# **PROJECT REPORT**

**Crop Suitability for the state of Andhra Pradesh using Machine Learning Techniques**

*BY*

**Akash Kumar Rao**

**Ishan Agarwal**

**Srish Srinivasan**

**Vishruth P Reddy**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

PES UNIVERSITY, BANGALORE

UNDER THE GUIDANCE OF

**Dr. S. RAMA SUBRAMONIAM**

**Scientist/Engineer ‘SF’**

**Prof. RAGHU B A RAO**

**Associate Professor, PES University**

In fulfilment of dissertation work carried out in   
Regional Remote Sensing Centre - South   
NRSC, ISRO  
Department of Space / Government of India  
ISITE Campus, Marathahalli, Outer Ring Road,  
Bangalore - 560 037  
KARNATAKA



**CERTIFICATE**

This is to certify that project entitled “Crop Suitability for the state of Andhra Pradesh using Machine Learning Techniques” is a bonafide work carried out by **Akash Kumar Rao, Ishan Agarwal, Srish Srinivasan and Vishruth P Reddy** of6th semester B.Tech in Computer Science and Engineering, PES University under my guidance during the period of January 2021 to July 2021. Certified further that to the best of our knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this to any other candidate.

**Dr. S. RAMA SUBRAMONIAM**

# 

# 

# 

# 

# 

# 

# 

# 

# **DECLARATION**

# 

# We hereby declare that this dissertation work “**CROP SUITABILITY FOR THE STATE OF ANDHRA PRADESH USING MACHINE LEARNING TECHNIQUES**", is a bonafide record of the work done by us, and no part of this has been presented for the award of any degree, diploma, associateship, fellowship or other similar titles of any other University or society.

# 

# 

# 

# **Akash Kumar Rao**

# **Ishan Agarwal**

# **Srish Srinivasan**

# **Vishruth P Reddy**

# 

# 

# 

# 

# 

# 

# 

# **ACKNOWLEDGEMENT**

We would like to express our gratitude to our external project guide Dr. S Rama Subramoniam, Scientist/Engineer ‘SF’, RRSC South, ISRO and our internal project guide Prof. Raghu B A Rao, Department of Computer Science and Engineering, PES University for their continuous guidance, assistance, and encouragement throughout the development of the project.

We would like to extend our gratitude to Dr. K Ganesha Raj, the General Manager of RRSC South, ISRO for accepting our request to work under the guidance of RRSC South, ISRO.

We also take this opportunity to thank Dr. U. Surendran, Scientist, CWRDM, Kozhikode for helping us in looking for data.

We would like to thank the management of PES University for giving us an opportunity to work on this project in collaboration with RRSC South, ISRO.

Finally, we would like to thank our parents and siblings for their complete cooperation and support throughout this project.

# 

# 

# 

# 

# **ABSTRACT**

Our project comes under the domain of Precision Agriculture. It helps farmers make informed decisions with regards to the kind of crop they must invest in to get good returns. 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 with the help of machine learning algorithms. A farmer or a horticulturist toils all day and keeps himself really busy throughout the year by looking after his crops and providing them with the right amount of water and nutrients expecting only one thing in return. And this is good crop yield. But if the farmer or the horticulturist makes a mistake in the very first step by choosing the wrong crop to cultivate, then all the remaining steps in the process are absolutely useless. So, our main goal in this project is to give the farmer or the horticulturist a sensible start to their cropping season by helping them choose the right crop to grow in their lands so that they are able to obtain a very good yield at the end of the harvest season. The focus is restricted to the state of Andhra Pradesh and we have made use of the historic crop yield data for that particular region and incorporated machine learning algorithms that ingest this data in order to build accurate models that could recommend the most ideal crop to be grown when fed with the properties associated with the soil of that region. We observed that the **Random Forest Classifier** and the **Gradient Boosting Classifier** were the **most accurate classifiers** among all the models that have been employed yielding accuracies in the range of 70% to 85%.

## 

## 

## 

## 

## 

## 

## 

## **TABLE OF CONTENT**

INTRODUCTION 6

PROBLEM STATEMENT 9

LITERATURE REVIEW 10

PROJECT REQUIREMENTS SPECIFICATION 29

HIGH LEVEL DESIGN DOCUMENT 32

IMPLEMENTATION AND PSEUDOCODE 35

CONCLUSION 56

LIMITATIONS 57

BIBLIOGRAPHY 58

ABBREVIATIONS 59

**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, 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 require 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.

**PROBLEM STATEMENT**

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 properties of a specific location. We will be narrowing down our focus onto a specific part of India. The properties of the soil that are to be taken into consideration could include its nutrient contents and their quantities. The nutrients could include both macronutrients as well as micronutrients. Macronutrients are those nutrients that are produced by the soil in relatively larger quantities. Some examples of macronutrients include N, P, K, Ca, S, Mg, C, O, and H. On the other hand, micronutrients are those nutrients that are produced by the soil in relatively smaller quantities. Some examples of micronutrients include Fe, B, Cl, Mn, Zn, Cu, Mo, and Ni. These elements are present in the form of salts beneath the soil and are absorbed by the plants in the form of ions. Other soil properties include pH, EC, moisture, temperature, etc. On collecting the required data, we plan to use machine learning algorithms in order to build robust models that can make forecasts with a very good accuracy level. Our project comes under the domain of precision farming. Precision farming can reduce the quantity of nutrients and other requirements used to a large extent while boosting the yield by a large margin. Farmers can thus obtain great returns on their investments and can also save big on fertilizer, pesticide, water and other resources. The next big advantage of practicing precision farming is the prevention of malefic impact on our environment by using the right quantity of chemicals and thereby conserving the quality of soil and groundwater.

**LITERATURE SURVEY**

**Paper1**

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 value 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:

Xi = vector of independent variables

hθ = hypothesis function

Yi = True value of the response variables

m = normalizing 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 of by this algorithm.

**Paper 2**

Yield rate of crops is dependent on two broad factors, the first being genetic development of seeds which lies more in 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:

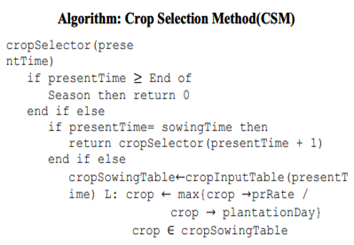
a) Seasonal crops, and

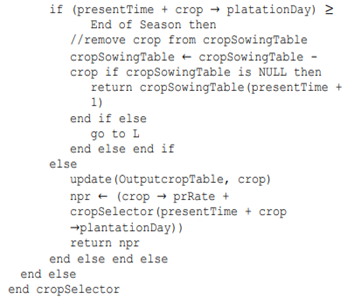
b) Whole year crops

c) Short time plantation crops, and

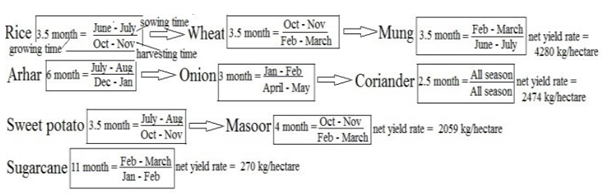
d) Long time plantation crops

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 a waste of farmers’ resources. Below is the algorithm used in this part of the research work.





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 the 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.

**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 |  |
| Mean Squared Error |  |
| Mean Absolute Error |  |

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. Rice is also

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.

**Paper 4**

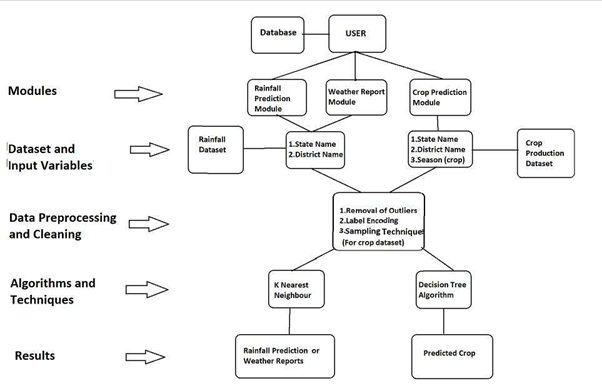
In this paper, the authors have taken into account the different ML algorithms which are used in crop prediction over various other studies and have tried to add more attributes to the system in order to improve the results. They have compared the prediction of the ideal crop from using different models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

* Seasonal
* Whole Year
* Short plantation period
* 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.



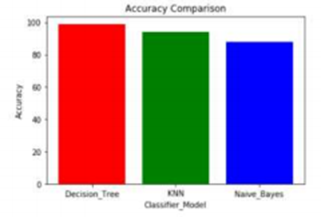
The data must include:

* Soil Parameters
  + 1. Soil Type
  + 2. Soil Ph
* Climatic Parameters like humidity, temperature, wind, rainfall
  + 1. Humidity
  + 2. Temperature
  + 3. Wind
  + 4. Rainfall
* Production
* Cost of cultivation
* Previous year yield results

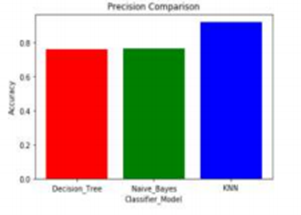
The data is pre-processed and fed into KNN, Decision tree and Naive Bayes classifier and the results from all of these are compared. (The selection attributes of the Decision tree being Gini index, entropy and information gain.)

The results were as follows:

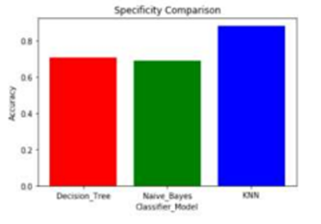
Accuracy:



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 tree for such datasets. The combination classification algorithms like naïve bayes and decision tree classifiers are better performing than the use of a single classifier model. Thus, we can make use of this study and also cross check the findings when using different models.

**Paper 5**

The paper throws light on the usage of deep neural networks to predict the crop yield using various parameters like the environmental conditions and the genotype of the crop. A root mean squared error of 12% was achieved which was reduced to 11% when accurate weather data was taken into account. The DNN model implemented gave better accuracy scores than the other models they tried to implement such as Lasso Regression, Regression Trees and Shallow Neural Networks. The main understanding from the paper was that environmental conditions (soil and atmospheric parameters) have more impact on the crop yield than the genotype. The crop under consideration was Maize.

The neural network they used had the following features:

* 21 hidden layers in each NN
* 50 neurons in each layer
* 3,00,000 maximum iterations
* Batch normalization and Adam Optimizer were used

The above hyperparameters reduced overfitting and improved the accuracy of the model. Genotype, soil and weather parameters were compared to see the extent of influence. The following table provides the details:

| Model | Training RMSE | Training Correlation Coefficient (%) | Validation RMSE | Validation Correlation Coefficient (%) |
| --- | --- | --- | --- | --- |
| DNN (G) | 21.74 | 20.26 | 21.72 | 15.09 |
| DNN (S) | 15.28 | 73.37 | 15.49 | 72.04 |
| DNN (W) | 14.26 | 76.98 | 14.96 | 72.60 |
| Average | 24.40 | 0.0 | 23.14 | 0.0 |

Here, DNN(x) refers to the performance of the models on individual parameters without the consideration of the other parameters.

From this, we can see that:

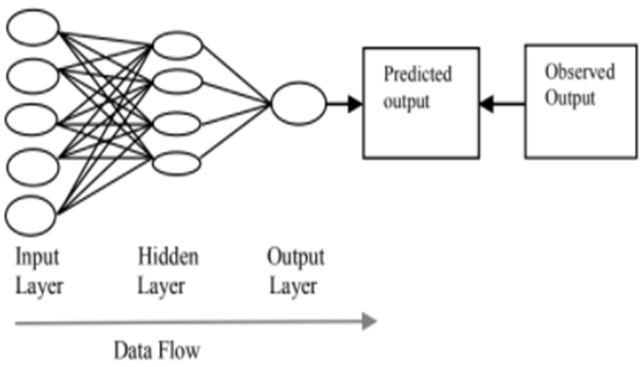
* Soil and weather DNNs have similar performance measures
* DNN(S) and DNN(W) have higher accuracies than DNN of genotype.

We can conclude that weather and soil parameters are essential and can’t be excluded as their impact on the yield has great significance. The black box property was eliminated by performing feature selection on the trained model making use of back propagation methodology.

**Paper 6**

The paper provides insights upon the need for prediction. Growing crops takes months of hard work and tremendous amounts of resources. If the right crop can be recommended, it will be very beneficial. Fuzzy systems, Genetic algorithms and Artificial neural Networks have provided great efficiency and accuracy in prediction. Reducing the scope of study to only ANN, the computational model they used was feed forward back propagation neural network.

The reason for choosing ANN was because it can find the factor that is most impacting the yield. Just like any other ANN, they have used a model that contains an input layer, the interconnected hidden layer and an output layer as shown below:



Using the back propagation algorithm, which uses the gradient descent method, they were able to achieve a high accuracy. They were able to predict cotton, sugarcane, jowar, bajra, soybean, corn, wheat, rice and groundnut using ANN with a good accuracy score.

The following are the requirements for the crops to grow:

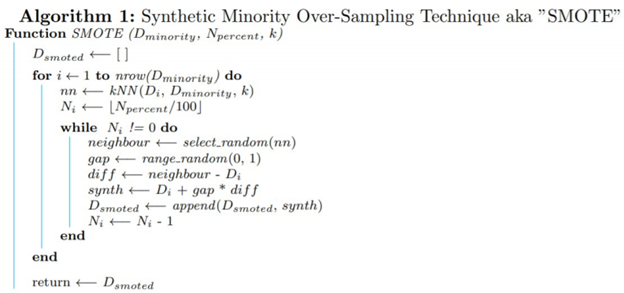
| Crop | pH | Nitrogen (ppm) | Depth (ppm) | Temp (Celsius) | Rainfall (cm) |
| --- | --- | --- | --- | --- | --- |
| Cotton | 7-8.5 | 40 | 30 | 27-33 | 700-1200 |
| Sugarcane | 6.5-7.5 | 40 | 60 | 20-50 | 750-1200 |
| Jawar | 6.0-8.5 | 132-180 | 50-20 | 25-30 | 800-1000 |
| Bajara | 7-8.5 | 120 | 15 | 28-32 | 400-750 |
| Soyabean | 6.5-7.5 | 37 | 15-20 | 25-33 | 700-1000 |
| Corn | 7.5-8.5 | 60-120 | 5 | 13-30 | 500-600 |
| Wheat | 6-8.5 | 80-150 | 15-20 | 16-22 | 25-180 |
| Rice | 5.5-8.5 | 50 | 50-20 | 22-25 | 1000-1500 |
| Groundnut | 6-7.5 | 25 | 20 | 24-27 | 500-1250 |

**Paper 7**

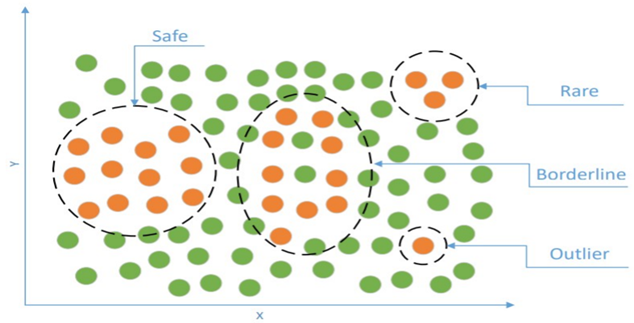
The author of this paper aims to understand the impact of instances belonging to the minority class on the performance of SMOTE by performing a selective preprocessing of these instances. Most contemporary classification techniques make an assumption that all the classes in the training dataset have almost equal number of instances under their name. However, this assumption need not always come true and in the case of an imbalance, the classifier will end up making predictions in the favour of the majority class. So, our goal must be to ensure that the classifier performs well when it comes to the minority classes and at the same time make sure that their performance with the majority classes doesn’t drop by a large margin.

The data-level solution to this problem aims at balancing the data that is used to train the model. The algorithm-level solution to this problem looks into the specifics of the algorithm being used and if a drawback or 2 are found, it tries to improve the algorithm in those areas. This problem is also being tackled by the use of an ensemble of classifiers which could put up a better show collectively as compared to working as individual entities.

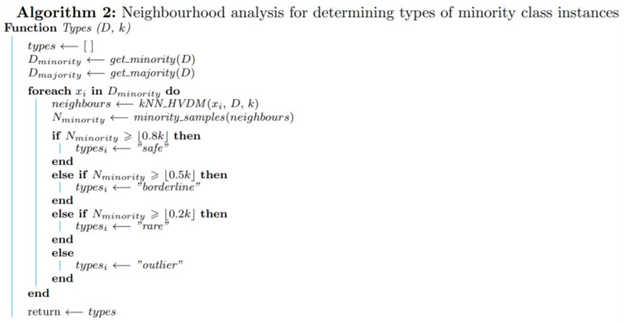
Data-level approaches work independent of the classifier being used. Under this category, SMOTE was a very popular algorithm that was being employed despite the fact it too had some drawbacks. One of them is to treat all minority instances equally. But being more selective when it comes to oversampling the minority instances yielded a better overall accuracy accordingly. Fortunately, there are few solutions that focus on this drawback of SMOTE. Borderline-SMOTE focused on those instances that sit close to the class demarcations. This idea of Borderline-SMOTE was further developed into a technique called ADASYN which could dynamically pick those instances that are more problematic to the classifier. Safe-Level-SMOTE allocates weights to instances based on how much the majority class influences these instances and makes use of these weights for the generation of artificial instances. A technique called SPIDER focuses on those instances that overlap with the majority class.



SMOTE is computationally quite complex and its memory requirements are quite significant. When working with imbalanced data, apart from the class disparity, the characteristics of the minority class instances are also imperative. In most cases, these minority class instances have a crystal-clear demarcation through the formation of heterogeneous structures. Some of these structure types are as follows.



The algorithm used to obtain the above segregation is as follows and the distance metric used here is the Heterogenous Value Difference Metric and not Euclidean distance.



After having determined the types of minority class instances, a base classifier was tested in different preprocessing configurations which is achieved by oversampling only selected types of minority class instances. In conclusion, the author stated that data-driven oversampling improved the performance of the classifier as compared to a uniform oversampling technique. Therefore, the imbalance ratio is not the only cause for learning obstacles, but the properties of the instances belonging to the minority classes need to be looked into as well and thereby enabling SMOTE to oversample only selected instances from the dataset.

**PROJECT REQUIREMENTS SPECIFICATION**

# **Introduction**

Our project comes under the domain of Precision Agriculture. It helps farmers make informed decisions with regards to the kind of crop they must invest in to get good returns.

**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. This results in greater yield of crops 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.

# **Product Perspective**

Usually, farmers and horticulturists don’t have a firm idea as to what is the best crop to be grown due to the limited knowledge of the soil parameters and the surrounding conditions. This often results in poor yield of crops which impacts the farmers financially thereby instigating the farmers to take extreme measures like committing suicide. Therefore, our project is a return of favour to all those hard-working men and women who slog all day at the fields so that we are able to consume nutritious food.

# **Product Features**

Our product ingests parameters that describe the soil properties such as N, P, K, pH, etc as input and outputs the name of the most suitable crop that could be grown in order to achieve maximum yield and have a successful harvest season.

# **Operating Environment**

Our plan is to build a web application that provides a simple user interface for the farmers to interact with in order to make informed decisions with regards to crop selection.

# **General Constraints**

* Good quality network connection to use the web application

**Risks**

The data which we are utilizing must be from a reliable source as farmers will be investing their time, efforts and resources in growing the crop recommended by our model with the aim of maximizing their profits. We must also ensure that the mobile application is lightweight so that it can function efficiently even if there’s a fluctuation in the network connectivity.

**Functional Requirements**

* After finishing with all the installation and setup, the user needs to input the soil parameters requested by the application.
* The application validates the parameters input by the user and raises an exception in case of an erroneous input. It then prompts the user to change the value and this continues until all the parameters are correctly input by the user.
* The application passes these values to the machine learning models and returns to the user the name and details of the crop that is most suitable to be grown based on the results obtained from the analytics.

**Hardware Requirements**

A good quality network connection is necessary for using the web application and also for downloading the mobile application. However, on successfully installing the mobile application, network connection is no longer required to use the same. The mobile running the application is required to have a minimum of 4 GB of RAM and must have android as its operating system.

**Performance Requirement**

* Windows 10
* 4 GB of RAM (Minimum)
* Good quality network connection (Wi-Fi or Ethernet)
* Web Browser like Chrome, Firefox

**Security Requirements**

The client will have to create an account in order to make use of our application. The client’s login credentials will be stored in an encrypted format and it will be made sure that no other user is able to compromise any other fellow user’s account.

**HIGH LEVEL DESIGN DOCUMENT**

**Introduction**

The main goal of our project is to predict the most suitable crop to a farmer or a horticulturist based on the soil parameter values that are entered in the mobile or web application. This document gives a detailed description of the website and the machine learning models being used to predict the right crops.

**Design Considerations**

**Design Goals**

* The existing applications require the farmer to create an account using email IDs and by providing card details.
* Some applications even charge for using their product.
* Our website is very simple as we do not collect any personal information from the user. All they have to do is enter their phone number and create a password.
* Once they are logged in, all that a farmer has to do is to just input the soil details and let the machine learning models do their job.
* No payment or addition of personal details is required.
* We respect the privacy of our users so we do not collect any personal details.
* Since we aren’t collecting any personal information like name, age, salary, card details and other details like farm location or farm ID, so security is also maintained.
* As soon as the user enters all the values asked for, the website immediately starts to process the data and predicts the perfect crop within a few seconds.

# 

# 

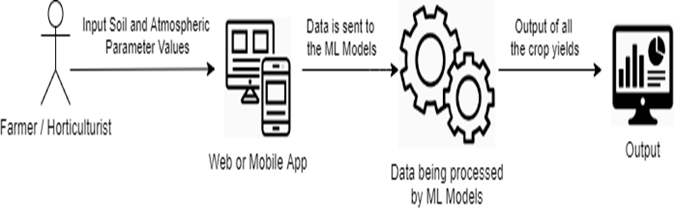
# 

# 

# **Constraints, Assumptions and Dependencies**

* **Interoperability requirements**: There are no interoperability requirements in our project.
* **Interface/protocol requirements**: All that is necessary is that the users have to enter the soil data on the website.
* **Data repository and distribution requirements**: All the soil details added by users will be stored in the firebase for training the models and improving their predictive capacity.  
  There will not be any platform related issues as it is a very simple website and easy to operate.
* **End-user environment**: In the end-user environment, the output of the prediction will be very clear and self-explanatory. It needs no technical knowledge.
* **Hardware or software environment**:
  + There is no hardware component for our project.
  + The software component will be the website that the users will use for prediction.

**High Level System Design**

****

# 

# **Design Details**

# **Novelty**

# The problem statement that we are solving here is not new and a few solutions have already been built in the past. But the difference in the approaches arises due to the different machine learning algorithms that have been employed to solve the problem. Instead of training the models on the entire state, we trained the models on each district separately to achieve better accuracy.

# **Innovativeness**

# We have incorporated innovativeness into our project by employing the most accurate machine learning technique and tried to minimize any computational overhead to the greatest extent possible. Our goal is to make the user interface extremely simple and user friendly.

# **Interoperability**

# We have ensured that the exchange of information between the machine learning models and the website is extremely smooth and quick.

# **Performance**

# The website functions very efficiently and thereby produces the required results in quick time.

# **Security**

# The application requires a very minimal set of user data, which includes First Name, Last Name, Contact Number and City. We will make sure that the user data is stored in an encrypted format and is denied access to any other person apart from the user himself.

* **Reliability**

We have made sure that the results generated by the website is very reliable by making use of extremely reliable data in order to build and train the machine learning models.

* **Maintainability**

We have designed our website in a modular approach. Every component of the website is a module in itself, therefore the website is maintenance friendly and in case of any modifications or additions, their implementations can be integrated with the existing website with absolute ease.

* **Portability**

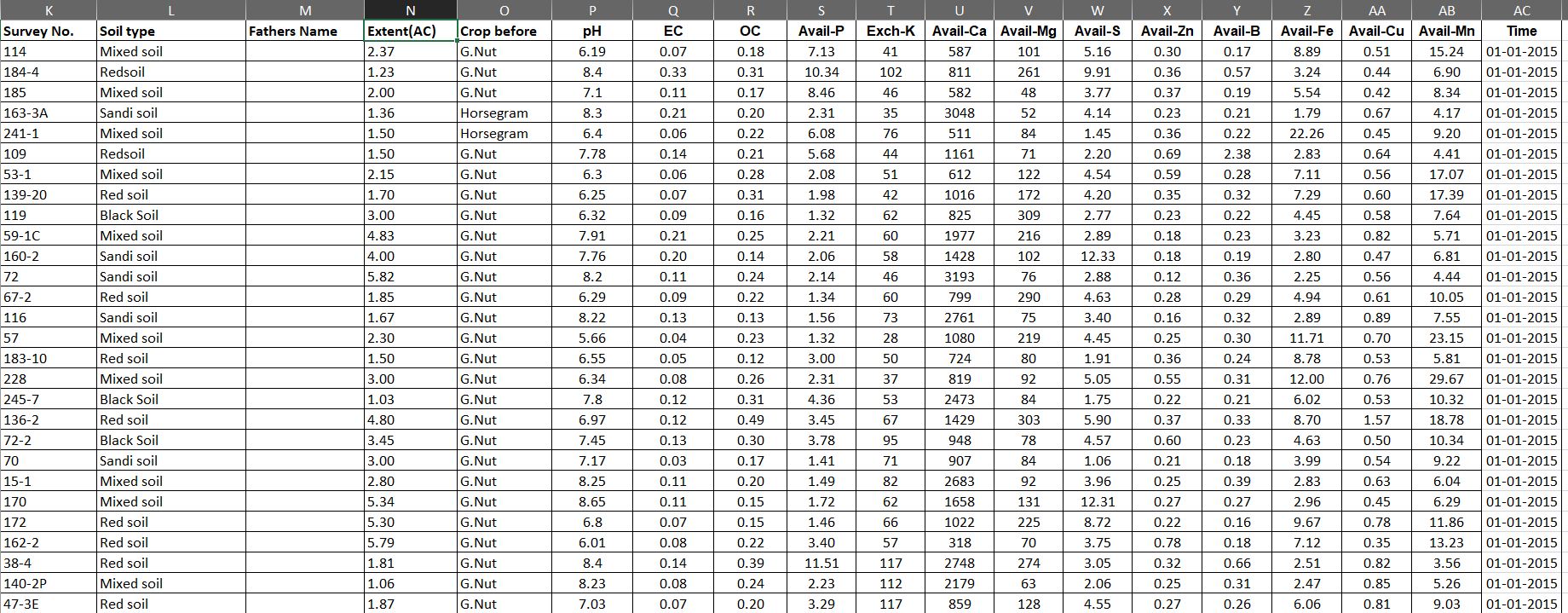
The website could be used both on PC browsers as well as android smartphone browsers. Therefore, the website is portable.

**IMPLEMENTATION AND PSEUDOCODE**

The following is a snapshot of the Crop Suitability dataset for the state of Andhra Pradesh covering 13 districts.

Number of columns: 19

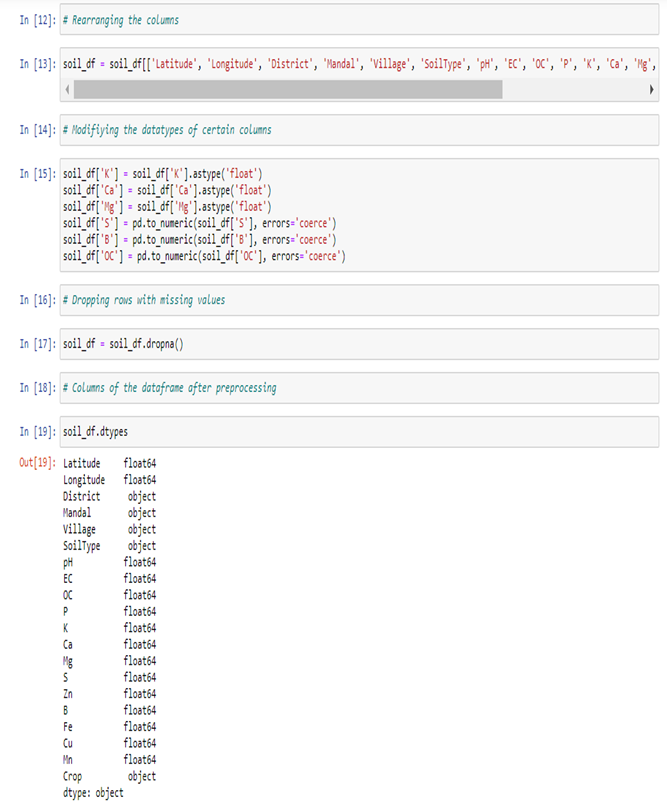
Number of rows: 4816

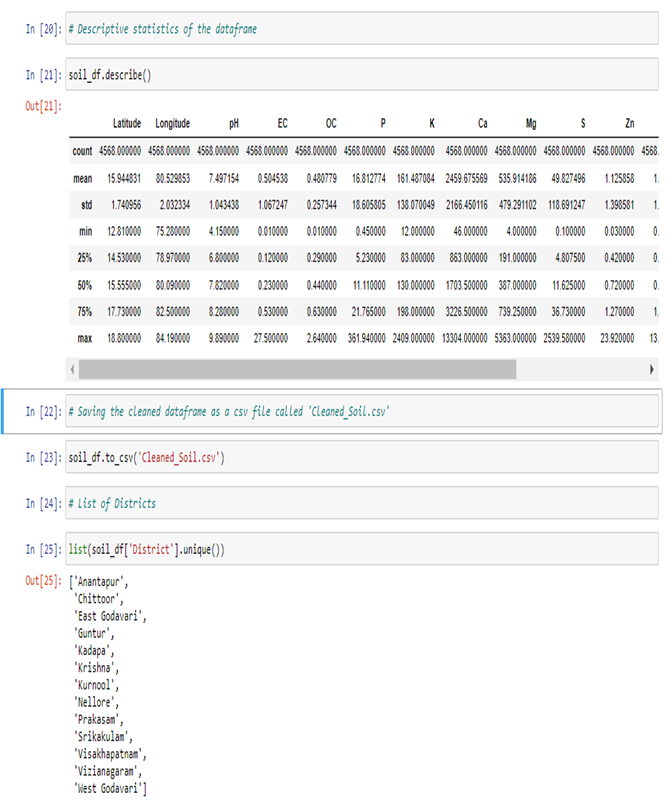


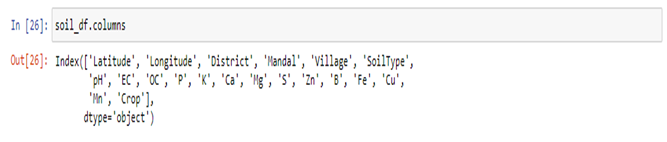
**Data Preprocessing**

****

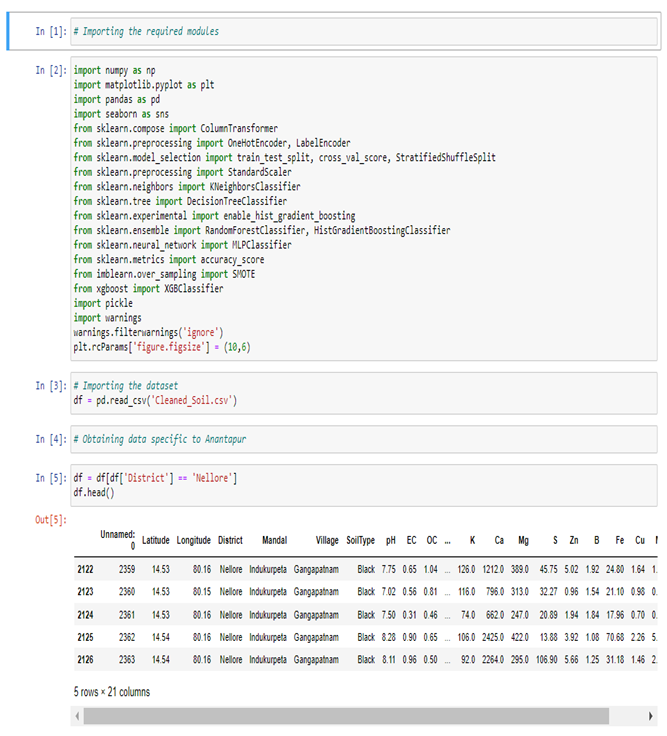
****

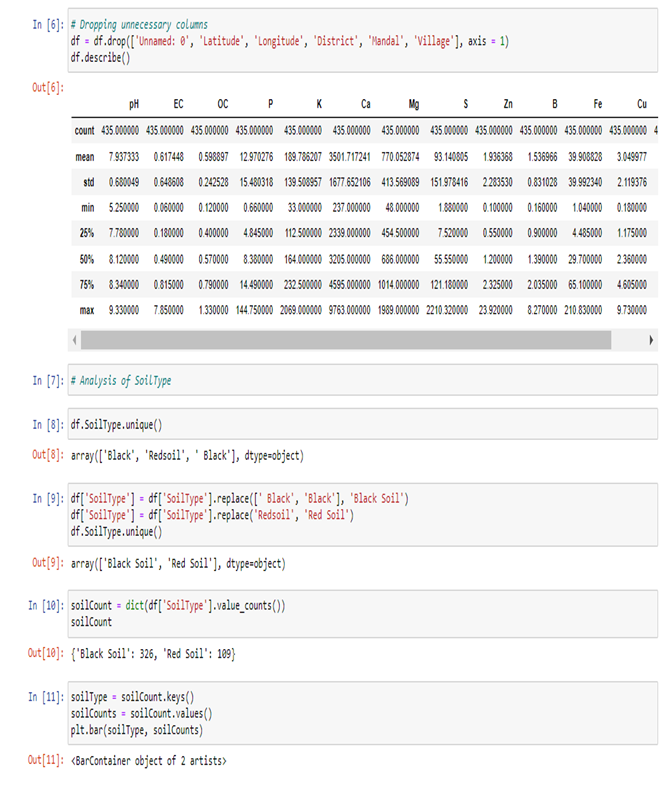


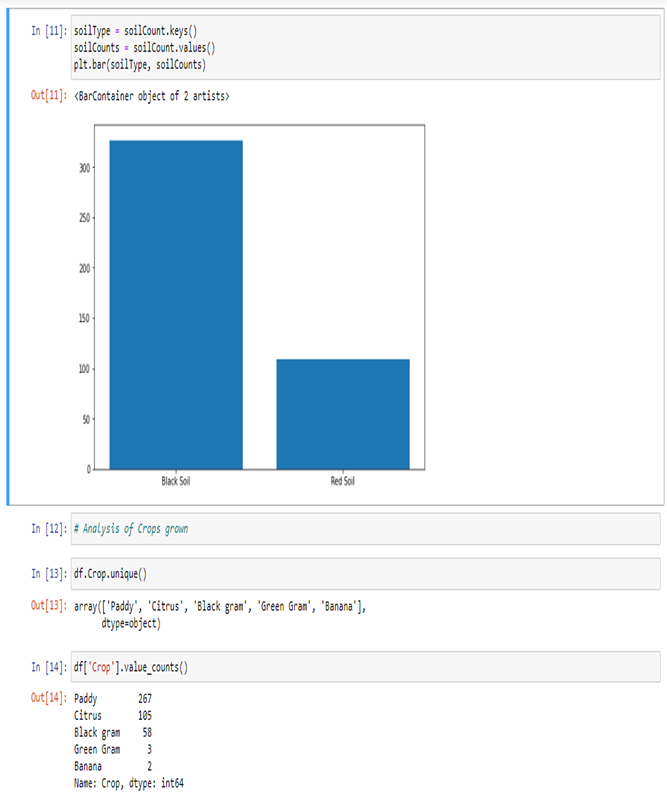


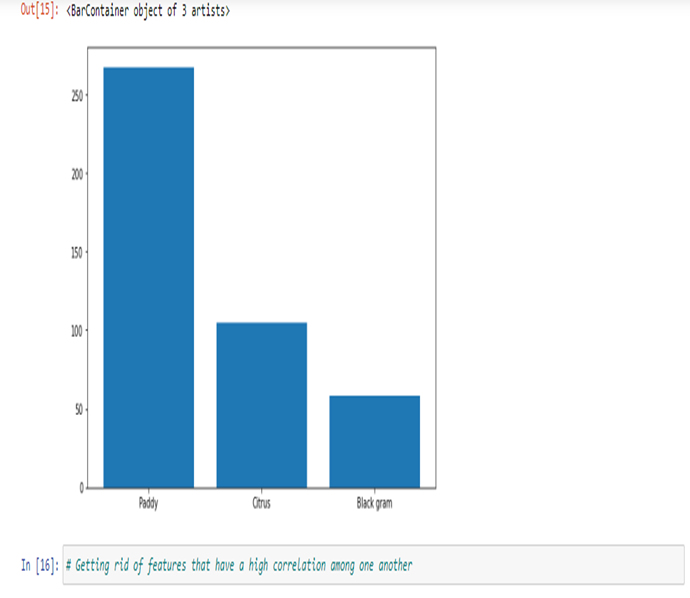


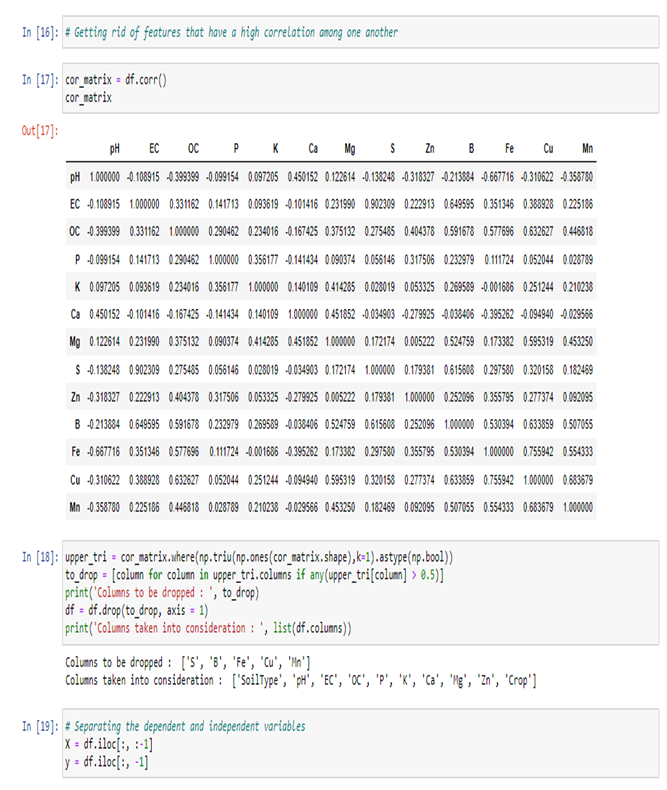
**Analysis for the district of Nellore**

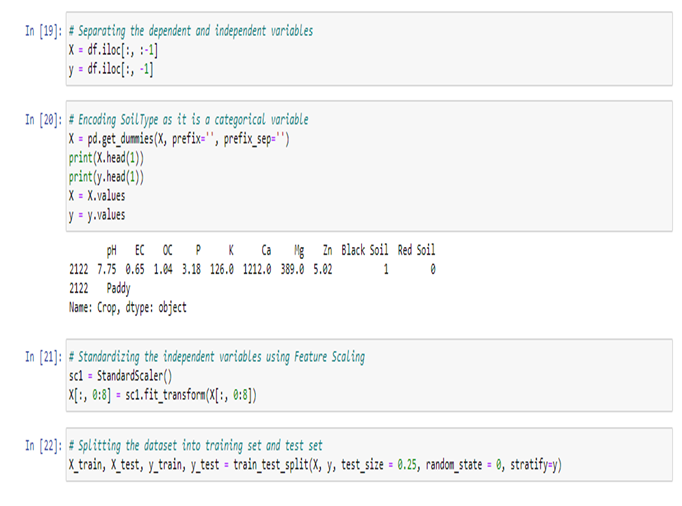
****

****

****

****

****

****

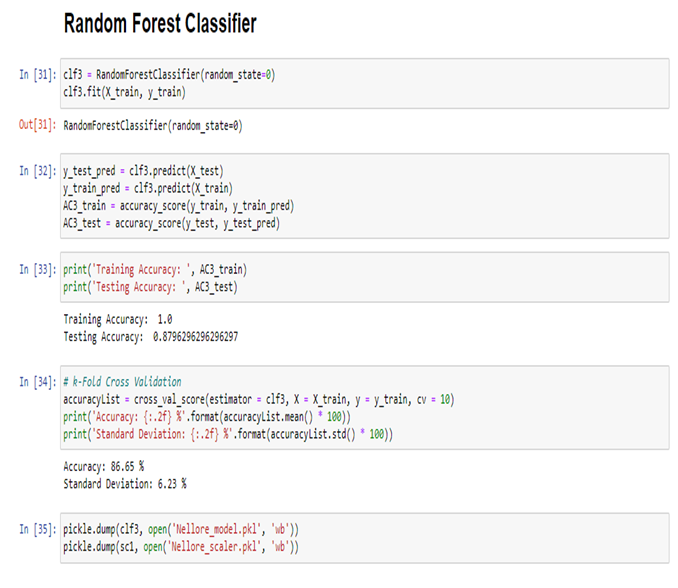
**Model 1: K-Nearest Neighbors Classifier**

****

**Model 2: Decision Tree Classifier**

****

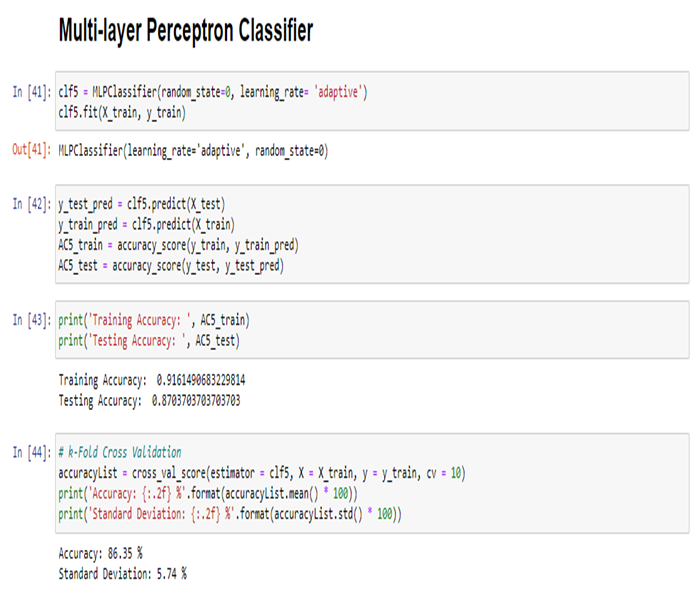
**Model 3: Random Forest Classifier**

****

**Model 4: Gradient Boosting Classifier**

****

**Model 5: Multi-layer Perceptron Classifier**

****

| ***Machine Learning Technique*** | ***Training Set Accuracy %*** | ***Testing Set Accuracy %*** |
| --- | --- | --- |
| **K-Nearest Neighbours Classifier** | **100** | **87.04** |
| **Decision Tree Classifier** | **100** | **80.60** |
| **Random Forest Classifier** | **100** | **88.00** |
| **Gradient Boosting Classifier** | **100** | **86.11** |
| **Multi-layer Perceptron Classifier** | **91.61** | **87.04** |

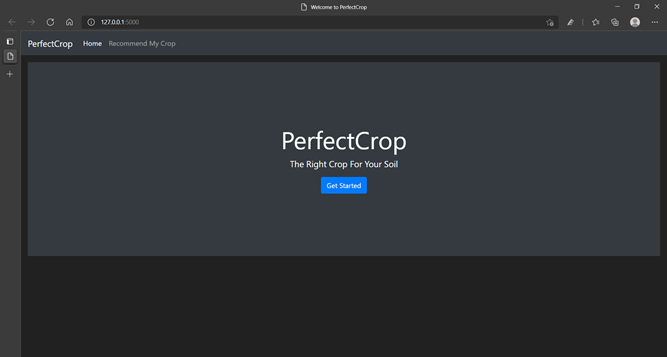
**Therefore, the Random Forest Classifier was chosen as it has the highest accuracy percentage of 88.**

A similar process was employed for the analysis of the remaining 12 districts and the results obtained were as follows. (The best model for each district along with their training and testing accuracy percentages are as follows)

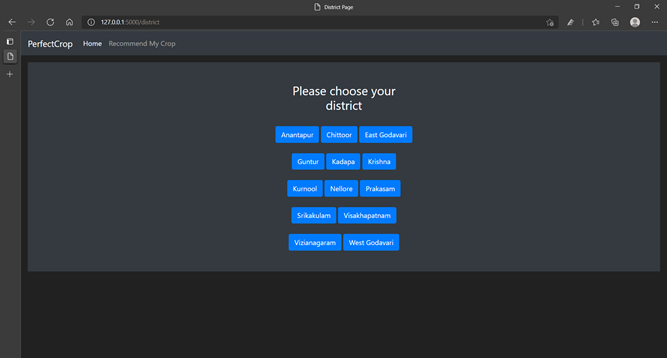
| ***District*** | ***Machine Learning Technique*** | ***Training Set Accuracy %*** | ***Testing Set Accuracy %*** |
| --- | --- | --- | --- |
| **Anantapur** | **Random Forest Classifier** | **100** | **84.84** |
| **Chittoor** | **Gradient Boosting Classifier** | **100** | **86.2** |
| **East Godavari** | **Gradient Boosting Classifier** | **92.86** | **68.42** |
| **Guntur** | **Random Forest Classifier** | **100** | **83.1** |
| **Kadapa** | **Random Forest Classifier** | **100** | **68.58** |
| **Krishna** | **Random Forest Classifier** | **100** | **72.13** |
| **Kurnool** | **Random Forest Classifier** | **99.66** | **79.41** |
| **Prakasam** | **Random Forest Classifier** | **100** | **77.12** |
| **Srikakulam** | **Random Forest Classifier** | **100** | **72.38** |
| **Visakhapatnam** | **Random Forest Classifier** | **100** | **81.2** |
| **Vizianagaram** | **Random Forest Classifier** | **100** | **85.03** |
| **West Godavari** | **Random Forest Classifier** | **100** | **82.9** |

**Navigation through the Website**

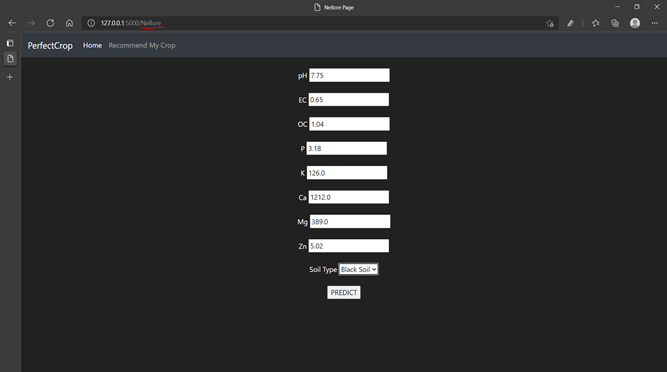
1. The following picture is a depiction of the web application’s home page. The user must click on the button with the text “**Get Started**” or could click on the option called “**Recommend My Crop**” present in the navigation bar in order to land up on the page where the district needs to be chosen.

****

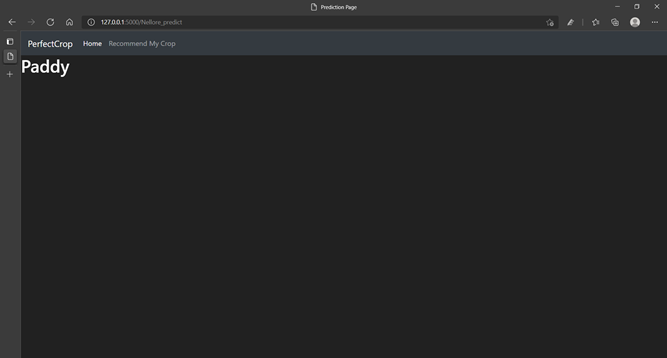
1. Now, the user will have to choose the district for which he or she wants to predict the most suitable crop to grow by simply clicking on the button that corresponds to the district of interest.



1. On clicking on any one of these districts, the user will be directed to a page with a form that takes in values pertaining to the levels of soil nutrients present in the soil and for some districts, even the soil type is taken as one of the inputs. In this example, we have depicted the scenario for the district of Nellore.



1. On entering all the required values and clicking the predict button, the user will be directed to a page where the name of the most suitable crop to be grown will be displayed on the screen.



**CONCLUSION**

We are a team of Computer Science and Engineering students, currently in our third year, interested in the field of Data Science and Data Analytics. At the same time, we wanted our project to have a social impact. Therefore, we decided to come up with an idea that would aid the farmers in making the right decision with regards to the choice of crops to be grown.

Through the literature survey that we conducted, we learnt about the various machine learning algorithms that have been employed to recommend the most suited crop given the required soil properties of a specific location. On thorough searching, we also came across a relevant dataset for the state of Andhra Pradesh covering 13 districts which also seemed to adhere to the guidelines given to us by our external guide. Therefore, we began to preprocess the dataset and make it suitable enough to perform various analyses for the same. At first, we began to build various machine learning models based on algorithms such as **k-Nearest Neighbors, Decision Trees, Random Forests, Gradient Boosting** and **Artificial Neural Networks.** But the accuracy percentages were not meeting our expectations. On taking a closer look at our implementation and the dataset, we realized that the number of crop labels were a lot and quite a few crops grew in similar conditions which lead to potential misclassification. So, we decided to perform a district specific analysis rather than clubbing all the districts under one umbrella. We then observed that the **Random Forest Classifier** and the **Gradient Boosting Classifier** were the **most accurate classifiers** among all that we had employed. This method helped us achieve very good test set accuracies for each district and therefore gave a meaningful end to our project with all our objectives completed with a good scientific backing.

**LIMITATIONS**

The crop suitability data that we have used in this project is restricted to only the state of Andhra Pradesh. Within the state of Andhra Pradesh, this data only covers 13 districts. So, our aim in the future would be to find more data which covers more states in the country. Additionally, further research can be done in this field by adding geospatial input to the existing models in order to try and obtain more accurate results.

**BIBLIOGRAPHY**

**[1]** Keerthan Kumar, T.G., Shubha, C. and Sushma, S.A., **Random Forest Algorithm for Soil Fertility Prediction and Grading Using Machine Learning**. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2019*.

**[2]** Kumar, R., Singh, M.P., Kumar, P. and Singh, J.P., 2015, May. **Crop Selection Method to maximize crop yield rate using machine learning technique**. In 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM) (pp. 138-145). IEEE.

**[3]** Palanivel, K. and Surianarayanan, C., 2019. **An approach for prediction of crop yield using machine learning and big data techniques**. International Journal of Computer Engineering and Technology, 10(3), pp.110-118.

**[4]** Patil, A., Kokate, S., Patil, P., Panpatil, V. and Sapkal, R., 2020. **Crop Prediction using Machine Learning Algorithms**. International Journal of Advancements in Engineering & Technology, 1(1), pp.1-8

**[5]** Khaki, S. and Wang, L., 2019. **Crop yield prediction using deep neural networks**. Frontiers in plant science, 10, p.621

**[6]** Dahikar, S.S. and Rode, S.V., 2014. **Agricultural crop yield prediction using artificial neural network approach**. International journal of innovative research in electrical, electronics, instrumentation and control engineering, 2(1), pp.683-686.

**[7]** Skryjomski, P. and Krawczyk, B., 2017, October. **Influence of minority class instance types on SMOTE imbalanced data oversampling**. In the first international workshop on learning with imbalanced domains: theory and applications (pp. 7-21). PMLR.

**ABBREVIATIONS**

pH: power of Hydrogen

EC: Electrical Conductivity

Mg: Magnesium

K: Potassium

AI: Artificial Intelligence

ANN: Artificial Neural Networks

N: Nitrogen

P: Phosphorus

K: Potassium

Ca: Calcium

S: Sulphur

Mg: Magnesium

C: Carbon

O: Oxygen

H: Hydrogen

Fe: Iron

B: Boron

Cl: Chlorine

Mn: Manganese

Zn: Zinc

Cu: Copper

Ni: Nickel

Mo: Molybdenum