number of neurons than the respective input

layer. The output layer was built with only one neuron that indicates the calculated crop yield.

It was found on comparing all of these that all regression and neural network models developed to estimate yield provided a good fit with measured yield.

This further proves that satellite images can be used to accurately determine yield of a crop and also neural networks provide better accuracy than machine learning models in most cases . Hence, The possibility of combining satellite images with climatologic or soil data to improve the performance of yield estimation, is the next step to be explored.

4. Methodology

Since previous works have used these technologies/methods to predict the crops individually, there is a potential of achieving better results when these methods are combined. Therefore, we provide our models with all the 3 types of data, that is, soil properties such as N-P-K, atmosphere properties such as moisture and vegetation indices obtained from satellite image data.

A flow chart of the methodology is shown in Fig.1

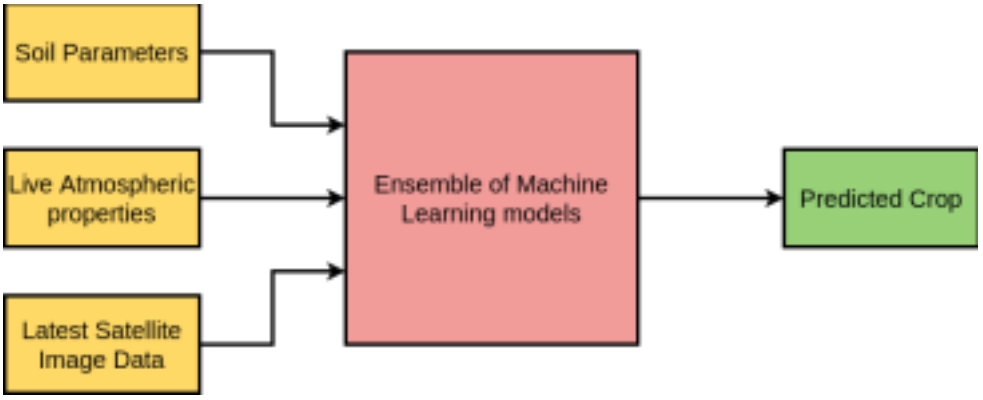


Fig.1

A prototype code for the generation of NDVI using LANDSAT data from USGS (United States Geological Survey) was run successfully. The index value obtained now has to be given as an input to our model that will be designed to verify the results.
The code is attached in Appendix A.
Fig.2 (Appendix A) shows the NDVI index calculated for shimoga. The data used to obtain this is collected from the USGS website, and is dated 14th April 2021. We see that the values are generously towards the positive side, making it very cultivable. Using this quantitative data regarding the vegetation, we aim to predict qualitative data regarding the type of crop to be grown.

Appendix A

Python3 notebook for generating NDVI using LANDSAT data of shimoga:

LANDSAT-8

In [21]:
```
shivmoga_nir_data_path = os.path.join("/home/ishan/Desktop/capstone/LC08_L2SP_145050_20210410_20210416_02_T1/LC0
shivmoga_red_data_path = os.path.join("/home/ishan/Desktop/capstone/LC08_L2SP_145050_20210410_20210416_02_T1/LC0
```

In [22]:
```
shivmoga_nir_data = rxr.open_rasterio(shivmoga_nir_data_path)
shivmoga_red_data = rxr.open_rasterio(shivmoga_red_data_path)
```

In [23]:
```
shivmoga_ndvi = es.normalized_diff(shivmoga_nir_data,shivmoga_red_data)
```

In [26]:
```
ep.plot_bands(shivmoga_ndvi,
              cmap='PiYG',
              scale=False,
              vmin=-1, vmax=1,
              title="shivmoga")
plt.show()
```

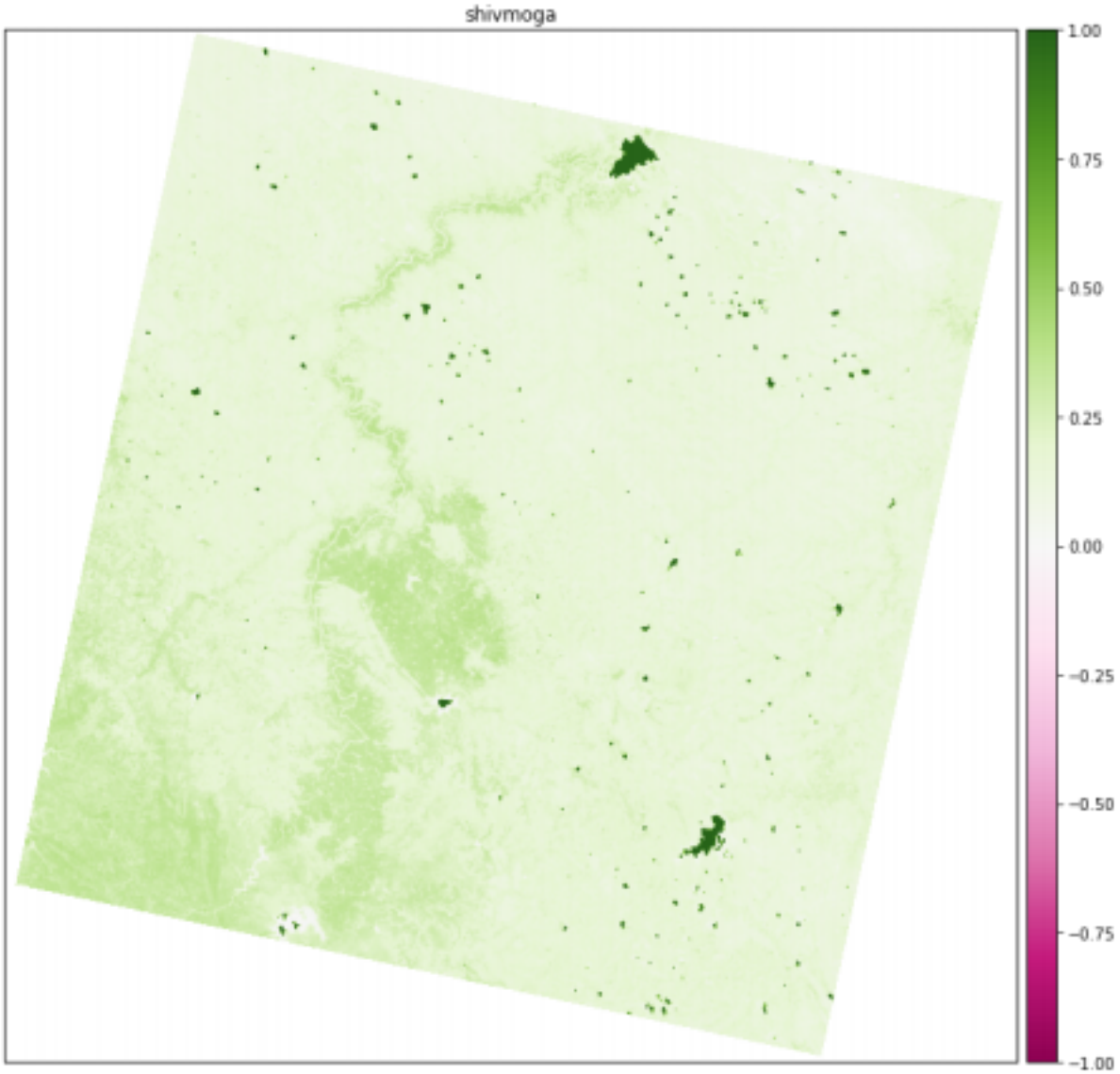


Fig. 2