**EFFICIENT CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHMS**

**ABSTRACT**

Starting stage of analytics will be the descriptive analysis. using this method we can identify about that what happened earlier and then it will come to know that past is the better analyzer for further. For agriculture production domain we are using descriptive analysis for the crop to find the most efficient crop field estimation. in this are using three datasets as Soil, Rainfall, Yield .then forming a dataset and on this it will apply various supervised techniques to find the estimated cost and the accuracy of many number of techniques. Here we are using three supervised methods are used like as random forest, linear regression. It is a comparative study which tells the accuracy of training proposed model and error rate. The accuracy of training model must be more and the error rate must be minimum.

**CHAPTER 1**

**PREAMBLES**

**1.1 INTRODCUTION**

Among all the industries in India, agriculture has the major role and is mostly depending up on it for rural support. Because of some issues like changing climate, unpredicted rainfall, low water level, usage of pesticides heavily etc. the status of agriculture in India is reduced. Hence to know the status of production we have shown descriptive analysis on the agriculture data. The objective of this work is to produce kind of methodology by that it can easily perform descriptive analysis on production of crop yield in a sufficient way. Moreover some studies says statistical data regarding the agriculture in India, few more studies have identified crop prediction depending up on the climatic and production data. ANNs accept been adopted to a new environment for various reasons which involves classification, clustering, agent quantization, arrangement association, action approximation, forecasting , and optimization.

Here we are using two kinds of algorithms which are random forest regression, linear regression. They are used to predict the crop yields depending up on the data which is provided from Telangana state in India and performing to get best results.

**1.2 PROBLEM STATEMENT**

The issue of agriculture is affected by several climate factors. As Humidity, wind speed, temperature, and moisture. Precipitation and soil and because of continuously change in climate condition everything is messed. In India farmers still follow the traditional technology which they adopted from their ancestor. But the problem is that in earliest time climate was very healthy everything was happened on time. But now most of the things have been changed due to global warming and many other factors. major issue with agriculture is lack of rainfall in season time. Winter season is been affected so Rabi crops are widely affected. Since few years the rainfall in winter season was high as expected. To overcome these above issues we need to develop a system which will able to find the hidden facts or results, patterns and insights. The farmer can predict which crop should sow so that he/she can get more benefit. In proposed system we are applying data analytics techniques on agriculture production based datasets and find the insights so that it can help to the farmers and their decision making. In this we are proposed a system which is based on descriptive analytics. By which farmers can know what happened in past time and what is going to occur.

* 1. **OBJECTIVES**

In our work the harvest yield arrangement will perform to classify based on yield efficiency and class marks will be low, mid, and high. Also, scope of efficiency will be characterized set up for assessing yield much before genuine collect of the harvests. By utilization of observational measurable models utilizing connection and relapse system crops yield are figure on an operational reason for the nation. Meteorological parameters at different harvest development arranges alongside mechanical patterns are utilized in the models. And this research will also helpful if in future we make a complete recommender system for farmers. Because here we are performing descriptive analytics which is the base or foundation of any recommender system. And relapse will be performed to get the real harvest yield assessed cost.

* 1. **SCOPE OF THE STUDY**

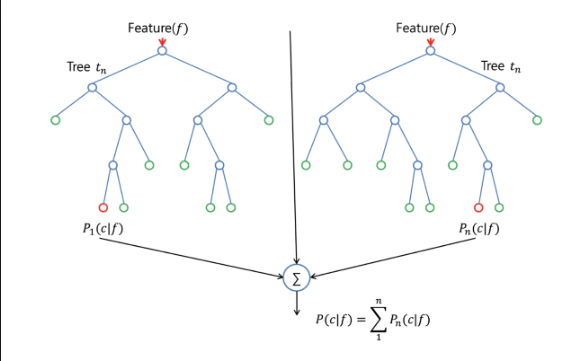
Agriculture is the backbone for a developing economy like India and there is an enormous need to maintain the agricultural sustainability. Hence it is a significant contribution towards the economic and agricultural welfare of the countries across the world.

* 1. **METHODOLOGY**

There are three dataset named as Soil dataset, Rainfall dataset, Yield dataset. These datasets includes several parameters which are helpful to know the condition of crops and classify the data into separate classes by performing supervised training on the dataset that are collected from agriculture domain. This system has the capability to perform both the classification as well as regression. In the classification step the data is classified into three classes (low, mid, and high), whereas in regression step the actual cost of yield production is estimated. We used three major algorithms of supervised learning such as random forest regression, linear regression

**RANDOM FOREST REGRESSION:-**

Random forest regression is a flexible, clean to apply system mastering algorithm that produces fantastic effects most of the time with minimum time spent on hyper-parameter tuning. It has gained recognition because of its simplicity and the reality that it may be used for each type and regression responsibilities. In this i will present in information the random wooded area regression version.



**LINEAR REGRESSION.**

Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple.

### ****Simple Linear Regression****

Easy linear regression is useful for finding dating between two continuous variables. One is predictor or independent variable and different is response or structured variable. It appears for statistical dating but now not deterministic dating. Relationship between two variables is stated to be deterministic if one variable may be appropriately expressed by way of the opposite. For example, the usage of temperature in degree celsius it is viable to correctly predict fahrenheit. Statistical relationship isn't always accurate in determining dating between two variables. For instance, dating between top and weight.

**Multiple linear regression (MLR),**

A multiple of linear regression (mlr), also recognized truely as multiple regression, is a statistical technique that uses numerous explanatory variables to are expecting the final results of a response variable. The purpose of a multiple of linear regression (mlr) is to model the linear dating among the explanatory (impartial) variables and reaction (established) variable. In essence, more than one regression is the extension of everyday least-squares (ols) regression that entails more than one explanatory variable.

**1.6 LITERATURE SURVEY**

On this paper [2] crop system is a multi-yr, multi-crop, daily time step cropping systems simulation model evolved to function an analytical tool to study the impact of weather, soils, and control on cropping systems productivity and the surroundings. Crop system simulates the soil water and nitrogen budgets, crop boom and development, crop yield, residue production and decomposition, soil erosion by water, and salinity. The improvement of crop system started out inside the early nineteen nineties, evolving to a collection of programs inclusive of a cropping structures simulator (crop system), a weather generator (climgen), gis-crop system cooperator application (arccs), a watershed model (crop system watershed), and numerous miscellaneous utility applications. Crop system and associated packages may be downloaded free of charge over the internet. One key characteristic of crop system is the implementation of a well-known crop simulator that allows the simulation of each yearly and multi-year crops and crop rotations through a unmarried set of parameters. Simulations can final a fraction of a 12 months to masses of years. The model has been evaluated in lots of international locations through comparing model estimates to facts accrued in field experiments.

In this paper [5] writer presenting a practical introduction to state area methods as carried out to unobserved components time series models, also called structural time series models, this paper introduces time series evaluation the usage of state space technique to readers who're neither familiar with time series evaluation, nor with state area methods. The most effective historical past required so that it will apprehend the material offered in the paperis a primary expertise of classical linear regression fashions, of which a quick evaluation is supplied to refresh the reader's know-how. Also, some sections expect familiarity with matrix algebra, but, these sections can be skipped without dropping the float of the exposition. The paper offers a step by step technique to the evaluation of the salient functions in time collection which include the fashion, seasonal, and abnormal components. Realistic troubles which includes forecasting and lacking values are handled in a few detail.

In this paper [8] recent studies by using agriculture researchers in pakistan have proven that attempts of crop yield maximization thru pro-pesticide kingdom regulations have brought about a dangerously high pesticide usage. Those research have pronounced a terrible correlation among pesticide utilization and crop yield in pakistan. Hence excessive use (or abuse) of insecticides is harming the farmers with detrimental financial, environmental and social influences. On this work we've shown that how information mining integrated agricultural facts such as pest scouting, pesticide usage and meteorological recordings is beneficial for optimization (and reduction) of pesticide usage. The facts used on this paintings has in no way been applied in this manner ever before. We've got accomplished unsupervised clustering of this facts through recursive noise removal (rnr) heuristic of abdullah and brobst (2003). These clusters monitor thrilling patterns of farmer practices together with pesticide usage dynamics and for this reason help become aware of the motives for this pesticide abuse.

In this paper [9] author advise a mechanism of performing the mapping from nominal to numeric values (in reality rating) based at the transmittance in addition to the statistical houses of the plant. Spectral analysis (the use of chemical method) is a tedious and time ingesting system, therefore difficult to copy, each and each time, for type of (numerically) unclassified cotton sorts. So a helping statistical method is proposed based totally on linear regression curve becoming the use of normalized nominal attributes. In the end a rank is assigned to the range primarily based on its r2 value and slope of the plot. This rank accordingly turns into the numeric equal of the nominal alphanumeric call of the range being considered. The most complicated problem of this approach is the column ordering used at the same time as generating the regression plots. Wide variety of orderings and column picks were tested based totally on leaf traits, plant characteristics and many others. But, the choice and ordering of columns primarily based on botanical function gave the excellent effects..

On this paper[11] author aimed to assess these new statistics mining strategies and apply them to a soil technology database to set up if meaningful relationships may be located. A big statistics set of soil database is extracted from the department of soil sciences and agricultural chemistry, s v agricultural university, tirupati, the database incorporates measurements of soil profile statistics from various places of chandragiri mandal, chittoor district. The research establishes whether soils are classified the use of diverse statistics mining strategies. In addition, contrast become made among naive bayes classification and analyse the most effective approach.

**CHAPTER 2**

**GENERAL ASPECTS**

**2.1 INTRODUCTION TO MACHINE LEARNING**

machine learning is the subsection of artificial intelligence(AI). the aim of machine emarning is to know the architecture of the information and fitting the information into models which can be known and used by people.

“A computer program is said to learn from experience E with some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.” -Tom M. Mitchell

Consider playing checkers.

E = the experience of playing many games of checkers

T = the task of playing checkers.

P = the probability that the program will win the next game.

Examples of Machine Learning

There are many examples of machine learning. Here are a few examples of classification problems where the goal is to categorize objects into a fixed set of categories.

Face detection: Identify faces in images (or indicate if a face is present).

Email filtering: Classify emails into spam and not-spam.

Medical diagnosis: Diagnose a patient as a sufferer or non-sufferer of some disease.

Weather prediction: Predict, for instance, whether or not it will rain tomorrow.

**Need of Machine Learning**

AI is a field which is raised out of Artificial Intelligence Applying AI, we needed to assemble better and keen machines. In any case, with the exception of multiple of insignificant errands, for example, finding the most brief way between point An and B, we were unfit to program progressively unpredictable and always developing challenges. There was an acknowledgment that the good way to have the capacity to accomplish this assignment was to give machine a chance to gain from itself. This sounds like a kid gaining from its self. So AI was created as another ability for PCs. Also, presently AI is available in such a significant number of portions of innovation, that we don't understand it while utilizing I.

Discovering designs in information on planet earth is conceivable just for human cerebrums. The information being extremely huge, the time taken to process is expanded, and this is the place Machine Learning comes energetically, to help individuals with huge information in least time.

In the event that enormous information and distributed computing are picking up significance for their commitments, AI as innovation breaks down those huge lumps of information, facilitating the errand of information researchers in a computerized procedure and increasing equivalent significance and acknowledgment.

The procedures we use for information digging have surrounded for a long time, however they were not compelling as they didn't have the aggressive capacity to run the calculations. On the off chance that you run profound learning with access to better information, the yield we get will prompt emotional leaps forward which is AI

**Kinds of Machine Learning**

There are three kinds of Machine Learning Algorithms.

a. Supervised Learning

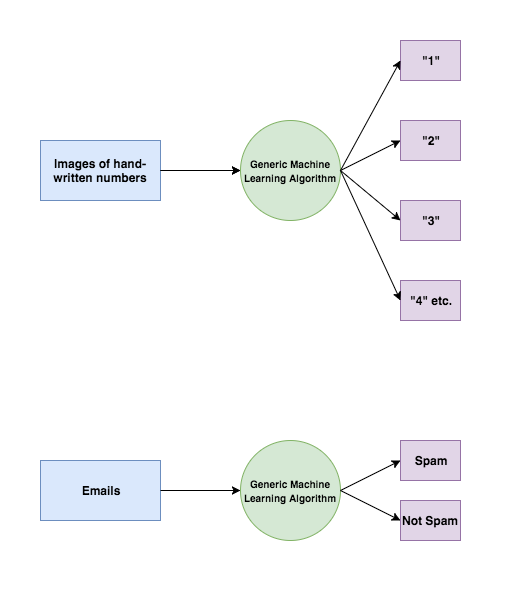
b. Unsupervised Learning

c. Reinforcement Learning

A.supervised learning.

dominant part of down to earth AI utilizes managed learning. In regulated learning, the framework attempts to gain from the past models that are given. (Then again, in unsupervised learning, the framework endeavors to discover the examples legitimately from the model given.)

Talking scientifically, regulated learning is the place you have both info factors (x) and yield variables(Y) and can utilize a calculation to get the mapping capacity from the contribution to.the yield

.

Managed learning issues can be additionally isolated into two sections, to be specific characterization, and relapse.

Characterization: An arrangement issue is the point at which the yield variable is a classification or a gathering, for example, "dark" or "white" or "spam" and "no spam".

Relapse: A relapse issue is the point at which the yield variable is a genuine esteem, for example, "Rupees" or "stature."

**Unsupervised Learning**

In this, the algorithms are left to themselves to discover interesting structures in the data.

Mathematically, unsupervised learning is when you are having have input data (X) and no related output variables.

This is called unsupervised learning because unlike supervised learning above, there are no given exact answers and machine automatically finds the answers. Unsupervised learning problems can be sectioned into association and clustering problems.

Association: association is a kind of problem where you identify the rules which explain the heavy portions of your data as “people that buy x also tends to y.

Clustering: clustering is a king of issue where you identify the inherent grouping in the data, like rouping customers by purchasing behaviour.

**Reinforcement Learning**

A computer program will be communicating with a dynamic environment in which it must perform a particular goal by making use of this algorithm, machines are trained as that it has the capacity of taking exact decisions. It works this way: the machine is exposed to an environment where it continuously trains itself using trial and error method.

**2.2 APPLICATIONS OF MACHINE LEARNING IN CROP YIELD PREDICTION**

We have an idea of the [tending concept of t farming](https://medium.com/sciforce/smart-farming-or-the-future-of-agriculture-359f0089df69) that makes agriculture which is very efficient and effective by making use of high-precision algorithms. The mechanism that drives it is Machine Learning - the scientific field that gives machines which has the capacity to learn apart from being strictly programmed..

Machine learning Is growing throughout the world passionately. It begins with a seed being planted in the soil  from the soil preparation, seeds breeding and water feed measurement   and it ends when robots pick up the harvest determining the ripeness by utilizing the computer vision.

Let’s discover how agriculture can benefit from Machine Learning at every stage:

**Species Recognition**

While the traditional human approach for plant classification would be to compare color and shape of leaves, machine learning can provide more accurate and faster results analyzing the leaf vein morphology which carries more information about the leaf properties.

**Soil management**

Our favourite, this application is so logical and yet so unexpected, because mostly you read about harvest prediction or ambient conditions management at later stages.

Species selection is a complex process of searching for specific genes that determine the effectiveness of water and nutrients use, adaptation to climate change, disease resistance, as well as nutrients content or a better taste. Machine learning, in a specific deep learning algorithms, take decades of field data to analyze crops performance in sevaral climates and new characteristics developed in the process. by depending on this information they can easily form a probability model which will predict what are the genes will be contributing a beneficial trait to plants.

For specialists involved in agriculture, soil is a heterogeneous natural resource, with complex processes and vague mechanisms. Its temperature alone can give insights into the climate change effects on the regional yield. Machine learning algorithms study evaporation processes, soil moisture and temperature to know the ecosystems and the impact in agriculture.

**Water Management**

Water management in agriculture impacts hydrological, climatological, and agronomical balance. So far, the most developed ML-based applications are connected with estimation of daily, weekly, or monthly evapotranspiration allowing for a more effective use of irrigation systems and prediction of daily dew point temperature, which helps identify expected weather phenomena and estimate evapotranspiration and evaporation.

**Yield Prediction**

Yield prediction is one of the most important and popular topics in precision agriculture as it defines yield mapping and estimation, matching of crop supply with demand, and crop management. major approaches have gone far beyond simple prediction based on the historical data, but incorporate computer vision technologies to provide data on the go and comprehensive multidimensional analysis of crops, weather, and economic conditions to make the most of the yield for farmers and population.

**Crop Quality**

The accurate detection and classification of crop quality characteristics can increase product price and reduce waste. In comparison with the human experts, machines can make use of seemingly meaningless data and interconnections to reveal new qualities playing role in the overall quality of the crops and to detect them.

**Disease Detection**

Both in open-air and greenhouse status, largely used practice in pest and disease control is to uniquely spray pesticides on the cropping area.. To be effective, this approach requires significant amounts of pesticides which results in a high financial and significant environmental cost. ML is used as a part of the general precision agriculture management, where agro-chemicals input is targeted in terms of time, place and affected plants.

**Weed Detection**

Apart from diseases, weeds are the major issues to crop production. The complex problem in weeds fighting is that they are difficult to detect and discriminate from crops. Computer vision and ML algorithms can improve detection and discrimination of weeds at low cost and with no environmental issues and side effects. In future, these technologies will drive robots that will destroy weeds, minimizing the requirement for which is used to destroy unwanted plants.

**Livestock Production**

Same as crop management, machine learning provides exact prediction and analyzation of farming parameters to minimize the financial efficiency of livestock production systems, like wise cattle and eggs production. For example, weight predicting systems can easily analyse the future weights 150 days prior to the slaughter day, allowing farmers to modify diets and conditions.

**Animal Welfare**

In present-day setting, the livestock is increasingly treated not just as food containers, but as animals who can be unhappy and exhausted of their life at a farm. Animals behaviour classifiers can connect their chewing signals to the need in diet changes and by their movement patterns, including standing, moving, feeding, and drinking, they can tell the amount of stress the animal is exposed to and predict its susceptibility to diseases, weight gain and production.

**Farmer’s Little Helper**

This is an application that can be called a bonus: imagine a farmer sitting late at night and trying to figure out the next steps in management of his crops. Whether he could sell more now to a local producer or head to a regional fair? He needs someone to talk through the various options to take a final decision. To help him, companies are now working on development specialized chatbots that would be able to converse with farmers and provide them with valuable facts and analytics. Farmers’ chatbots are expected to be even smarter than consumer-oriented Alexa and similar helpers, since they would be able not only to give figures, but analyze them and consult farmers on tough matters.

**CHAPTER 3**

**REQUIREMENT ANALYSIS AND SPECIFICATIONS**

**3.1 Functional requirements**

A functional requirement demonstrates the function of software application or its component and is also explained as group of input, behaviour and outputs. The proposed system is achieved by dividing the data into chunks and the APs applying chunk scheduling algorithm. Applying carrier selection algorithm by APs leads to carry&forward transmission of the data chunks to the destination.

**3.2 NON-FUNCTIONAL REQUIREMENTS**

**3.2.1 EFFICIENCY**

Efficiency is improved by increased network capacity which leads to efficient bandwidth usage resulting in complete information downloading using chunk scheduling algorithms.

**3.2.2 PERFORMANCE**

The way the infrastructure is deployed will have high focus on the efficiency of the cooperative download framework. Not only the position, but also the number of the APs can impact the performance of the vehicular cooperative download.

**3.3 Hardware Requirements**

The hardware requirements performs as the keyrole for a contract for the implementation of the system and must be complete and consistent specification of the total system. They are used by software engineers as the initial point for the system design. It shows what the system does and not how it should be implemented.

**3.4 SOFTWARE REQUIREMENTS**

* Operating system : Windows XP.
* Coding Language : PYTHON 2

**CHAPTER 4**

**SYSTEM ANALYSIS AND DESIGN**

Design Engineering deals with the various UML [Unified Modeling language] diagrams for the implementation of project. Design is a meaningful engineering representation of a thing that is to be built. Software design is a process through which the requirements are translated into representation of the software. Design is the place where quality is rendered in software engineering. Design is the means to accurately translate customer requirements into *finand* the software development process. The UML uses mostly graphical notations to express the design of software projects.

**4.1 USE CASE DIAGRAM**

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

USER

**SEQUENCE DIAGRAM**

### A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

### 

**ACTIVITY DIAGRAM**

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

Start

user

preprocessing

missing values

classification models

SVM, KNN

REGRESSION

EVALUATION

**4.2 Existing System**

Agriculture is the major industrial sectors in India and the country’s economy is highly based up on it for rural supportability. Because of some elements as climate changes, unpredicted rainfall, low water level, use of pesticides in a high amount etc. The level of agriculture in India is reduced. To know the level of production we performed descriptive analytics on the agriculture data.

**4.3 Proposed System**

In our proposed framework the harvest yield order will perform to arrange based on yield efficiency and class marks will be low, mid, and high. What's more, scope of profitability will be characterized and relapse will be performed to get the genuine harvest yield assessed cost. This is the thought process to build up this framework. In light of harvest climate considers, crop yield gauge models are set up for evaluating yield much before genuine reap of the yields. By utilization of exact measurable models utilizing connection and relapse system crops yield are figure on an operational reason for the nation

**4.4 Module Description**

1. **preprocessing and feature extraction**

In this we will find missing values of the dataset.

1. **classification**

In this classification technique we are using svm(support vector machine), knn(k nearest neighbour) ,lssvm (least square support vector machine) . these plays very important role in the classification.

1. **regression**

In this regression we use svm we are using svm(support vector machine), knn(K nearest neighbour), lssvm(least square support vector machine) .

,

**4.evaluation**

In the final evaluation method we are using R-squared metrix and confusion matrix here.

**CHAPTER 5**

**IMPLEMENTATION**

**5.1 Dataset (description:Particular Study Area(Telangana), About Crop)**

In this phase, we collect data from various sources and prepare datasets. And these dataset are used for analytics (descriptive and diagnostic). There are several online abstracts sources such as Data.gov.in and indiastat.org. We will use annual abstracts about a crop for at least ten years period. These dataset usually accept behavior of anarchic time series. Combined the primary and necessary abstracts (collect at least 10 years data) - for sugarcane. Originally data of agriculture (Sugarcane specific) and Rainfall data. humidity, potassium, sodium, phosphorous abundance appropriate for sugarcane crop, minimum and best temperature for crop, RH ethics morning and evening), Clay parameter(depth of soil, PH, comestible agreeable acclimated for crop see table below), acclimate forecast, date of sawing, diseases attacked & pesticides/fertilizers acclimated for sugarcane crop, crop in tones/acres,

**5.2 Each algorithms working and its steps(with flowchart and its Description)**

“Support Vector Machine” (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) which can be used for both classification or regression challenges. However,  it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_1.png)

Support Vectors are simply the co-ordinates of individual observation. Support Vector Machine is a frontier which best segregates the two classes (hyper-plane/ line).

You can look at [support vector machines](https://www.analyticsvidhya.com/blog/2014/10/support-vector-machine-simplified/?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) and a few examples of its working here.

 How does it work?

Above, we got accustomed to the process of segregating the two classes with a hyper-plane. Now the burning question is “How can we identify the right hyper-plane?”. Don’t worry, it’s not as hard as you think!

Let’s understand:

* **Identify the right hyper-plane (Scenario-1):**Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.  
  
* You need to remember a thumb rule to identify the right hyper-plane: “Select the hyper-plane which segregates the two classes better”. In this scenario, hyper-plane “B” has excellently performed this job.
* **Identify the right hyper-plane (Scenario-2):**Here, we have three hyper-planes (A, B and C) and all are segregating the classes well. Now, How can we identify the right hyper-plane?



Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**. Let’s look at the below snapshot:[[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_4.png)](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_4.png)Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

* **Identify the right hyper-plane (Scenario-3):**Hint:Use the rules as discussed in previous section to identify the right hyper-plane

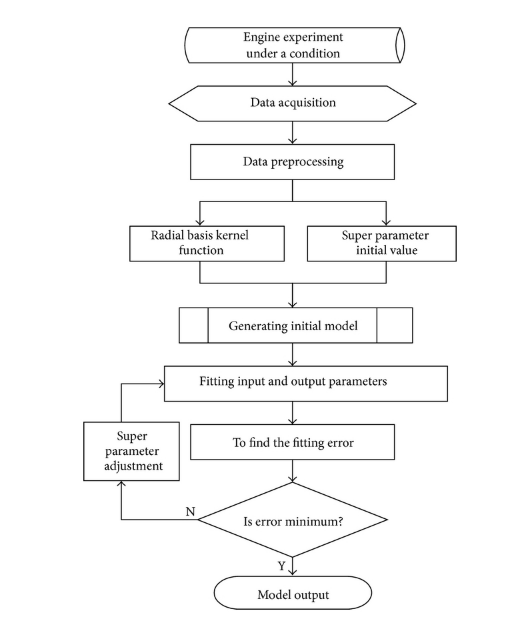
**[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_5.png)**

Some of you may have selected the hyper-plane **B**as it has higher margin compared to **A.**But, here is the catch, SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is **A.**

* **Can we classify two classes (Scenario-4)?:**Below, I am unable to segregate the two classes using a straight line, as one of star lies in the territory of other(circle) class as an outlier.  **[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_61.png)**
* As I have already mentioned, one star at other end is like an outlier for star class. SVM has a feature to ignore outliers and find the hyper-plane that has maximum margin. Hence, we can say, SVM is robust to outliers.  
  **[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_71.png)**
* **Find the hyper-plane to segregate to classes (Scenario-5):**In the scenario below, we can’t have linear hyper-plane between the two classes, so how does SVM classify these two classes? Till now, we have only looked at the linear hyper-plane.**[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_8.png)**
* SVM can solve this problem. Easily! It solves this problem by introducing additional feature. Here, we will add a new feature z=x^2+y^2. Now, let’s plot the data points on axis x and z:  
  [[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_9.png)](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_9.png)In above plot, points to consider are:
  + All values for z would be positive always because z is the squared sum of both x and y
  + In the original plot, red circles appear close to the origin of x and y axes, leading to lower value of z and star relatively away from the origin result to higher value of z.

In SVM, it is easy to have a linear hyper-plane between these two classes. But, another burning question which arises is, should we need to add this feature manually to have a hyper-plane. No, SVM has a technique called the [**kernel**](https://en.wikipedia.org/wiki/Kernel_method)**trick**. These are functions which takes low dimensional input space and transform it to a higher dimensional space i.e. it converts not separable problem to separable problem, these functions are called kernels. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then find out the process to separate the data based on the labels or outputs you’ve defined.

When we look at the hyper-plane in original input space it looks like a circle:  
[](https://www.analyticsvidhya.com/wp-content/uploads/2015/10/SVM_10.png)



svm algorithm construction

**KNN ALGORITHM**

Knn is a non-parametric supervised learning technique in which we try to classify the data point to a given category with the help of training set. In simple words, it captures information of all training cases and classifies new cases based on a similarity.

Predictions are made for a new instance (x) by searching through the entire training set for the K most similar cases (neighbors) and summarizing the output variable for those K cases. In classification this is the mode (or most common) class value.

**How KNN algorithm works**  
  
Suppose we have height, weight and T-shirt size of some customers and we need to predict the T-shirt size of a new customer given only height and weight information we have. Data including height, weight and T-shirt size information is shown below -

|  |  |  |
| --- | --- | --- |
| **Height (in cms)** | **Weight (in kgs)** | **T Shirt Size** |
| 158 | 58 | M |
| 158 | 59 | M |
| 158 | 63 | M |
| 160 | 59 | M |
| 160 | 60 | M |
| 163 | 60 | M |
| 163 | 61 | M |
| 160 | 64 | L |
| 163 | 64 | L |
| 165 | 61 | L |
| 165 | 62 | L |
| 165 | 65 | L |
| 168 | 62 | L |
| 168 | 63 | L |
| 168 | 66 | L |
| 170 | 63 | L |
| 170 | 64 | L |
| 170 | 68 | L |

**Step 1 : Calculate Similarity based on distance function**  
  
There are many distance functions but Euclidean is the most commonly used measure. It is mainly used when data is continuous. Manhattan distance is also very common for continuous variables.

|  |
| --- |
| <https://2.bp.blogspot.com/-VSgHTe470Po/WjhZimqmDgI/AAAAAAAAGiA/j60crf319gM5QVkqMxyCkIPNA78sQkj-ACLcBGAs/s1600/distance.PNG> |
| Distance Functions |

The idea to use distance measure is to find the distance (similarity) between new sample and training cases and then finds the k-closest customers to new customer in terms of height and weight.

**New customer named 'Monica' has height 161cm and weight 61kg.**  
  
Euclidean distance between first observation and new observation (monica) is as follows -

*=SQRT((161-158)^2+(61-58)^2)*

Similarly, we will calculate distance of all the training cases with new case and calculates the rank in terms of distance. The smallest distance value will be ranked 1 and considered as nearest neighbor.  
  
**Step 2 : Find K-Nearest Neighbors**  
  
**Let k be 5.**Then the algorithm searches for the 5 customers closest to Monica, i.e. most similar to Monica in terms of attributes, and see what categories those 5 customers were in. If 4 of them had ‘Medium T shirt sizes’ and 1 had ‘Large T shirt size’ then your best guess for Monica is ‘Medium T shirt. See the calculation shown in the snapshot below -

|  |
| --- |
| <https://2.bp.blogspot.com/-46qXZZx6OKM/WjfXtoYZ8cI/AAAAAAAAGhw/73azNjuB5AoJcSDZ4SSxWVt7n-UDzxO1QCEwYBhgL/s1600/knn.PNG> |
| Calculate KNN manually |

In the graph below, binary dependent variable (T-shirt size) is displayed in blue and orange color. 'Medium T-shirt size' is in blue color and 'Large T-shirt size' in orange color. New customer information is exhibited in yellow circle. Four blue highlighted data points and one orange highlighted data point are close to yellow circle. so the prediction for the new case is blue highlighted data point which is Medium T-shirt size.

|  |
| --- |
| [https://1.bp.blogspot.com/-KfZqdgwJAHg/WjhfBziLKCI/AAAAAAAAGiU/9gUJHU8ztoMjMXQaAvdtkR-OZdzRrNKfgCLcBGAs/s1600/knn%2Bchart.PNG](https://1.bp.blogspot.com/-KfZqdgwJAHg/WjhfBziLKCI/AAAAAAAAGiU/9gUJHU8ztoMjMXQaAvdtkR-OZdzRrNKfgCLcBGAs/s1600/knn+chart.PNG) |
| KNN: Visual Representation |
| K nearest neighbour flow chart |

**CHAPTER 6**

**RESULT ANALYSIS AND DISCUSSION**

**6.1 Screenshots**

**CHAPTER 7**

**CONCLUSION AND FUTURE WORKS**

The report includes the idea to implement the concept of descriptive analytics in agriculture domain. This work provides the information about the application of data analytics on sugarcane crop datasets. There are three dataset named as Soil dataset, Rainfall dataset, Yield dataset. These datasets consists of many parameters which is useful to know the status of crops and classifying the information into several classes by performing supervised training on the dataset which is collected from agriculture domain. This system has the capability to perform both the classification as well as regression. In the classification step the data is classified into three classes (low, mid, and high), whereas in regression step the actual cost of yield production is analyzed. here we are making the use of three major algorithms of supervised learning such as KNN, SVM and LS-SVM to train and build a model.

This research work can be enhancing to the next level. We can build a recommender system of agriculture production and distribution for farmer. By which farmers can make decision in which season which crop should sow so that they can get more benefit. This system is work for structured dataset. In future we can implement data independent system also. It means format of data whatever, our system should work with same efficiency.

**REFERENCES**

[1] J. Ramirez-Villegas and A. Challinor,“Assessing relevant climate data for agricultural applications,” Agricultural Forest Meteorology, 2012, vol. 161(3), pp. 26–45.

[2] C. O. Stockle., S. A. Martin and G. S. Campbell, “CropSyst, a cropping systems simulation model: water/nitrogen budgets and crop yield,” Agricultural Systems, 1994, vol. 46(3), pp. 335-359.

[3]. X. K. Chen and C.H. Yang, “Characteristic of agricultural complex giant system and national grain output prediction,” System Engineering Theory and Practice, 2002, vol. 6(6), pp. 120-125.

[4]. J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,”Operating Systems Design & Implementation, 1989, vol. 51(1), pp. 147–152.

[5]. J. Durbin, “Introduction to state space time series analysis,” State Space & Unobserved Component Models, 2004, pp. 3-25.

[6]. Wu X, Kumar V, Quilan JR, Ghosh J, Yang Q, Motoda H, McLanchlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D, Top 10 algorithms in data mining. KnowlInfSyst14 : 1-37, 2008.

[7]. Abdullah, A., Brobst, S., M.Umer M. 2004. "The case for an agri data ware house: Enabling analytical exploration of integrated agricultural data".Proc. of IASTED International Conference on Databases and Applications. Austria. Feb

[8]. Abdullah, A., Brobst, S, Pervaiz.I.,Umer M.,A.Nisar. 2004. "Learning dynamics of pesticide abuse through data mining". Proc. of Australian Workshop on Data Mining and Web Intelligence, New Zealand, January.

[9]. Abdullah, A., Bulbul.R., Tahir Mehmood. 2005. "Mapping nominal values to numbers by data mining spectral properties of leaves". Proc. of 3 rd International Symposiumm on Intelligent Information Technology in Agriculture. Beijing, China. Oct, 2005

[10]. Georg Ruß, Rudolf Kruse, Martin Schneider, and Peter Wagner. Estimation of neural network parameters for wheat yield prediction. In Max Bramer, editor, Artificial Intelligence in Theory and Practice II, volume 276 of IFIP International Federation for Information Processing, pages 109–118. Springer, July 2008

[11]. Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils P.Bhargavi, Dr.S.Jyothi, IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009 117

[12]. A. Mucherino, A. Urtubia, Consistent Biclustering andApplications to Agriculture, IbaI Conference Proceedings, Proceedings of the Industrial Conference on Data Mining (ICDM10), Workshop “Data Mining in Agriculture” (DMA10), Berlin, Germany,105-113, 2010.

[13]. Tripathi S, Srinivas VV, Nanjundiah RS Downscaling of precipitation for climate change scenarios: a Support Vector Machine approach. J Hydrol ss330:621–640, 2006

[14]. Fagerlund S Bird species recognition using Support Vector Machines. EURASIP J Adv Signal Processing, Article ID 38637, p 8, 2007.

[15]. Yue Jin Hai, Song Kai, 2010. "IBLE Algorithm in agricultural disease diagnosis". In third International Conference on Intelligent Networks and Intelligent Systems held at Shenyang, Liaoning China during November 01- November 2003