Crop Yield Prediction using Mining and Machine Learning Techniques

Submitted in partial fulfilment of the requirements for the degree of

Bachelor of Technology

in

BCA

by

KAMESH R (17BCA0091)

JAGAN T (17BCA0035)

MOHAMMED ARHAM RAYAN J (17BCA0068)

Under the guidance of

Dr. A. Anny Leema

SITE

VIT, Vellore.

May, 2020

DECLARATION

I hereby declare that the thesis entitled "Crop Yield Prediction using Mining and Machine Learning Techniques" submitted by me, for the award of the degree of *Bachelor of Technology in BCA* to VIT is a record of bonafide work carried out by me under the supervision of Dr. A. Anny Leema.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place : Vellore

Date: 25.05.2020

R. K

KAMESH R

Jagan. T

JAGAN T

Mohammed A

Mohammed Arhaam Rayyan J

Signature of the Candidates

CERTIFICATE

This is to certify that the thesis entitled “Crop Yield Prediction using Mining and Machine learning Techniques” submitted by Kamesh R, 17BCA0091, Jagan T (17BCA0035), Mohammed arham rayyan J (17BCA0068) SITE school, VIT, for the award of the degree of *Bachelor of Technology in BCA*, is a record of bonafide work carried out by him / her under my supervision during the period, 01. 12. 2018 to 30.04.2019, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfils the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place : Vellore

Date :20.05.2020

Signature of the Guide

Internal Examiner ***External Examiner***

Head of the Department

Programme

ACKNOWLEDGEMENTS

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many faculties and individuals. I would like to extend my sincere thanks to all of them.

I would like to express my gratitude to Dr. G. Viswanathan (Chancellor), Dr. Sekar Viswanathan (VP), Mr. Sankar Viswanathan (VP), Dr. Anand A. Samuel (Vice-Chancellor), Dr. S. Narayanan (Pro-Vice Chancellor), and Dr. Balakrushna Tripathy (Dean), School of Information Technology and Engineering, for providing with an environment to work in and for his inspiration during the tenure of the course.

In jubilant mood I express ingeniously my whole-hearted thanks to Prof. Mala Serene (Associate Professor and HOD), all teaching staff and members working as limbs of our university for their not self-centered enthusiasm coupled with timely encouragements showered on me with zeal, which prompted the acquirement of the requisite knowledge to finalize my course study successfully. I would like to thank my parents for their support.

I am highly indebted to Dr. A. Anny Leema for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project with the perfection and the completeness.
I would like to express my gratitude and sincereness towards my team mates for their kind co-operation and encouragement which help me in completion, the smooth work flow and the researches of this project.

I would like to express my special gratitude and thanks to Faculty members for giving me such attention and time and help me complete with the standard level and better opinions and comments & work to be done for the betterment of the project.
My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

R. K

KAMESH R

Jagan. T

JAGAN T

mohammed A

Mohammed Arhaam Rayyan J

ABSTRACT

The world is exceedingly reliant on the internet. Nowadays, farming is rapidly decreasing in our country among younger generations. An essential issue for agricultural planning intention is the accurate yield estimation for the numerous crops involved in the planning. Data mining techniques are necessary approach for accomplishing practical and effective solutions for this problem. Environmental conditions, variability in soil, input levels, combinations and commodity prices have made it all the more relevant for farmers to use information and get help to make critical farming decisions. The main objective of this paper is collecting agricultural data which can be stored and analyzed for useful crop yield forecasting. To predict the crop yield with the help of data mining technique, advanced methods can be introduced to predict crop yield and it also helps the farmer to choose the most suitable crop, thereby improving the value and gain of the farming area.

EXECUTIVE SUMMARY

This project involved the design and construction of a Crop yield prediction model - an excessively complex system of mining and machine learning techniques modeled to complete a relatively tedious task. The primary aim of the task was to develop and code such a model which functioned to successfully Predict the yield of a particular crop for a year around an acre of land with higher accuracy and yield.

Following evolution of these designs, some modules were constructed and examined both separately and in compounding with other modules. Early paradigms of these modules included the use of algorithms and console on datasets.

The process is been done through the collection of the data and the pre processing, it has been made to be scaled in order to bring all the set of values to the same range. Followed by the Multiple linear regression which includes two predicates and one predictor yield. Then DBSCAN is performed where the data points have been scattered and then clustered based on the density of it. Finally the paper is concluded with the accuracy rate of multiple linear regression and the clustering conciseness of the DBSCAN.

CONTENTS

Page No.

.

List of Figures

List of Abbreviations

MLR	Multiple Linear Regression
DBSCAN	Density Based Spatial Clustering And Noise Application
OLS	Ordinary Least Squares
CAPM	Capital Asset Pricing Model
PSM	Project Schedule Management

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION ABOUT AREA

Agriculture is the main support and the major sector of the Indian economy. The production of agronomy is far too low. As the demand for food is growing exponentially, the researchers, analysts, farmers, scientists, specialists and government try to place further effort and strategies to increase agricultural production to accommodate the needs.

Data Mining is the process of extract helpful and significant information from huge sets of data. Data Mining in agriculture field is a comparatively novel research field. Yield prediction is a very important agricultural problem. Any farmer is interested in knowing how much yield he is concerning to be expecting. In the earlier period, yield prediction was performing by considering farmer's experience on particular field and crop. In any of Data Mining actions the training data is to be collected from past data and the gathered data is used in terms of training which has to be exploited to study how to categorize future yield predictions. Crop models and decision tools are more and more used in agricultural field to improve production efficiency. The combination of higher technology and agriculture to improve the production of crop yield is becoming more interesting newly. Due to the rapid development of new high technology, crop models and predictive tools might be predictable to become a crucial element of agriculture.

Crop yield is a combined bio-socio-system comprised of complex interaction among the soil, the air, the water, and the crops grown in it, where a comprehensive model is necessary which are possible only through classical engineering expertise. As define by the Food and Agriculture of the United Nations, crop forecasting is the art of predict crop yields and production before the harvest in fact takes place, typically a couple of months in advance. Based on atmospheric and agronomy data, several indices are gained which are held to be pertinent variables in determining crop production, for illustrate crop water satisfaction, surplus and excess moisture, average soil moisture, etc. Linear Regression model characterizes the mathematical relationships intrinsic to the data set from previous experiments. This method can produce results under various situations assuming extensive information used to expand and test the model. Though, in agricultural data, information is rather sparse and incomplete. Because of this limitation, the linear regression approach is the common approach for predicting yield across large area.

It is a fascinating method of estimating crop and the quantity of yield in an advance way before the harvest essentially takes place. Foreseeing the crop yield can be tremendously valuable for farmers. It tells them the indication of when and how to harvest crops certainly.

The contribution of agricultural professionals and re- searchers in the prediction of crop yield leads to issues like nascent farmers about natural occurrence's, the negation of personal awareness and exhaustion etc. such problems can be altered by using crop yield techniques.

Data mining technique aim at finding those patterns or information in the data that are together valuable and interesting to the farmer. The most frequent specific problem to take place is yield prediction.

1.2 MOTIVATION

A crop prediction is a widespread trouble that takes place. During the upcoming season, a farmer had curiosity and patience knowing how much outcome he is about to expect. In the earlier times, this yield forecast become a matter of fact relied on.

Farmer's long-term see for specific yield, crops and weather conditions. Farmer directly goes for yield forecast rather than bearing on crop prediction with the existing system. Unless the correct crop is forecasted how the yield will be better and moreover with existing systems pesticides, overall environmental and atmospherical parameter relation to crop is not taken into consideration. Considering and solacing the agricultural production at a more quickly pace is one of the needy situation for agricultural development. Any crop's production see the path either by concern of domain or improvement in yield or both. In India, the prospect of extending the district under any crop does not exist except by re-establishing to rise cropping intensity or crop replacement. So, variations in crop productivity persist to problem the area and create rigorous distress. So, there is need to attempt the best technique for crop prediction in order to get over existing problem.

1.3 LITERATURE SURVEY

Kusum Lata et al uses Classification algorithms based on rule based machine learning K-Nearest Neighbour etc.Genetic Algorithm (Fitness Function), Back Propagation Algorithm with Feed Forward Neural Network and Artificial Neural Network [1].

Ananthi et al approaches Linear Regression, K-NN and LR-Proposed algorithm to have an accurate crop yield prediction that gives us the accuracy rate for each of these applied algorithms [2].

Tavva Sasikanth et al utilises ID3 (Iterative Dichotomiser) Algorithm and CART (Classification and Regression Trees) algorithms. A small change in the data can cause a large change in the structure of the decision tree causing instability. For a Decision tree sometimes calculation can go far more complex compared to other algorithms [3].

Ramesh A et al uses K-Means, K-Nearest Neighbor(KNN), Artificial Neural Networks(ANN) and Support Vector Machines(SVM). Unexplained behaviour of the network: This is the most important problem of ANN. When ANN produces a probing solution, it does not give a clue as to why and how. This reduces trust in the network [4].

Surya A. Venkaiah makes use of ID3 (Iterative Dichotomiser) algorithm and CART (Classification and Regression Trees) algorithm. Decision Tree algorithm is inadequate for applying regression and predicting continuous values [5].
Sajidullah et al advances WEKA tool, J48, Bayes Net, KStar and Random Tree, Machine learning. Weka tool it can only handle small datasets. Whenever a set is bigger than a few megabytes an Out Of Memory error occurs. The object of this thesis is to alter **Weka** in such a way that it can handle ”all” datasets, up until a few gigabytes [6].

Chavan A et al encourages Prediction Algorithms, Regression Analysis(non linear regression, support vector regression). A strong sensitivity to outliers, need to be set key parameters correctly to achieve the best classification results for any given problem [7].

Dr. C. Yamini et al promotes K-Nearest Neighbour classifier, Decision tree and Bayesian network. The main disadvantage of the KNN algorithm is that it is a lazy learner, i.e. it does not learn anything from the training data and simply uses the training data itself for classification. They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree [8].

Latha G et al employs DBSCAN and Linear Regression algorithms to compute the yield prediction. But it is tend to be known from the research is it is sensitive to outliers [9].

Prof. Anil Kumar Mishra et al approaches K-Means and Multiple Linear Regression. The major drawback that we have come through here is that it is difficult to predict K-Value and sensitive to outliers [10].

1.4 BACKGROUND

CLASSIFICATION

Classification is an information mining capacity that relegates things in an accumulation to target classifications or classes. The objective of characterization is to precisely anticipate the objective class for each case in the information.

APPLICATIONS IN AGRICULTURE

There are several applications in the field of agriculture. Some of them are listed below.

CROP SELECTION AND CROP YIELD PREDICTION

To increase the crop yield, selection of the particular crop that will be sown plays an important role. It reckon on various factors like the variety of soil and its composition, climate, geography of the area, crop production, market prices etc. Techniques like Artificial neural networks, K-nearest neighbours and Decision Trees have carved a niche for themselves in the context of crop selection which is based on various factors. Crop excerption based on the effect of natural disasters or calamities like famines has been made based on machine learning (Washington Okori, 2011). The use of artificial neural networks to choose the crops based on soil and climate has been shown by researchers (Obua, 2011). A plant nutrient management system has been proposed based on machine learning methods to meet the needs of soil, maintain its fertility levels, and hence improve the crop yield (Shivnath Ghosh, 2014). A crop excerption method called CSM has been suggested which helps out in crop selection based on its yield prediction and other agents (Kumar, 2009).

WEATHER FORECASTING

Indian farming mainly depends on seasonal rains for irrigation. Therefore, a precise prediction of weather can deduce the enormous toil faced by farmers in India including crop choice, watering and harvesting. As the farmers have poor access to the Internet as a result of digital-divide, they have to rely on the little information available regarding weather reports. Up-to-date as well as accurate weather information is still not available as the weather changes dynamically over time. Researchers have been working on improving the accuracy of weather predictions by using a variety of algorithms. Artificial Neural networks have been adopted extensively for this purpose. Likewise, weather prediction based on machine learning technique called Support Vector Machines had been proposed (M.Shashi, 2009). These algorithms have shown better results over the conventional algorithms.

SMART IRRIGATION SYSTEM

Farming sector consumes a vast assign of water in India. The levels of ground water are dropping down day-by-day and global warming has resulted in climate changes. The river water for irrigation is a big issue of dispute among many states in India. To combat the scarcity of water, many companies have come up with sensor based technology for smart farming which uses sensors to monitor the water level, nutrient content, weather forecast reports and soil temperature. EDYN Garden sensor is another example (Gupta, 2016). However, the high cost of such devices deters the small land owners and farmers in India to use them. These smart devices are being designed on the principles of machine learning. The nutrient content of soil can also be recorded using the sensors and hence used for supplying fertilizers to the soil using smart irrigation systems. This will also reduce the labour cost in the fields, which is a huge crisis being faced by the Indian farmers these days.

1.5 PROBLEM STATEMENT

GENERAL PROBLEM ANALYSIS

Early prediction of crop yield is essential for planning and adopting various policy determinations. Many countries use the conventional technique of data collection for crop monitoring and yield prediction based on ground based visits and reports. These methods are subjective, very costly and time consuming. The coarse problem in existing crop yield prediction methods are listed below,

- The most significant problem of existing crop yield prediction method is precise and time devouring problem.

- In surviving time series crop yield prediction method does not respond to variations that take place for cycles and seasonal effects.

- Needs extensive information to develop and test the model and also available information in agriculture is sparse and incomplete in existing simulation model.

- Limited studies have been built in crop yield prediction using living decision tree proficiency.

- Prediction error value also crucial problem in crop yield forecast or estimation methods.

- These are the main drawbacks of various existing works, which motivate us to do this research on crop yield prediction.

PROBLEM STATEMENTS IN ALGORITHMS

With the evolution of the algorithms in data mining, the prediction process is changing in terms of speed with the use of data mining techniques and new algorithms. But the existing systems lack in terms of speed and Efficiency due to implementation of techniques with high time complexity and implementation of primitive algorithms. Even if a particular website tries its best to grab any customers, there is a huge competition from the market. Website owners are thus, unable to understand the user’s personal needs and as a result are failing to meet their demands. And it finds difficult to update the dataset often due to this there are problem arise finding the accurate result. The existing system also lacks predicting the variety of crops.

CHAPTER 2

PROJECT DESCRIPTION AND GOALS

2.1 EXISTING SYSTEM

With the evolution of the algorithms in data mining, the prediction process is changing in terms of speed with the use of data mining techniques and new algorithms. But the existing systems lack in terms of speed and Efficiency due to implementation of techniques with high time complexity and implementation of primitive algorithms. Even if a particular website tries its best to grab any customers, there is a huge competition from the market. Website owners are thus, unable to understand the user's personal needs and as a result are failing to meet their demands. And it finds difficult to update the dataset often due to this there are problems that arise finding the accurate result. The existing system also lacks predicting the variety of crops.

2.2 OBJECTIVE

The main aim of agricultural production is to obtain maximum crop yield. Initial discovery and management of complexities like crop yield can assist amplify return yield and assuring profits. If regional weather patterns are influenced, large scale weather events can have a substantial effect on crop production. Crop managers can use predictions to minimize damage in critical conditions. Furthermore, these forecasts could be used to make full use of the crop forecast if the potential for favourable conditions of growth exists. In this part we are attending to follow through the methodologies which have been used by professors, scientists and data specialist for predicting crop yield.

The main objective is to fetch the output of Multiple Linear Regression and the DBSCAN pattern to verify whether the results in crop prediction are accurate and constellate. This paper uses crop yield prediction techniques to forecast the appropriate crop by identifying different crop seasons and geographic condition parameters. With MLR, we could be able to estimate the accuracy of the production when compared to the real time values. This paper demonstrates the ability of the multiple linear regression algorithm to monitor and predict crop yields in remote areas and cities.

2.3 SCOPE OF THE PROJECT

SCOPE AND IMPORTANCE OF AGRICULTURE IN INDIA AND TAMILNADU

With a 16% share to the gross domestic product (GDP), agriculture still provides living affirm to about two-thirds of country's population.

The sector grants employment to 58% of country's work force and is the single largest private sphere occupation.

Agriculture bills for about 15% of the total export incomes and provides raw material to a large number of Factories (textiles, silk, sugar, rice, flour mills, milk products).

Rural areas are the biggest markets for low-priced and middle-priced consumer products, including consumer durables and rural domestic preservations are a significant source of resource mobilization.

The farming sector acts as a surround in holding food security and in the process, national security as well been included.

The allied and confederative sectors like horticulture, animal husbandry, dairy and fisheries, have an essential role developing the complete economic conditions and health and nutrition of the rural masses.

To assert the ecological balance, there is indigence for sustainable and balanced growth of agriculture and allied sectors.

Agriculture's eyes and minds are comforted by active changes from brown (bare soil) to green (growing crop) to golden (mature crop) and bumper reaps.

Plateauing of agricultural productivity in irrigated areas and in some causes the diminishing trend warrants attention of scientists.

Agriculture helps to the growth of the community consisting of dissimilar castes and communities to a better social, cultural, political and economical life. Agriculture holds a biological equilibrium in nature. Satisfactory agricultural production brings peace, prosperity, harmony, health and wealth to each and every individuals of a nation by driving away distrust, discord and anarchy.

INDIAN AGRICULTURE AND ECONOMY

Indian Agriculture is one of the most substantial contributors to the Indian economy. Agriculture is the one and only way of living for nearly 60% of the employed class in India. The agriculture sector of India has covered around 43% of India's geographical area. Agriculture is still the only largest subscriber to India's GDP (16%) even after a descent in the same in the agriculture share of India. Agriculture also plays an important role in the development of socio-economic sphere in India. In the older times, India was largely relied upon food imports, but the successive story of the agriculture sector of Indian economy has done it self-sufficing in grain production. The country also has significant reserves for the same. India trusts heavily on the agriculture sector, especially on the food production unit after the 1960 crisis in food sphere. Since then, India has commit a lot of attempt to be self-sufficient in the food production and this endeavour of India has guided to the Green Revolution.

CHAPTER 3

TECHNICAL SPECIFICATION

3.1 REQUIREMENTS SPECIFICATION

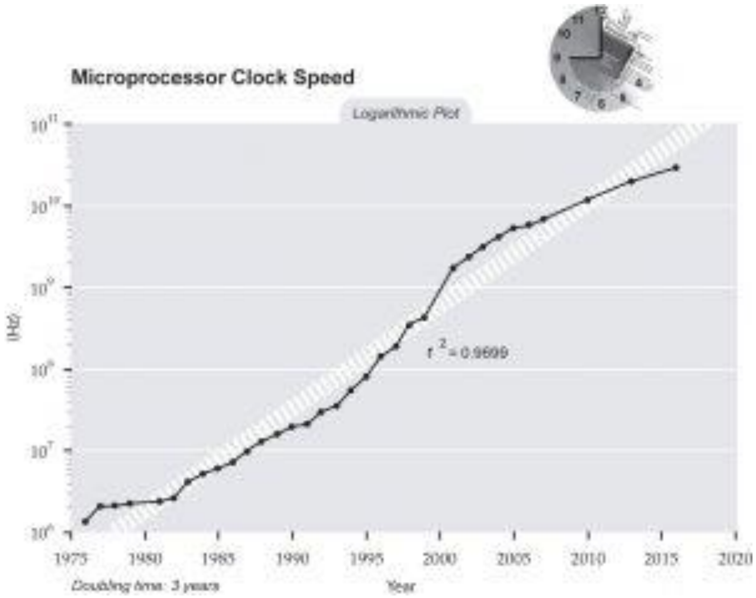
OVERVIEW AND INTRODUCTION

Chips are integral to your computer because they're the brain: processors deal with all of the instructions that other hardware and software throw around. When we're talking specifically about data mining and machine learning, the processor has the role of executing the logic in a given algorithm. If we're performing Gradient Descent to optimize the cost function, the processing unit is directing and executing it. That means running the basic mathematical computations (matrix multiplication) that drive the algorithm.

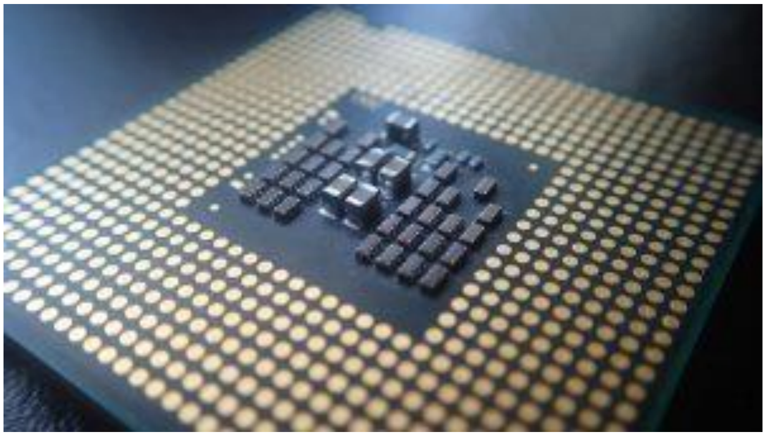
The CPU (Central Processing Unit)

The OG processing unit is the CPU, which was first developed by Intel in the early 1970s (pictured below).

As time went on, CPUs grew in speed and capability. For some context, according to Computer Hope, "the first microprocessor was the Intel 4004 that was released on 15 November 1971, and had 2,300 transistors and performed 60,000 operations per second. The Intel Pentium processor has 3,300,000 transistors and performs around 188,000,000 instructions per second."



Most processors were designed with one core (one CPU), which meant it could only perform one operation at once. IBM released the first dual-core processor in 2001, which was able to "focus" on two tasks at once. Since then, more and more CPUs have been crammed into microprocessors: some modern supercomputers can have more than 40.



Even with recent advances, the fact remains that most computers only have a few cores at most. CPUs are designed for complex computations: they're very good at rapidly parsing through a detailed and intertwined set of commands. And for most of the tasks a computer needs to do, like swimming aimlessly through a sea of Chrome tabs, that's exactly what you want. But with machine learning, things can get a bit different.

Data Mining Poses a New Type of Challenge for Processing

The strength of the CPU is executing a few complex operations very efficiently, and data mining and machine learning presents the opposite challenge. Most of the computation in the training process is matrix multiplication, which is a simple but broad task—the calculations are very small and easy, but there are *a ton* of them. Effectively, the CPU is often overpowered but understaffed.

Advances in data storage are some of the major drivers of the explosion of machine learning over the past decade, and they've also compounded this problem. Today we're training algorithms on more data than ever before, which means more and more small calculations that max out our CPUs.

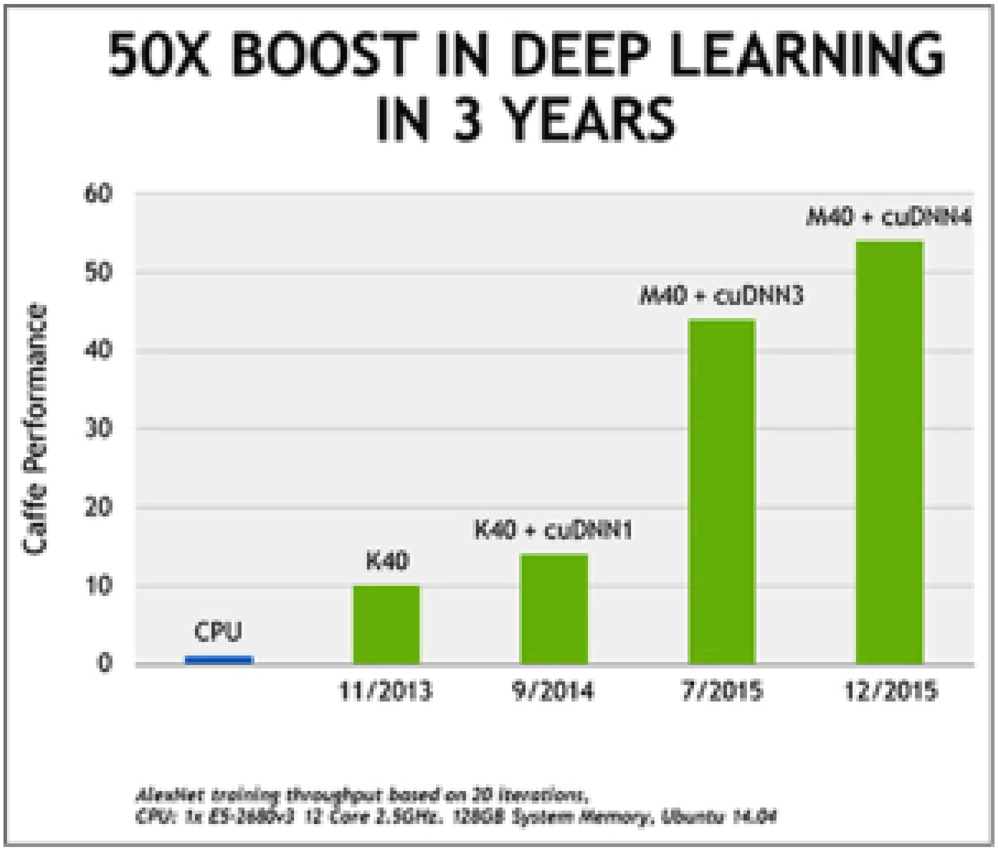
A far better-optimized chip for machine learning is actually another major processor that's mass manufactured—something that only has the core complexity to do basic operations, but it can do them at scale all at the same time. Luckily, that chip has already been sitting in our computers for years, and it's called a GPU.

GPUs Have Risen to the Occasion

GPUs, or Graphics Processing Units, have been around in gaming applications since the early 1970s. The late 80s saw GPUs being added into consumer computers, and by 2018 they're absolutely standard. What makes a GPU unique is how it handles commands—it's the exact opposite of a CPU.

GPUs utilize parallel architecture: while a CPU is excellent at handling one set of very complex instructions, a GPU is very good at handling many sets of very simple instructions.

A few years ago, groups in the machine learning community started to realize that these architectural features—that GPUs are excellent for parallel processing of simple operations—might lend well to using them for algorithms. Over time, GPUs started to show massive improvements over CPUs for training models, often in the 10x ballpark for speed. And the stock price of Nvidia, the most well-known manufacturer of these kinds of chips, shows as much (from Google Finance):



Nvidia isn’t the only manufacturer of GPUs, but it’s certainly the default one. There are other vendors like AMD, but the software that’s used to integrate with the chips is far behind Nvidia software. Cuda, Nvidia’s platform, is a usable platform for machine learning applications.

The degree to which GPUs have become popular is hard to overstate. They’re in high demand right now, for the original video game applications as well as for machine learning (and even cryptocurrency mining), and prices have been skyrocketing. The price for a standard Nvidia GPU manufactured last year is now **higher than it was when released.** Algorithmia is the only major vendor that supports serverless execution (FaaS) on GPUs.

ASICs: The Black Horse

ASICs, or Application Specific Integrated Circuits, are the next level of chip design—it’s a processor designed *specifically for one type of task*. The chip is built to be very good at executing a specific function or type of function.

An awesome and relevant example is cryptocurrency mining, which people seem to attack with endless creativity. If you need an introduction to cryptocurrency, head over to Coindesk's Blockchain 101 section. But for our purposes, all you need to know is this: mining these currencies on your computer involves brute-force guessing of a specific number. It's a pretty simple function, but the faster you do it, the higher chance you have at winning; and there's no variation: you just keep guessing until you get it right.

Google has also gotten into the ASICs game, but with a focus on machine learning. It's called a **TPU (Tensor Processing Unit)**, and it's a Google-designed and manufactured chip specifically made for machine learning with Tensorflow, Google's open-source deep learning framework. TPUs are much faster than the best CPUs and GPUs for programming neural nets, Google claims, but there has been some dispute as to how accurate that figure really is. A third-party benchmark was recently released and found that TPUs can be significantly more efficient than comparable GPUs.

As machine learning becomes more and more integrated into all the applications we use on a daily basis, expect more research to be done on how to create chips tailored for these tasks.

How to Access And Use CPUs, GPUs, and TPUs

Moving from the theoretical and architectural, in 2018 it's finally possible to access all of these kinds of chips for your machine learning applications.

CPUs

CPUs make up the bulk of modern public cloud offerings. If you train and run models on normal AWS (Amazon), GCP (Google), or Azure (Microsoft) instances, they'll be using CPUs. Unless you're using specific deep learning frameworks that target GPUs on your machine, running algorithms locally will also use your CPUs.

GPUs

Because of the skyrocketing demand and relatively modest supply, GPUs weren't always that easy to access. Thankfully, the major cloud platforms have bought up enough of them to make it possible today. AWS allows for a GPU-only setup (p3 instance), as well as Google and Microsoft. Algorithmia's **Serverless AI Layer** integrates both GPUs and CPUs, and makes it easy to deploy on either.

ASICs

The only widely available ASIC for machine learning applications is Google’s TPU program. As of recently, TPUs are available to the public for compute.

Type of Processor Best Suits

As with anything, the answer is it depends. Projects and products involving machine learning often have varying priorities, ranging from speed to accuracy to reliability. As a general rule, if you can get your hands on a state-of-the-art GPU, it’s your best bet for fast machine learning.

GPU compute will usually be about 4 times as expensive as CPU compute, so if you’re not getting 4 times improved speed, or if speed is less of a priority than cost, you might want to stick with CPUs. Additionally, your training needs will be different than your implementation needs. A GPU may be necessary for training a large **Neural Net**, but it’s possible that a CPU is more than powerful enough to use the model once built.

TPUs are still experimental, so we’ll need more data before making judgements about tradeoffs.

Difference between CPU and GPU - “*GPUs can handle graphics better because graphics include thousands of tiny calculations that need to be conducted. Instead of sending those tiny equations to the CPU, which could only handle a few at a time, they’re sent to the GPU, which can handle many of them at once.*”

HARDWARE REQUIREMENTS

- Physical server or virtual machine.
- CPU: 2 x 64-bit, 2.8 GHz, 8.00 GT/s CPUs or better. **Verify machine architecture.**
- Memory: minimum RAM size of 32 GB, or 16 GB RAM with 1600 MHz DDR3 installed, for a typical installation with 50 regular users. **Verify memory requirements.**
- Storage: Recommended minimum of 100 GB, or 300 GB if you are planning to mirror both Anaconda Repository, which is approximately 90 GB, and the PyPI repository, which is approximately 100 GB, or at least 1 TB for an air gapped environment. Additional space is recommended if Repository is used to store packages built by your organization. **Verify storage requirements.**
- Internet access to download the files from Anaconda Cloud, or a USB drive containing all of the files you need with alternate instructions for air gapped installations.

HARDWARE VERIFICATION

MACHINE ARCHITECTURE

Repository is built to operate only on 64-bit computers.

To verify that you have a 64-bit or x86_64 computer, in a terminal window, run:

```
arch
```

This command displays what your system is: 32-bit “i686” or 64-bit “x86_64.”

MEMORY REQUIREMENTS

You need a minimum RAM size of 32 GB, or 16 GB RAM with 1600 MHz DDR3.

In a terminal window, run:

```
free -m
```

This command returns the free memory size in MB.

STORAGE REQUIREMENTS

To check your available disk space—hard drive or virtual environment size—use the built-in Linux df utility with the -h parameter for human readable format:

```
df -h
```

SOFTWARE REQUIREMENTS

- RHEL/CentOS 6.5 to 7.4, Ubuntu 12.04+
- Ubuntu users may need to install cURL.
- Client environment may be Windows, macOS or Linux
- MongoDB 2.6 (provided)
- Anaconda Repository license file
- Cron entry to start the repo on reboot
- Linux system accounts
- mongod (RHEL) or mongodb (Ubuntu)
- anaconda-server

SECURITY REQUIREMENTS

- Privileged access OR sudo capabilities

- Open HTTP(S) port
- SELinux policy edit privileges (SELinux does not have to be disabled for Anaconda Repository operation)
- Optional: Ability to make iptables modifications
- Optional: SSL certificate

NETWORK REQUIREMENTS (TCP PORTS)

- Inbound HTTP: TCP 8080, 8443 (Anaconda repository)
- Optional Inbound SSH: TCP 22 (SSH)
- Optional Outbound HTTPS: TCP 443
- repo.anaconda.com
- anaconda.org
- conda.anaconda.org
- binstar-cio-packages-prod.s3.amazonaws.com
- 820451f3d8380952ce65-4cc6343b423784e82fd202bb87cf87cf.ssl.cf1.rackcdn.com
- Optional Outbound SMTP: TCP 25 (if not using AD/LDAP) email notifications
- Optional Outbound LDAP(s): TCP 389/636 for authentication integration

OTHER REQUIREMENTS

- License file provided to you by Anaconda at the time of purchase
- Installation tokens for binstar and anaconda-server channels provided by Anaconda at the time of purchase. Not applicable for air gapped installs.
- Optional: Your Anaconda Cloud (anaconda.org) account credentials. Not applicable for air gapped installs.

OTHER VERSIONS OF THE LINUX ENVIRONMENTS

Please contact us by filing a GitHub issue if you have problems with a version other than Redhat, CentOS or Ubuntu. Prompts may vary slightly depending on your version.

cURL access for Ubuntu users

RedHat and CentOS Linux distributions have cURL pre-installed, but Ubuntu does not.

To verify cURL access, in a terminal window, run:

```
curl --version
```

If cURL is not found, Ubuntu users can use the Advanced Packaging Tool (APT) to get and install cURL:

```
sudo apt-get install curl
```

TIP: If you already have Miniconda or Anaconda installed, in all versions of Linux you can use the conda command:

```
conda install curl
```

MongoDB version 2.4+ installed

MongoDB version 2.4 or higher must installed as root and running. Versions through 3.4 are supported. To check for the existence of MongoDB and its version number, in a terminal window, run:

```
mongod --version
```

If you get a “not found” message or if the MongoDB version is 2.3 or earlier, then install MongoDB 2.4 or higher using the official installation instructions. Remember to install as root with the sudo command.
MongoDB must always be running before Repository can be started.

To start MongoDB:

```
sudo service mongod start
```

To verify that MongoDB is running:

```
mongo --eval 'db.serverStatus().ok'
```

bzip2 is installed

To check for the existence of bzip2 and its version number, in a terminal window, run:

```
bzip2 --version
```

SECURITY VERIFICATION

ROOT ACCESS AND SUDO PREVILEGES

The Repository installation process cannot be completed without root access.

To verify that you have sudo privileges, in a terminal window, run:

```
sudo -v
```

Enter your root password when prompted and press Enter.

If you receive a message like the following, contact your system administrator for root access:

```
Sorry, user [username] may not run sudo on [hostname].
```

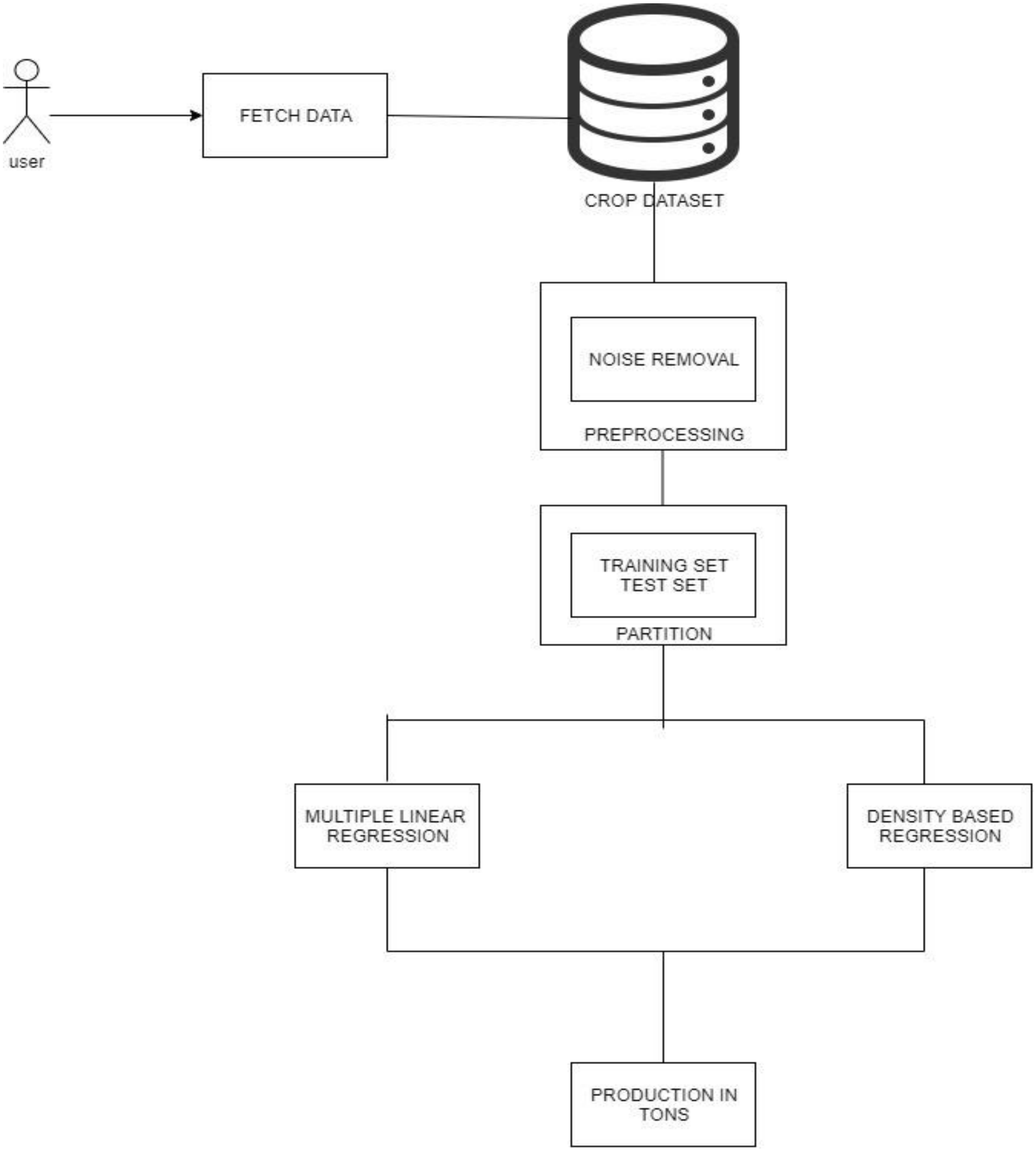
CHAPTER 4

DESIGN APPROACH AND DETAILS

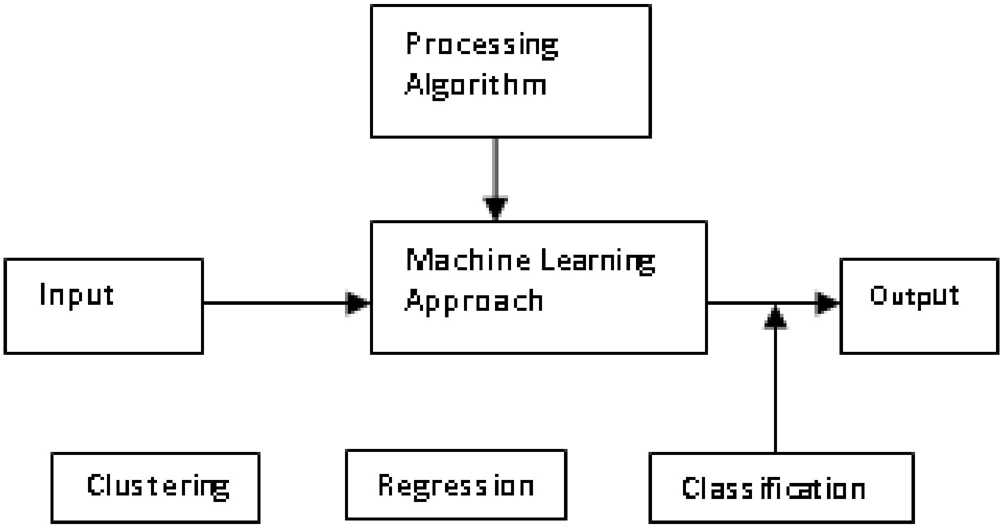
4.1 MATERIALS AND METHODS

SYSTEM MODULES

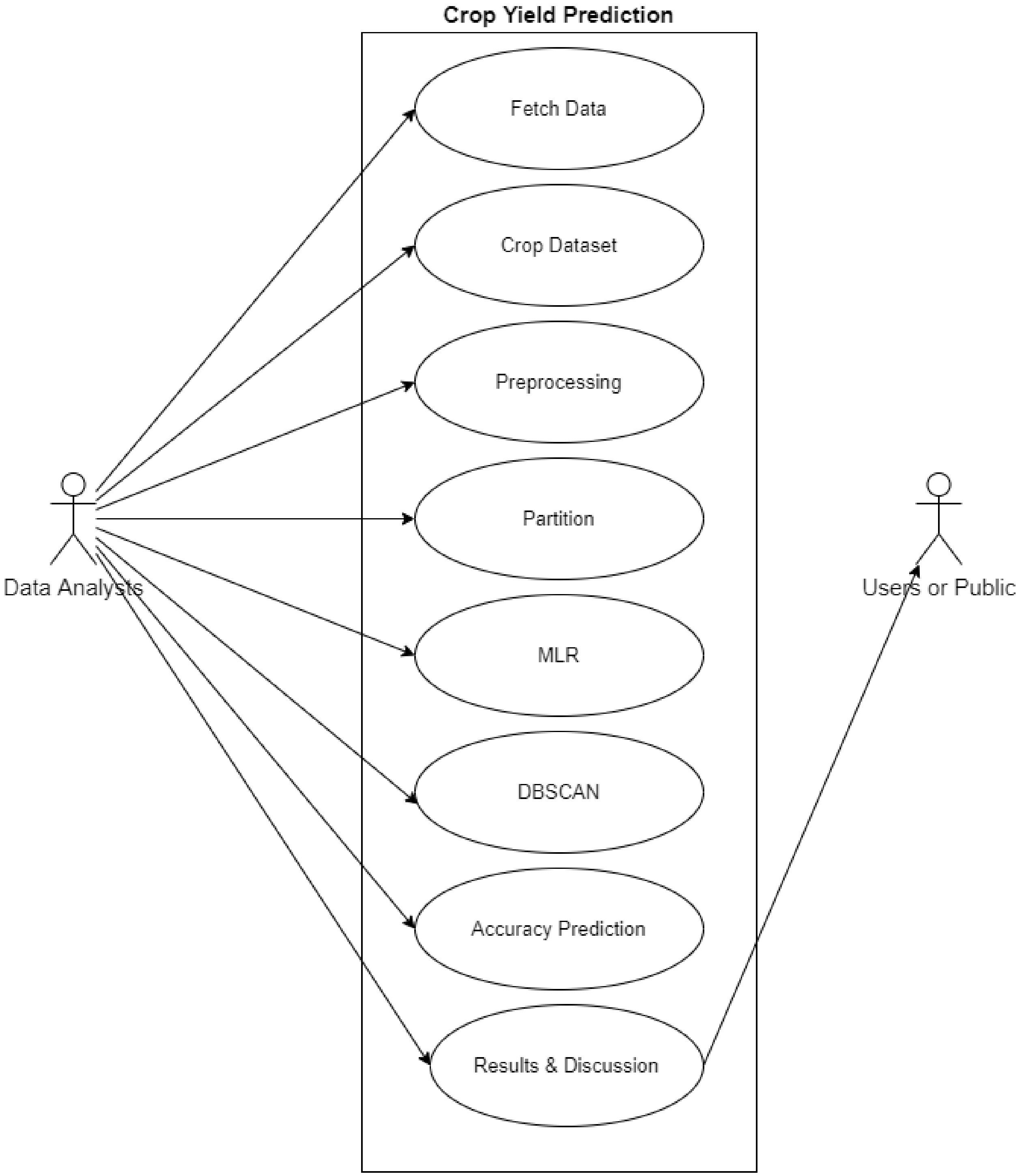
Here the system module is been divided into two major sector. They are the Website and the Mining & Machine Learning Algorithms sectors. The sector involves the real-time access of the users worldwide.



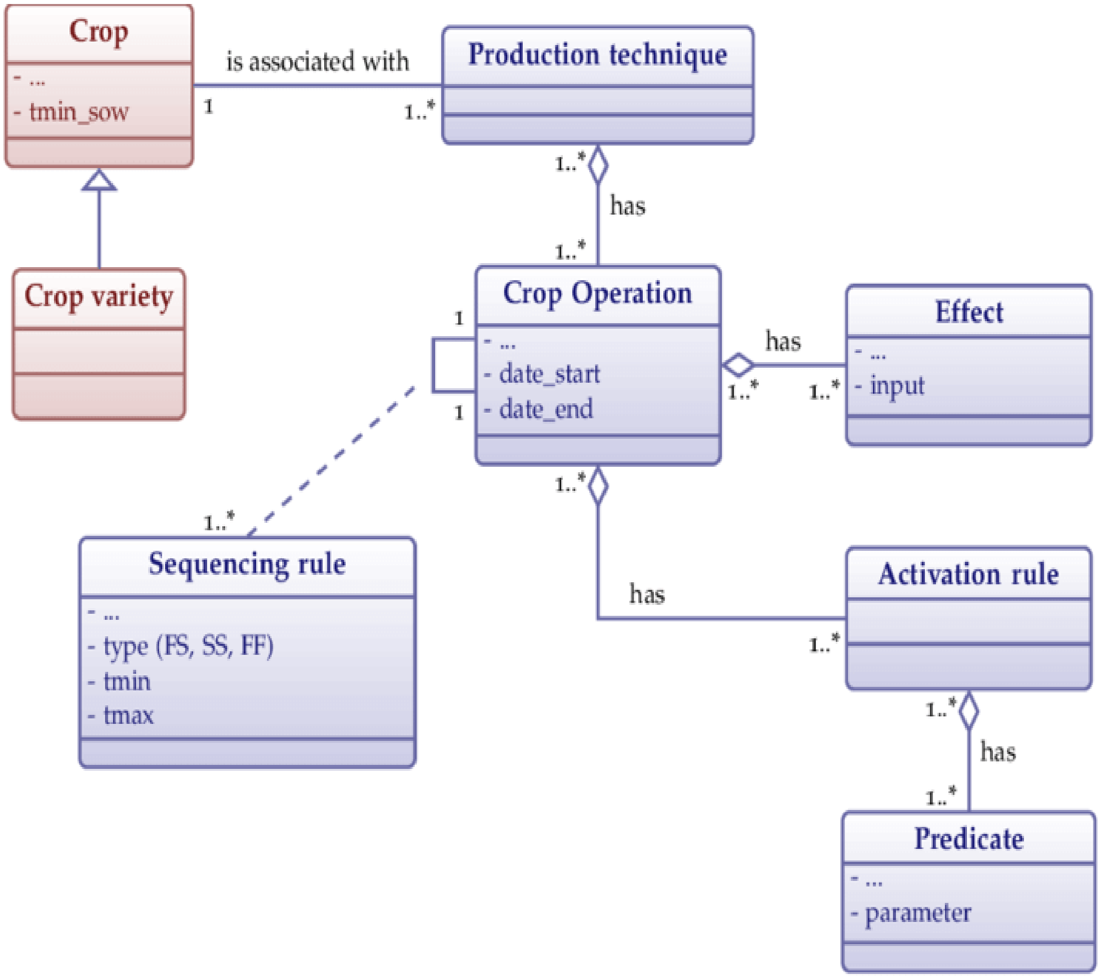
System Architecture Diagram



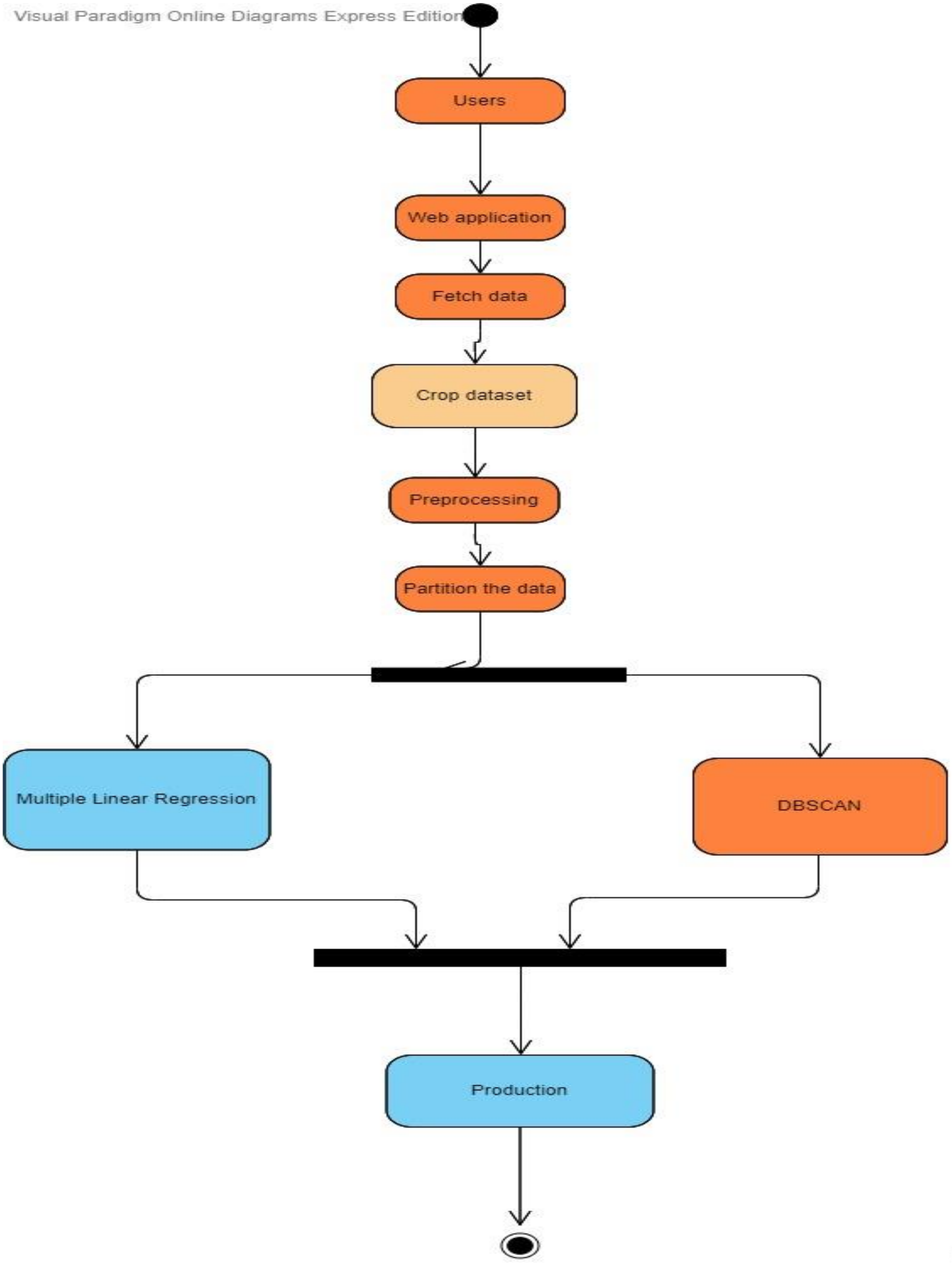
System Work Flow Diagram



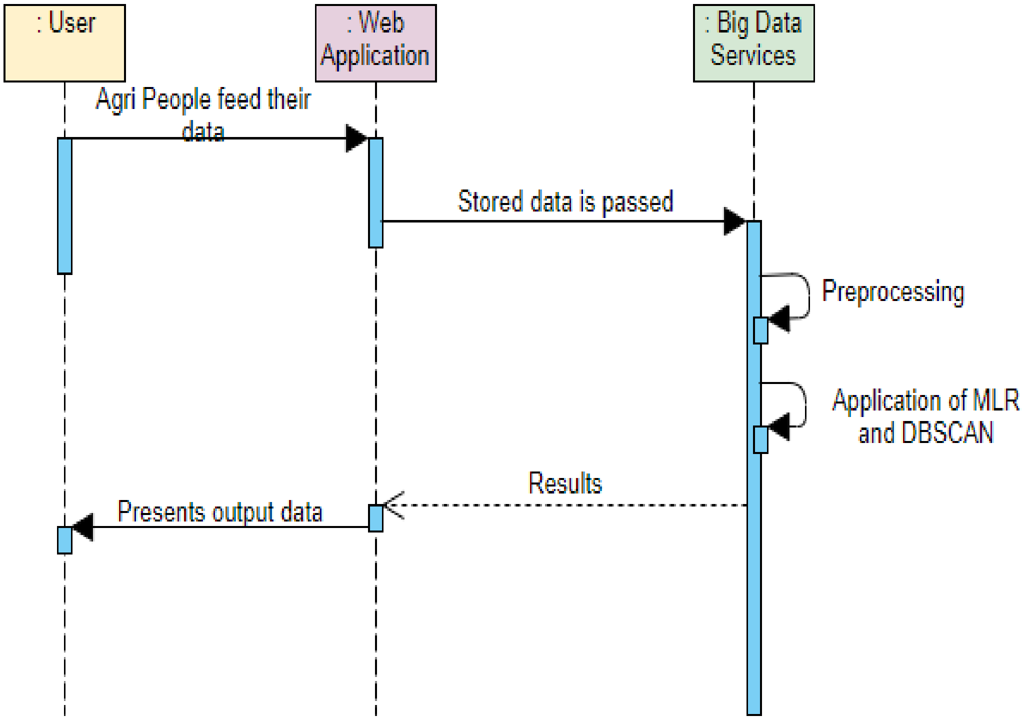
Use Case Diagram



Class Diagram



Activity Diagram



Sequence Diagram

TECHNOLOGY USED

REGRESSION

Regression is a statistical technology that is been used in finance, investing, and other applications that seeks to determine the intensity and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Regression assists investment and financial managers to determine assets and interpret the relationships between variables, such as **commodity prices** and the livestock of businesses that is been dealing in those commodities.

Regression can aid finance and investment professionals as well as professionals in other businesses. Regression can also help out to predict sales for a company or a sector based on weather, previous sales, GDP growth, or other types of cases. The capital asset pricing model (CAPM) is an often-used regression framework in finance for pricing assets and identifying costs of capital.

Types of Regression

The two introductory types of regression are simple linear regression and multiple linear regression, although there are non-linear regression exemplar for more complex data and analysis. Simple linear regression employs one independent variable to describe or predict the value of the dependent variable Y, while multiple linear regression exercises two or more independent variables to forecast the outcome.

The Difference Between Linear and Multiple Regression

Linear (OLS) regression equates the reply of a dependent variable applied a change in some explanatory variable. However, it is rare that a dependent variable is described by only one variable. In this case, an analyst uses multiple regression, which attempts to explain a dependent variable using more than one independent variable. Multiple regressions can be linear and nonlinear.

Multiple regressions are based on the assumption that there is a linear relationship between both the dependent and independent variables. It also assumes no major correlation between the independent variables.

MULTIPLE LINEAR REGRESSION

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical proficiency that employs several explanatory variables to anticipate the result of a response variable. The aim of multiple linear regression (MLR) is to exemplar the linear relationship between the explanatory (independent) variables and reaction (dependent) variable.

In essence, multiple regression is the elongation of ordinary least-squares (OLS) regression that demands more than one explanatory variable.

The Formula for Multiple Linear Regression Is
$yi = \beta 0 + \beta 1xi1 + \beta 2xi2 + ... + \beta pxip + \epsilon$
where, for **$i=n$** observations:
yi=dependent variable
xi=expanatory variables
$\beta 0$=y-intercept (constant term)
βp=slope coefficients for each explanatory variable
ϵ=the model's error term (also known as the residuals)

Explaining Multiple Linear Regression
A simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

The multiple regression model is based on the following assumptions:

- There is a linear relationship between the dependent variables and the independent variables.
- The independent variables are not too highly correlated with each other.
- y_i observations are selected independently and randomly from the population.
- Residuals should be normally distributed with a mean of 0 and variance σ.

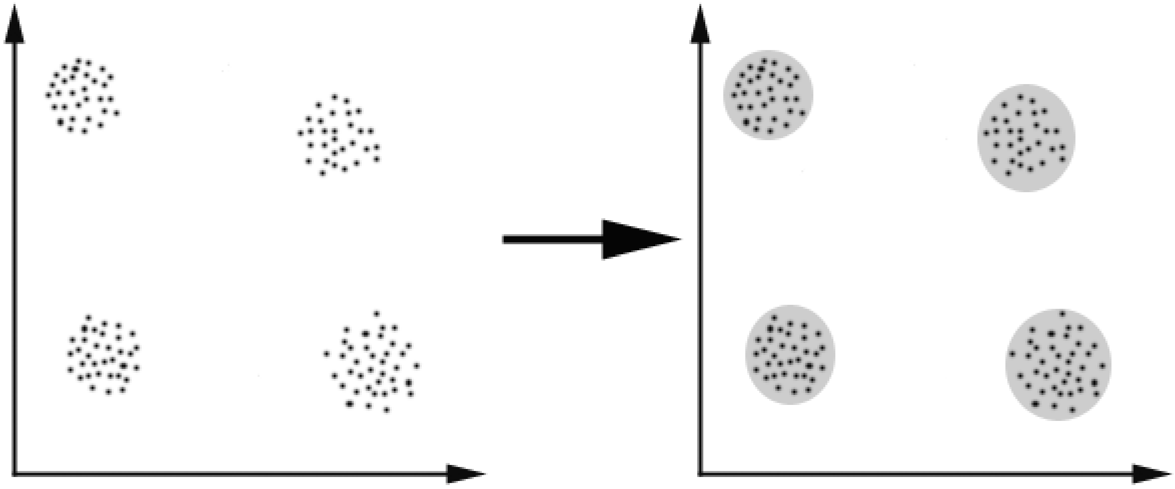
The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the MLR model even though the predictors may not be related to the outcome variable.

R^2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

When interpreting the results of a multiple regression, beta coefficients are valid while holding all other variables constant ("all else equal"). The output from a multiple regression can be displayed horizontally as an equation, or vertically in table form.

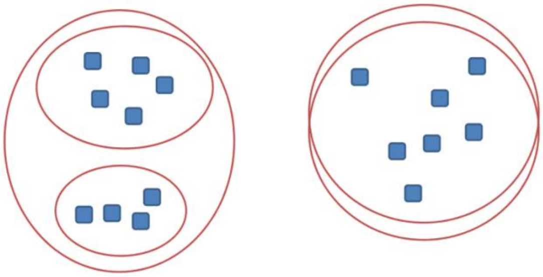
CLUSTERING IN MACHINE LEARNING

Clustering can be consolidated the most significant *unsupervised learning* problem; so, as every other problem of this type, it addresses with detecting a *structure* in a collection of unlabeled data. A loose description of clustering could be "the operation of organizing data points into groups whose members are likely in some way". A *cluster* is therefore a group of objects which are "similar" between them and are "dissimilar" to the objects owing to other clusters.



THE GOALS OF CLUSTERING

The aim of clustering is to ascertain the internal grouping in a set of unlabeled data. But how to consider what accounts for a good clustering? It can be viewed that there is no absolute "best" criterion which would be independent of the concluding aim of the clustering. Consequently, it is the user who should distribute this criterion, in such a way that the outcome of the clustering will suit their essentials.



To find a specific clustering solution, we need to fix the similarity values for the clusters.

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is the most well-defined density-based clustering algorithm, **initially introduced in 1996 by Ester et. al.** Owing to its importance in both theory and practical, this algorithm is one of three algorithms awarded the **Test of Time Award at SIGKDD 2014**.

Unlike K-Means, DBSCAN does not involve the number of clusters as a delimiter. Rather it interprets the number of clusters based on the data, and it can find clusters of arbitrary shape (for comparison, K-Means usually discovers spherical clusters). As I said earlier, the ε-neighborhood is fundamental to DBSCAN to approximate local density, so the algorithm has two parameters:

- *ε*: The radius of our neighborhoods around a data point *p*.
- *minPts*: The minimum number of data points we want in a neighbourhood to define a cluster.

Using these two parameters, DBSCAN categories the data points into three categories:

- *Core Points*: A data point *p* is a *core point* if **Nbhd**(*p*,*ε*) [ε-neighbourhood of *p*] contains at least *minPts* ; |**Nbhd**(*p*,*ε*)| >= *minPts*.
- *Border Points: A data point *q* is a *border point* if **Nbhd**(*q*, *ε*) contains less than *minPts* data points, but *q* is *reachable* from some *core point p*.
- *Outlier*: A data point *o* is an *outlier* if it is neither a core point nor a border point. Essentially, this is the “other” class.

These definitions may seem abstract, so let’s cover what each one means in more detail.

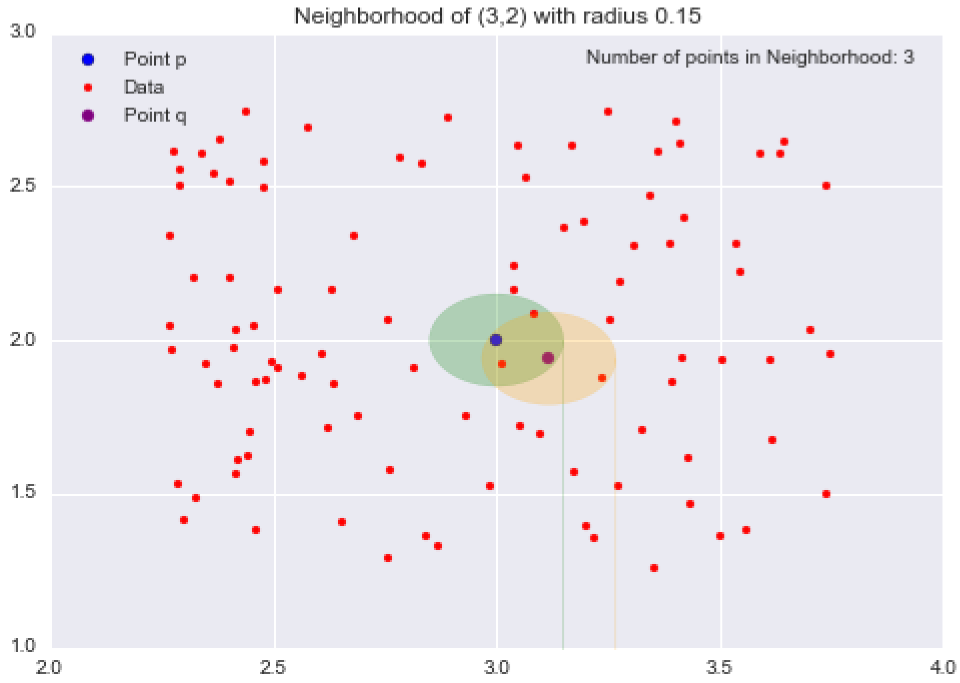
Core Points

Core Points are the creations for our clustered data points are based on the density estimation discussed earlier. We use the same ε to compute the neighborhood for each point, so the volume of all the neighborhoods is the same. However, the number of other points in each neighborhood is what differs. Recall that I said we can think of the number of data points in the neighborhood as its *mass*. The volume of each neighborhood is constant, and the mass of neighborhood is variable, so by assigning a threshold on the minimum measure of mass needed to be *core point*, we are significantly setting a minimum density threshold. Therefore, core points are data points that meet a minimum density requirement. Our clusters are built across our *core points* (hence the *core* part), so by aligning our *minPts* parameter, we can fine-tune how dense our clusters cores must be.

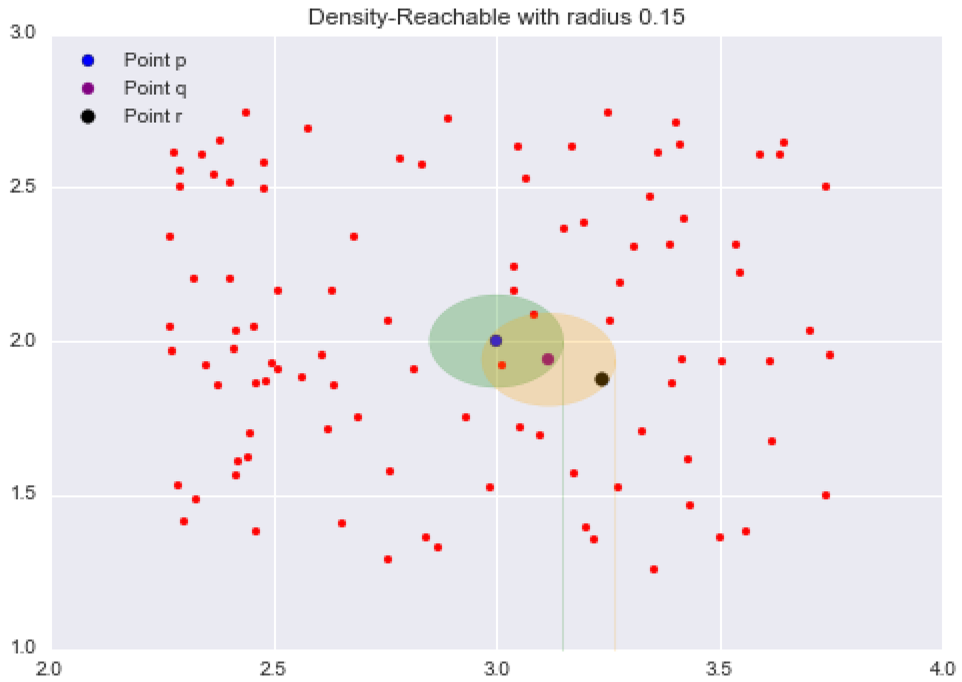
Border Points

Border Points are the points in our clusters that are not core points. In the resolution above for *border points*, I applied the term *density-reachable*. I have not described this term yet, but the concept is simple. To describe this method, let's revisit our neighborhood example with epsilon = 0.15. Consider the point *r* (the black dot) that is outside of the point *p*'s neighborhood.

All the points inside the point *p*'s neighborhood are said to be *directly reachable* from *p*. Now, let's explore the neighborhood of point *q*, a point *directly reachable* from *p*. The yellow circle represents *q*'s neighborhood.



Now while our target point *r* is not our starting point *p*'s neighborhood, it is contained in the point *q*'s neighborhood. This is the idea behind density-reachable: If I can get to the point *r* by jumping from neighborhood to neighborhood, starting at a point *p*, then the point *r* is density-reachable from the point *p*.



As an analogy, we can conceive of *density-reachable* data points as being the “friends of a friend”. If the *directly-reachable* of a *core point p* are its “friends”, then the *density-reachable* points, points in locality of the “friends” of *p*, are the “friends of its friends”. One thing that may not be acquit is *density-reachable* points is not fixed to just two adjacent neighborhood jumps. As long as you can arrive at the point causing “neighborhood jumps”, starting at a *core point p*, that point is *density-reachable* from *p*, so “friends of a friend of a friend … of a friend” are admitted as well.

It is important to keep in mind that this idea of *density-reachable* is dependent on our value of ε. By picking larger values of ε, more points become *density-reachable*, and by choosing smaller values of ε, less points become *density-reachable*.

Outliers

Finally, we head to our “other” class. *Outliers* are data points that are neither *core points* nor are they close enough to a group to be *density-reachable* from a *core point*. *Outliers* are not given to any cluster and, rely on the context, may be treated anomalous points.

Now that I have covered all the preliminaries, we can finally talk about how the algorithm works in practice.

Advantages of DBSCAN:

- Is great at classifying clusters of high density versus clusters of low density within an applied dataset.
- Is great with handling outliers within the dataset.

IMPLEMENTATION

Steps Involved in MLR

- Compute a Multiple Linear Regression Model using Multiple independent variables.
- Compute the residual values.
- Compute the residual sum of Squares and obtain the R Square.
- Implementation in the console and view the difference between the values and estimate the accuracy.

Steps Involved in DBSCAN

- Import the libraries.
- Load the Dataset.
- Install sklearn, matplotlib and pandas packages in the python application (If already installed ignore it).
- Select the particular column values needed for algorithm.
- Replace the missing values with the false or 0.
- Scale the values to a short range. This will keep all the data values to a standard normal distribution.
- Compute dbscan with the use of DBSCAN function.
- Visualize the cluster.

4.2 CODES & STANDARDS

WEBSITE CODING

LOGIN PAGE

```
<?php include('server.php') ?>
<!DOCTYPE html>
<html>
<head>
  <title>Registration system PHP and MySQL</title>
  <link rel="stylesheet" type="text/css" href="style.css">
</head>
<body style="background-color: coral;">
  <center><img src="logo.png" style="margin-top: 30px; margin-bottom: 0px;">
  <div class="header" style="margin-top: 30px;">
    <h2>Login</h2>
  </div>
	</center>
  <form method="post" action="login.php">
	<?php include('errors.php'); ?>
	<div class="input-group">
		<label>Username</label>
		<input type="text" name="username" >
	</div>
	<div class="input-group">
		<label>Password</label>
		<input type="password" name="password">
```

```
        </div>

        <div class="input-group">

                <button type="submit" class="btn" name="login_user">Login</button>

        </div>

        <p>

                Not yet a member? <a href="register.php">Sign up</a>

        </p>

  </form>

</body>

</html>
```

SERVER PAGE

```
<?php

session_start();


// initializing variables

$username = "";

$email    = "";

$errors = array();


// connect to the database

$db = mysqli_connect('localhost', 'root', '', 'registration');


// REGISTER USER

if (isset($_POST['reg_user'])) {

  // receive all input values from the form
```

```
$username = mysqli_real_escape_string($db, $_POST['username']);

$email = mysqli_real_escape_string($db, $_POST['email']);

$password_1 = mysqli_real_escape_string($db, $_POST['password_1']);

$password_2 = mysqli_real_escape_string($db, $_POST['password_2']);


// form validation: ensure that the form is correctly filled ...

// by adding (array_push()) corresponding error unto $errors array

if (empty($username)) { array_push($errors, "Username is required"); }

if (empty($email)) { array_push($errors, "Email is required"); }

if (empty($password_1)) { array_push($errors, "Password is required"); }

if ($password_1 != $password_2) {

        array_push($errors, "The two passwords do not match");

}


// first check the database to make sure

// a user does not already exist with the same username and/or email

$user_check_query = "SELECT * FROM users WHERE username='$username' OR email='$email' LIMIT 1";

$result = mysqli_query($db, $user_check_query);

$user = mysqli_fetch_assoc($result);


if ($user) { // if user exists

  if ($user['username'] === $username) {

    array_push($errors, "Username already exists");

  }
```

```
    if ($user['email'] === $email) {
      array_push($errors, "email already exists");
    }
  }

  // Finally, register user if there are no errors in the form
  if (count($errors) == 0) {
          $password = md5($password_1);//encrypt the password before saving in the database

          $query = "INSERT INTO users (username, email, password)
                              VALUES('$username', '$email', '$password')";
          mysqli_query($db, $query);
          $_SESSION['username'] = $username;
          $_SESSION['success'] = "You are now logged in";
          header('location: index.php');
  }
}

// ...

// LOGIN USER
if (isset($_POST['login_user'])) {
  $username = mysqli_real_escape_string($db, $_POST['username']);
  $password = mysqli_real_escape_string($db, $_POST['password']);
  if (empty($username)) {
    array_push($errors, "Username is required");
  }
  if (empty($password)) {
    array_push($errors, "Password is required");
  }

  if (count($errors) == 0) {
    $password = md5($password);
    $query = "SELECT * FROM users WHERE username='$username' AND password='$password'";
    $results = mysqli_query($db, $query);
    if (mysqli_num_rows($results) == 1) {
      $_SESSION['username'] = $username;
      $_SESSION['success'] = "You are now logged in";
      header('location: index.php');
    }else {
      array_push($errors, "Wrong username/password combination");
    }
  }
}
?>
```

AGRI INFO COLLECT PAGE

```
<!DOCTYPE html>
```

```
<html lang="en">
<head>
<title>AgroArc | AgroData</title>
<meta charset="utf-8">
<link rel="stylesheet" href="css/style.css">
<script src="js/jquery-1.7.1.min.js"></script>
<script src="js/superfish.js"></script>
<script src="js/jquery.easing.1.3.js"></script>
<script src="js/tms-0.4.1.js"></script>
<script src="js/slider.js"></script>
<!--[if lt IE 9]>
<script src="js/html5.js"></script>
<link rel="stylesheet" href="css/ie.css">
<![endif]-->
</head>
<body>
<div class="main-bg">
  <!-- Header -->
  <header>
    <div class="inner">
      <h1 class="logo"><a href="index.html">AgroArc - Agriculture company</a></h1>
      <nav>
        <ul class="sf-menu">
          <li><a href="index.html">home</a></li>
          <li><a href="feed.html">feed</a></li>
          <li> <a href="technology.html">technology</a></li>
          <li class="current"><a href="expert.html">farm expert</a></li>
          <li><a href="contacts.html">contacts</a></li>
        </ul>
      </nav>
      <div class="clear"></div>
    </div>
  </header>
  <!-- Content -->
  <section id="content">
    <div class="container_24">
      <div class="wrapper">
        <div class="grid_24 content-bg">
          <div class="wrapper">
            <article class="grid_15 suffix_1 omega">
              <h2><center>Agricultural Info Form:</center></h2>
              <form action="server.php" id="contact-form" method="post">
                <fieldset>
                  <label class="name"> <span>Full name:</span>
                    <input type="text" name="name">
                  </label>
                  <label class="name"> <span>State Name:</span>
                    <input type="text" name="sname">
                  </label>
                  <label class="name"> <span>District Name:</span>
                    <input type="text" name="dname">
                  </label>
                  <label class="name"> <span>Area:(in acres)</span>
                    <input type="text" name="area">
                  </label>
                  <label class="name"> <span>Season:</span>
                    <select name="season">
                      <option value="kharif">Kharif</option>
                      <option value="rabi">Rabi</option>
                      <option value="zaid">Zaid</option>
                    </select>
                  </label>
                  <label class="name"> <span>Crop Year:</span>
                    <input type="number" min="1990" max="2020" step="1" value="2019"
                  name="year">
                  </label>
                  <label class="name"> <span>Crop Cultivated:</span>
                    <select name="crop">
                      <option value="rice">Rice</option>
                      <option value="wheat">Wheat</option>
                      <option value="Groundnut">Groundnut</option>
                      <option value="Millets">Millets</option>
                      <option value="Pulses">Pulses</option>
                      <option value="Cotton">Cotton</option>
                      <option value="Tea">Tea</option>
```

```
                    <option value="Coffee">Coffee</option>

                    <option value="Maize">Maize</option>

                    <option value="Turmeric">Turmeric</option>

                    <option value="Barley">Barley</option>

                    <option value="Mustard">Mustard</option>

                    <option value="Soyabean">Soyabean</option>

                    <option value="peas">Peas</option>

                    <option value="Red Chillies">Red Chillies</option>

                    <option value="Sugarcane">Sugarcane</option>

                    </select>

                  </label>

                  <label class="name"> <span>Production:(in tonne)</span>

                   <input type="text" name="prod">

                  </label>

                  <div class="btns"> <button name="clear" class="button" style="margin:0 0 0
35px;box-shadow: 0px 0px 0px transparent;border: 0px solid transparent; text-shadow: 0px 0px
0px transparent;">Clear</button> <button class="button" name="submit" style="margin:0 0 0
35px;box-shadow: 0px 0px 0px transparent;border: 0px solid transparent; text-shadow: 0px 0px
0px transparent;">Submit</button> </div>

                 </fieldset>

                </form>

               </article>

              </div>

             </div>

            </div>

           </div>

          </section>
```

```
<!-- Footer -->
<footer>
  <div class="container_24">
    <div class="wrapper">
      <div class="grid_24 footer-bg">
        <div class="hr-border-2"></div>
        <div class="wrapper">
          <div class="grid_7 suffix_1 prefix_1 alpha">
            <div class="copyright"> &copy; 2020 <strong class="footer-logo">AgroArc</strong>
          </div>
          </div>
          <div class="grid_4">
            <h5 class="heading-1">Archives:</h5>
            <ul class="footer-list">
              <li><a href="#">October 2020</a></li>
              <li><a href="#">September 2020</a></li>
              <li><a href="#">August 2020</a></li>
              <li><a href="#">July 2020</a></li>
            </ul>
          </div>
          <div class="grid_4">
            <h5 class="heading-1">Links:</h5>
            <ul class="footer-list">
              <li><a href="#">Documentation</a></li>
              <li><a href="#">Plugins</a></li>
```

```
                <li><a href="#">Suggest Ideas</a></li>
                <li><a href="#">Support Forum</a></li>
              </ul>
            </div>
            <div class="grid_4">
              <h5 class="heading-1">Support:</h5>
              <ul class="footer-list">
                <li><a href="#">Special Proposition</a></li>
                <li><a href="#">Free Phone</a></li>
                <li><a href="#">Solutions</a></li>
              </ul>
            </div>
            <div class="grid_2 suffix_1 omega">
              <ul class="social-list">
                <li><a href="#"><img src="images/social-icon-1.png" alt=""></a></li>
                <li><a href="#"><img src="images/social-icon-2.png" alt=""></a></li>
                <li><a href="#"><img src="images/social-icon-3.png" alt=""></a></li>
              </ul>
            </div>
          </div>
        </div>
      </div>
    </div>
  </footer>
</div>
```

```
</body>
</html>
```

MULTIPLE LINEAR REGRESSION

```
# -*- coding: utf-8 -*-
"""
@author: vitians
"""
# importing the packages for the execution
# Of the plotting and importing the dataset
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
# importing the dataset using the pandas package
df = pd.read_csv('pro.csv')
"""
pairplot plot a pairwise relationships in a dataset. The pairplot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column.
"""
sns.pairplot(df)
# assigning colors to the each of the relations

sns.heatmap(df.corr(),linewidth = 0.2, vmax=1.0,square=True, linecolor='red',annot=True)

features = df.iloc[:,[2,3]].values
```

```
labels = df.iloc[:,-1].values

# partitioning the dataset into two parts
# one is the test set and the other is the training set
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(features,labels,test_size =0.3,random_state=0)
# preprocessing the dataset
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X_train=sc.fit_transform(X_train)
X_test=sc.transform(X_test)
from sklearn.linear_model import LinearRegression
# Regression is done using the train data
regressor = LinearRegression()
regressor.fit(X_train,y_train)

# predicting the y value
y_pred = regressor.predict(X_test)
# printing the coefficient and intercept
print(regressor.coef_)
print(regressor.intercept_)

# printing the test and the train score
# Then comparing the test and the train scores
```

```
print('Train score:',regressor.score(X_train,y_train))

print('Test score:',regressor.score(X_test,y_test))

y_output0 = regressor.intercept_ + regressor.coef_[0]*X_test[0][0] + regressor.coef_[1]*X_test[0][1]

plt.scatter(x=features[:,0],y=features[:,1],color="red",alpha=0.75)

plt.plot(features,color="blue",alpha=0.75)
```

DBSCAN

```
# -*- coding: utf-8 -*-

"""

@author: vitians

"""

# importing the packages

from sklearn.cluster import DBSCAN

from sklearn.preprocessing import StandardScaler

import pandas as pd

import matplotlib.pyplot as plt

# importing the crop dataset

X = pd.read_csv('E:\clg documents\capstone project\coding\MLR\pro.csv', header=[0])

print(X.head())

# Dropping the CUST_ID column from the data

X = X[["Year", "Production " ,"Area ", "Yield"]]


# Handling the missing values

X.fillna(method ='ffill', inplace = False)
```

```
print(X.head())

# preprocessing the dataset to transform

scaler = StandardScaler()

X_scaled = scaler.fit_transform(X)

# cluster the data into five clusters

dbscan = DBSCAN(eps=0.3, min_samples = 2)

clusters = dbscan.fit_predict(X_scaled)

# plot the cluster assignments

plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=clusters, cmap="plasma")

# labelling the x and y axes

plt.xlabel("Feature 0")

plt.ylabel("Feature 1")
```

4.3 CONSTRAINTS, ALTERNATIVES AND TREDEOFFS

IMPACT OF SCOPE AND QUALITY CONSTRAINTS

The development of software projects usually requires frequent changes. This is probably because users think that software changes only require programmers to change the code, and the cost of modification is not large. But this is not the case. Changes in the user's thoughts result in a change in requirements, and the project manager is unable to reject the user's changes for a variety of reasons. This will invisibly affect the progress of the project. If the quality of a sub-project completed is not as expected, for example, there are often loopholes, maintenance difficulties, and so on. This subproject must be rewritten, which wastes a lot of time by wasting human resources and other inputs. On the other hand, it also delays the progress of the project. Therefore, the quality of the subproject affects the progress of the entire project, and the subproject with higher priority will affect the priority to the subproject.

IMPACT OF RESOURCE AND BUDGET CHANGES CONSTRAINTS

The resources here refer to human resources. In some subprojects, there may be insufficient team members, or members of individual sub-projects in the whole project will be called into other sub-projects or one person is responsible for multiple sub-projects at the same time. Another resource refers to information resources. The legal standards of each country are different. The income of citizens in each city is different. The standards of each industry are not uniform. These information resources are rarely provided by customers. Team members collect, if the information resources are not available on time, it will also affect the project's needs analysis, overall design and programming work. In add-on, other resources refer to growth equipment and evolution environment software. These resources will not affect the progress of the entire project in time. The budget can also be said to be a resource. The amount of budget that affects other resources will ultimately affect the overall progress. For example, a high-budget development environment that uses a high-performance environment will speed up the team's completion of the project and vice versa.

CHAPTER 5

SCHEDULE, TASKS AND MILESTONES

The progress is the work plan date table specified for the executed activities and milestones. The project schedule management is also anticipated project time management and time handling. It refers to the progress of the work in each stage and the final completion of the project during the project implementation process. The management of the term is to ensure the necessary management process for the project to be completed on time. project schedule management is one of the important measures to ensure the project is completed on time and rationally arrange the supply of resources and save engineering costs.

Project schedule management is to use the corresponding method to analyze the activities and interrelationships between the projects, estimate the time required for each subproject, and arrange the subprojects and the reasonable time within the time limit specified by the project. Control the start and end time of the subproject. PSM（Project schedule management） can be summarized as the following six main parts: Activity definition: The parent project is divided into multiple subprojects, and each subproject must have deliverables. Detailed definition of the specific activities of the subproject. Activity sequencing: Determine the dependencies between each subproject and finally generate a document form. Activity resource estimating: Estimate the quantity and type of resources, equipment, and other resources required for each subproject. Activity duration estimating: Estimate the time at which a single subproject is completed. Schedule development: The analysis is performed first and then the project schedule is developed for the sequence of activities, time and resource requirements. Schedule control: Supervise the status of project activities, control changes to project schedules, and finally ensure that projects are completed within the required timeframe.

Efficiently control the progress, and then analyze the factors that affect the progress one by one and take necessary precautions through pre-judgment. Ensure that the actual progress is minimally offset from the planned progress and proactively control the project. There are many factors that influence software project development, technical factors, environmental factors, funding factors, human factors, and so on. The human factor is the most important factor in the implementation of software development projects, and the technical factors are in other words human factors. The principal problem of software development project schedule control management is still pondered in the thoughtfulness of various elements.

CHAPTER 6

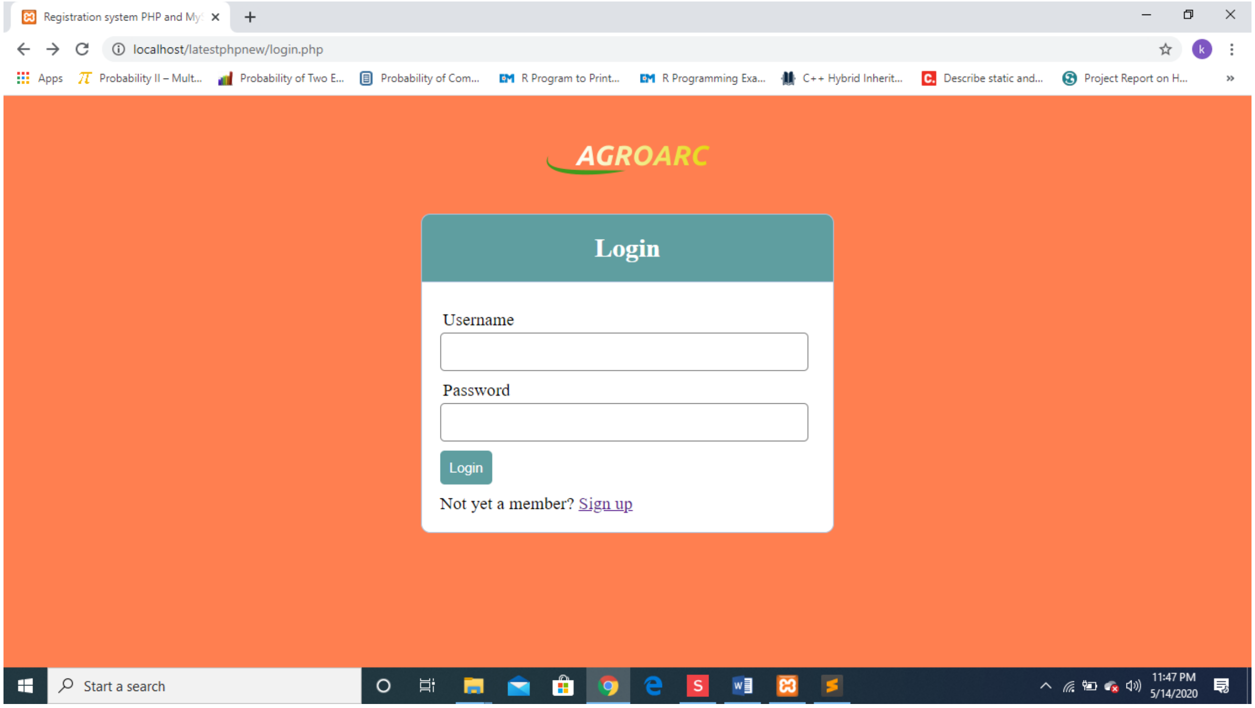
PROJECT DEMONSTRATION

WEBSITE

AgroArc

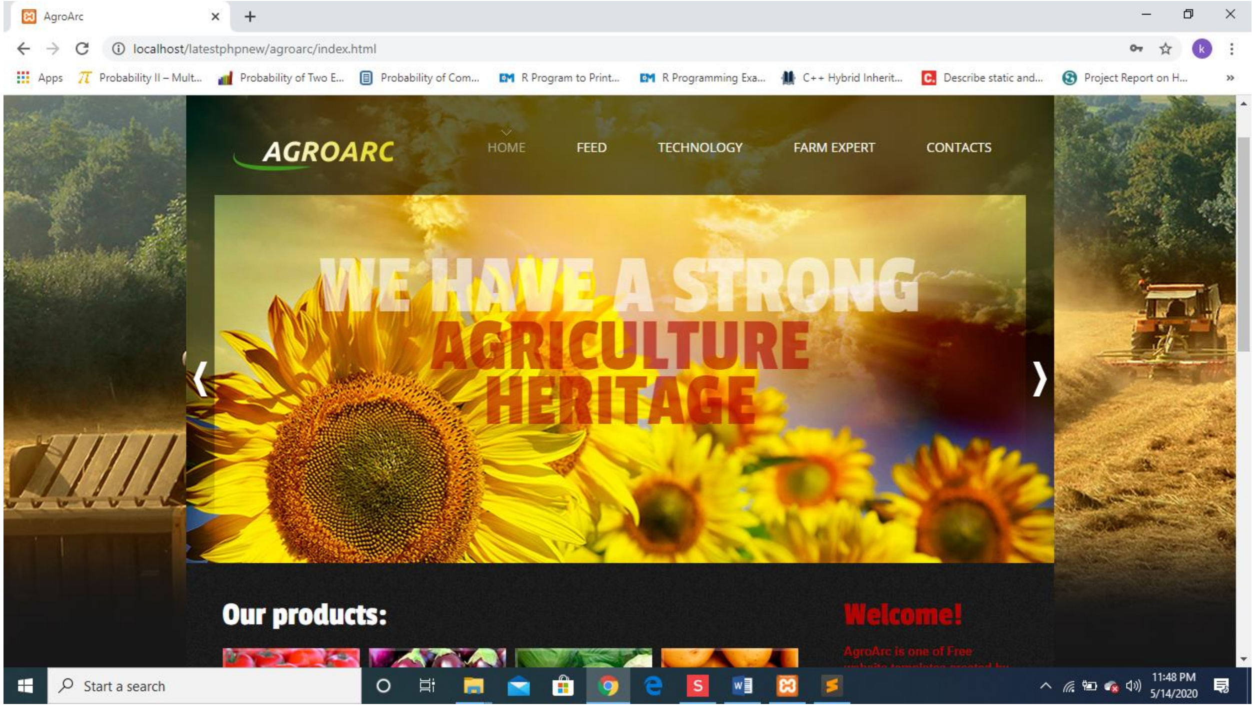
Login Page

Login page to get into our page with the credentials and the sign up page to give all our details to get stored and activate our session on login.



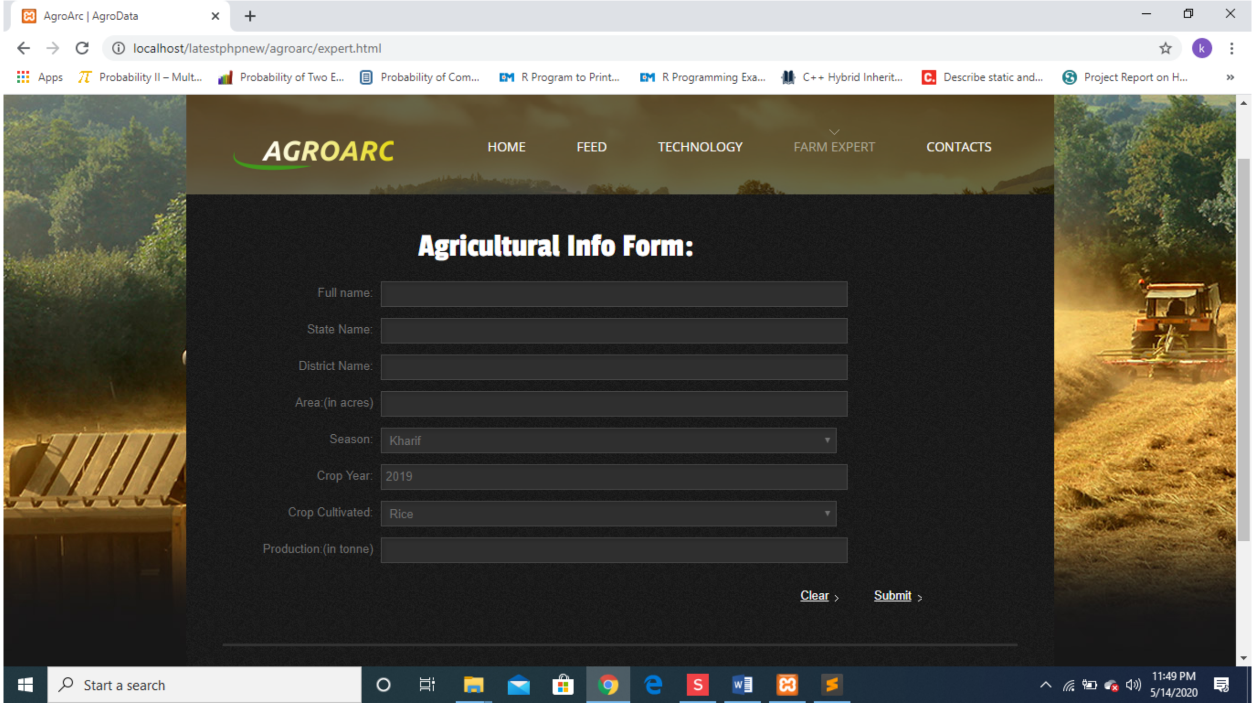
Home Page

Main page of the website to get to know about farming to the budding farmers.



Agricultural Info Form

This page is used to fetch information from agriculturalists about their land and crop details in order to create our dataset.

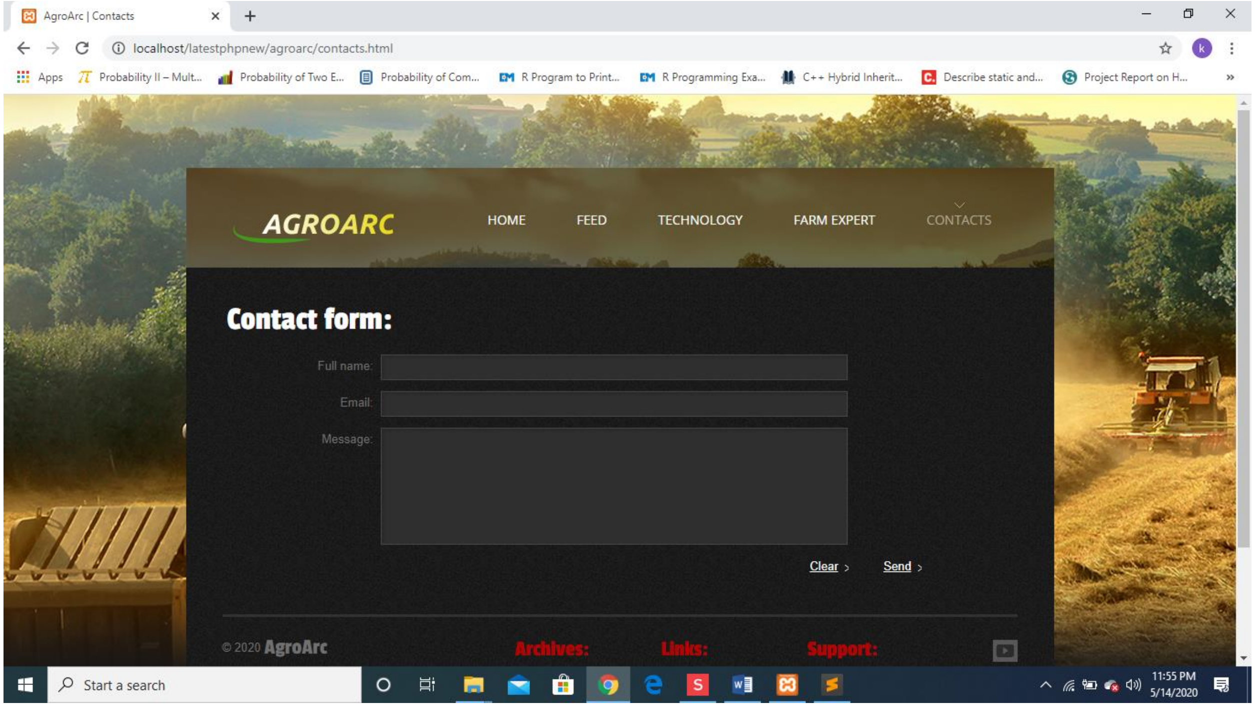


Agricultural Info Form:

Full name:
State Name:
District Name:
Area:(in acres)
Season: Kharif
Crop Year: 2019
Crop Cultivated: Rice
Production:(in tonne)

Contact Form

This is for the users to contact the admin and the other agricultural experts for any of the queries.



Contact form:

Full name:

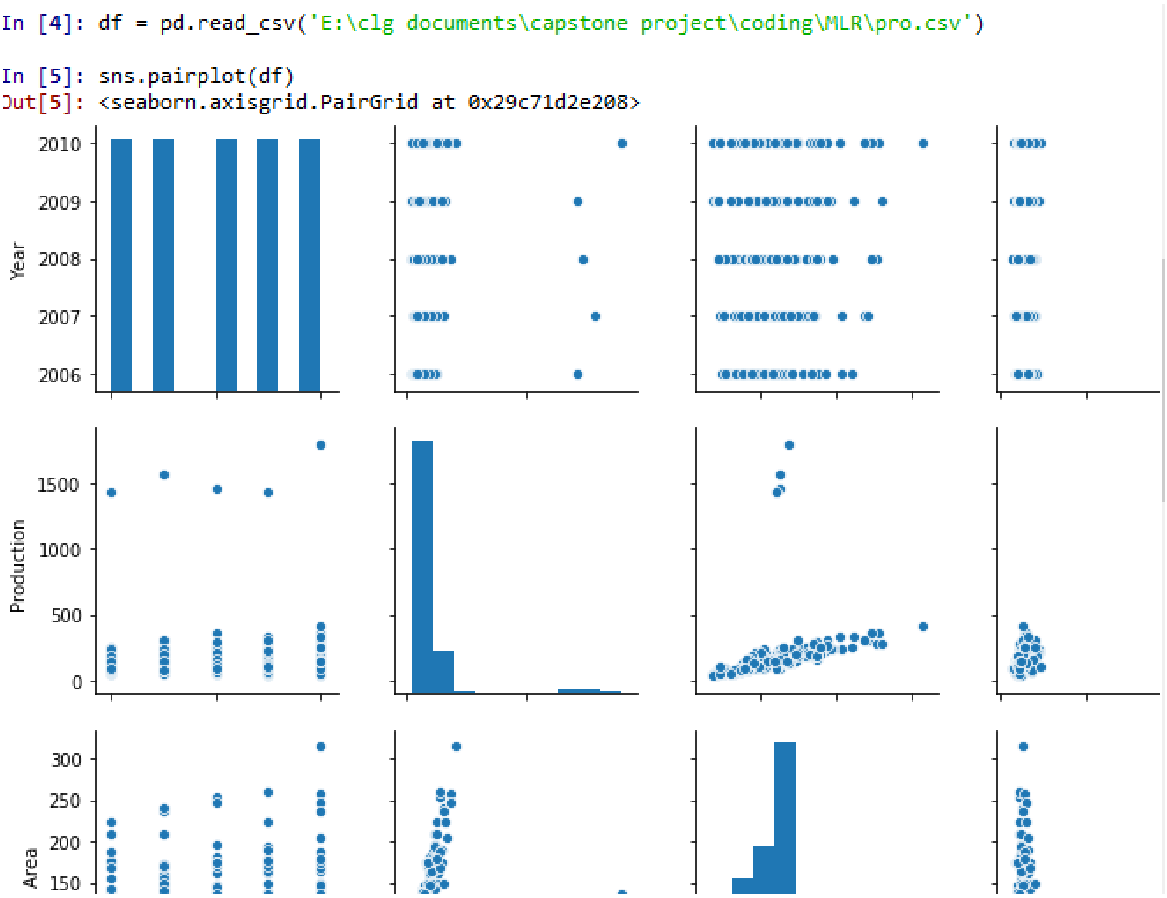
Email:

Message:

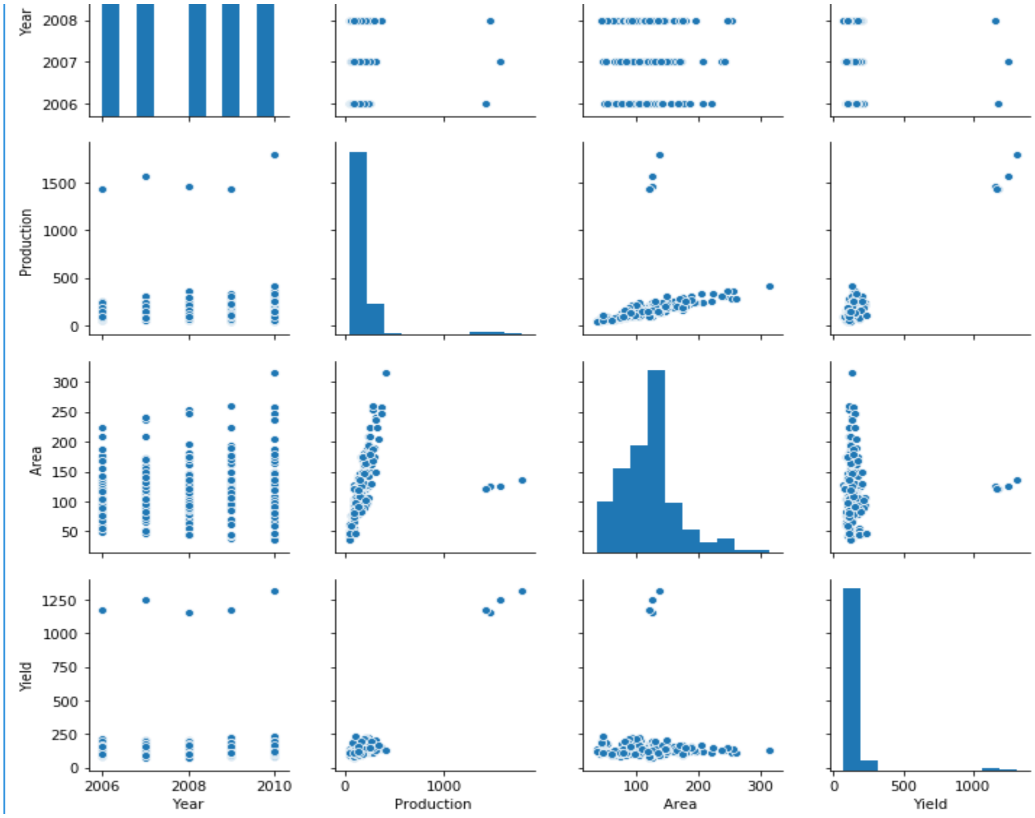
MULTIPLE LINEAR REGRESSION

First we are importing the dataset and the packages.

Then we are on to the pairplot function. It is a function from the seaborn class. It plots a pairwise relationships in a dataset. Here we have the four columns or relations in our dataset to be plotted as a matrix wise plots.

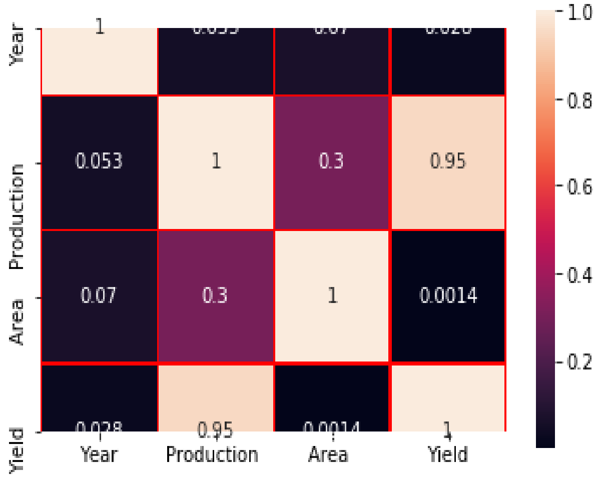


```
In [4]: df = pd.read_csv('E:\clg documents\capstone project\coding\MLR\pro.csv')

In [5]: sns.pairplot(df)
Out[5]: <seaborn.axisgrid.PairGrid at 0x29c71d2e208>
```



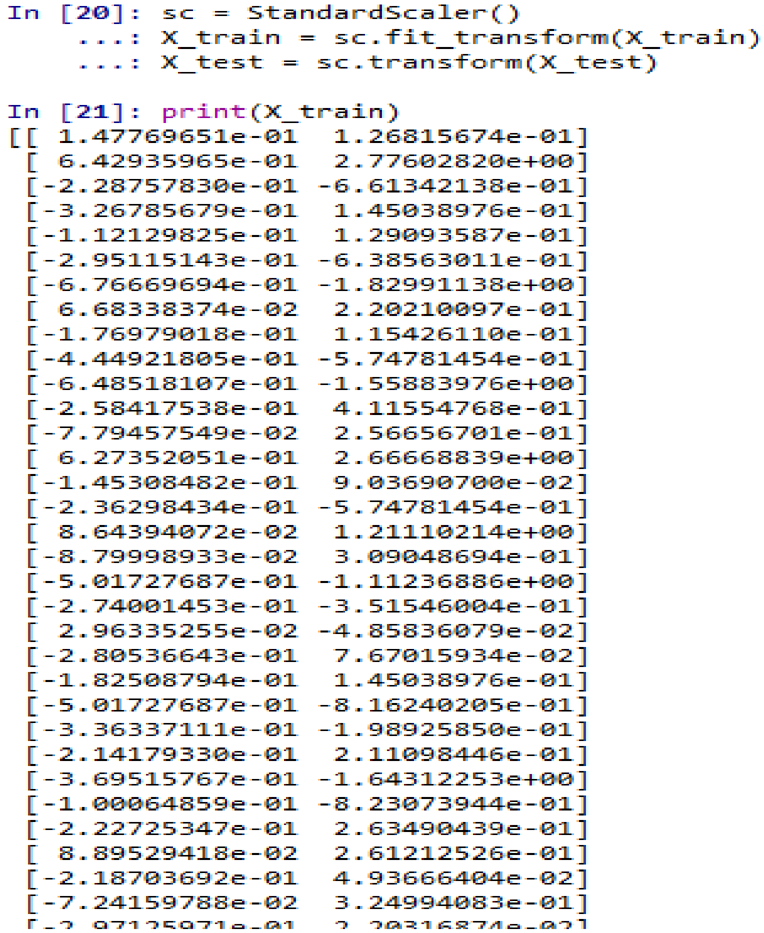
Next we are using heatmap function from the same class seanborn. A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors.

```
In [6]: sns.heatmap(df.corr(), linewidth=0.2, vmax=1.0, square=True, linecolor='red', annot=True)
Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x29c72b53bc8>
```



We have partitioned the dataset into the two groups one is the train set and the other is the test set. Now we have the train set values.

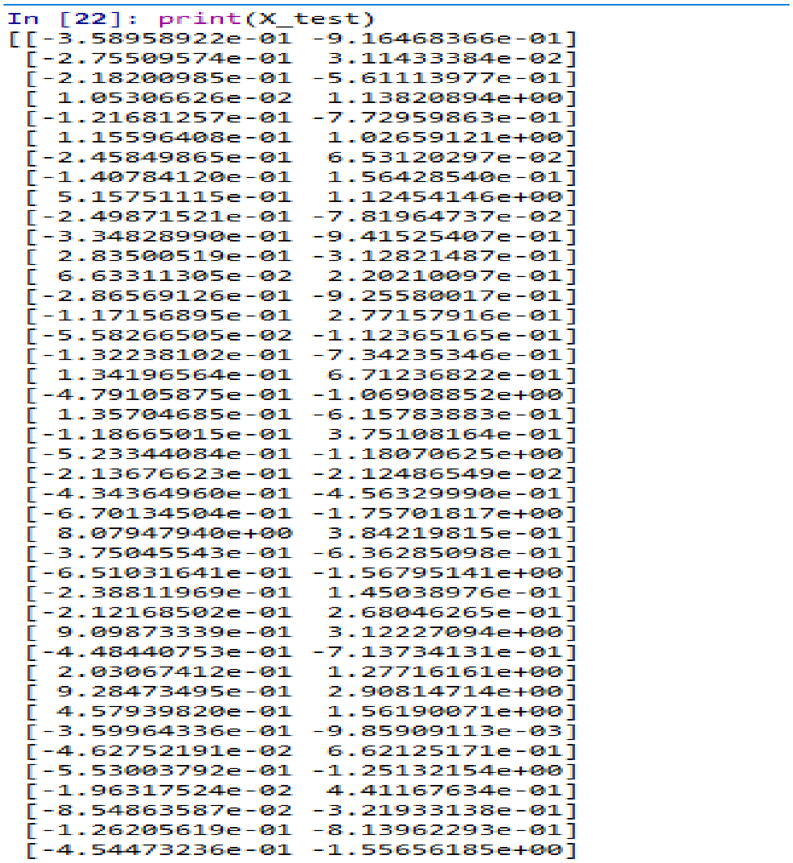
Print(X_train)



```
In [20]: sc = StandardScaler()
    ...: X_train = sc.fit_transform(X_train)
    ...: X_test = sc.transform(X_test)

In [21]: print(X_train)
[[ 1.47769651e-01  1.26815674e-01]
 [ 6.42935965e-01  2.77602820e+00]
 [-2.28757830e-01 -6.61342138e-01]
 [-3.26785679e-01  1.45038976e-01]
 [-1.12129825e-01  1.29093587e-01]
 [-2.95115143e-01 -6.38563011e-01]
 [-6.76669694e-01 -1.82991138e+00]
 [ 6.68338374e-02  2.20210097e-01]
 [-1.76979018e-01  1.15426110e-01]
 [-4.44921805e-01 -5.74781454e-01]
 [-6.48518107e-01 -1.55883976e+00]
 [-2.58417538e-01  4.11554768e-01]
 [-7.79457549e-02  2.56656701e-01]
 [ 6.27352051e-01  2.66668839e+00]
 [-1.45308482e-01  9.03690700e-02]
 [-2.36298434e-01 -5.74781454e-01]
 [ 8.64394072e-02  1.21110214e+00]
 [-8.79998933e-02  3.09048694e-01]
 [-5.01727687e-01 -1.11236886e+00]
 [-2.74001453e-01 -3.51546004e-01]
 [ 2.96335255e-02 -4.85836079e-02]
 [-2.80536643e-01  7.67015934e-02]
 [-1.82508794e-01  1.45038976e-01]
 [-5.01727687e-01 -8.16240205e-01]
 [-3.36337111e-01 -1.98925850e-01]
 [-2.14179330e-01  2.11098446e-01]
 [-3.69515767e-01 -1.64312253e+00]
 [-1.00064859e-01 -8.23073944e-01]
 [-2.22725347e-01  2.63490439e-01]
 [ 8.89529418e-02  2.61212526e-01]
 [-2.18703692e-01  4.93666404e-02]
 [-7.24159788e-02  3.24994083e-01]
 [-2.97125971e-01  2.20316874e-02]
```

Now we have the test set values.

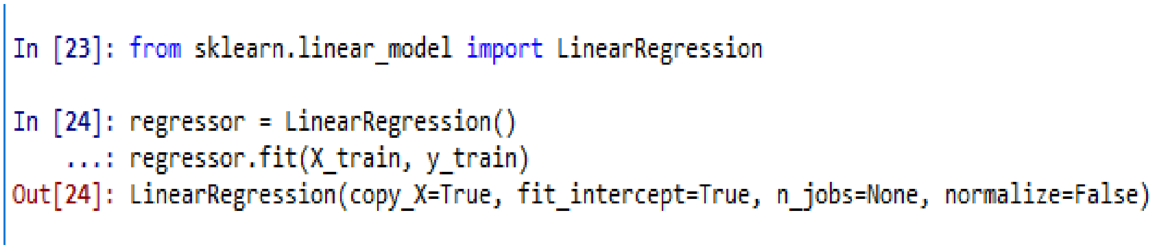
Print(X_test)



```
In [22]: print(X_test)
[[-3.58958922e-01 -9.16468366e-01]
 [-2.75509574e-01  3.11433384e-02]
 [-2.18200985e-01 -5.61113977e-01]
 [ 1.05306626e-02  1.13820894e+00]
 [-1.21681257e-01 -7.72959863e-01]
 [ 1.15596408e-01  1.02659121e+00]
 [-2.45849865e-01  6.53120297e-02]
 [-1.40784120e-01  1.56428540e-01]
 [ 5.15751115e-01  1.12454146e+00]
 [-2.49871521e-01 -7.81964737e-02]
 [-3.34828990e-01 -9.41525407e-01]
 [ 2.83500519e-01 -3.12821487e-01]
 [ 6.63311305e-02  2.20210097e-01]
 [-2.86569126e-01 -9.25580017e-01]
 [-1.17156895e-01  2.77157916e-01]
 [-5.58266505e-02 -1.12365165e-01]
 [-1.32238102e-01 -7.34235346e-01]
 [ 1.34196564e-01  6.71236822e-01]
 [-4.79105875e-01 -1.06908852e+00]
 [ 1.35704685e-01 -6.15783883e-01]
 [-1.18665015e-01  3.75108164e-01]
 [-5.23344084e-01 -1.18070625e+00]
 [-2.13676623e-01 -2.12486549e-02]
 [-4.34364960e-01 -4.56329990e-01]
 [-6.70134504e-01 -1.75701817e+00]
 [ 8.07947940e+00  3.84219815e-01]
 [-3.75045543e-01 -6.36285098e-01]
 [-6.51031641e-01 -1.56795141e+00]
 [-2.38811969e-01  1.45038976e-01]
 [-2.12168502e-01  2.68046265e-01]
 [ 9.09873339e-01  3.12227094e+00]
 [-4.48440753e-01 -7.13734131e-01]
 [ 2.03067412e-01  1.27716161e+00]
 [ 9.28473495e-01  2.90814714e+00]
 [ 4.57939820e-01  1.56190071e+00]
 [-3.59964336e-01 -9.85909113e-03]
 [-4.62752191e-02  6.62125171e-01]
 [-5.53003792e-01 -1.25132154e+00]
 [-1.96317524e-02  4.41167634e-01]
 [-8.54863587e-02 -3.21933138e-01]
 [-1.26205619e-01 -8.13962293e-01]
 [-4.54473236e-01 -1.55656185e+00]
```

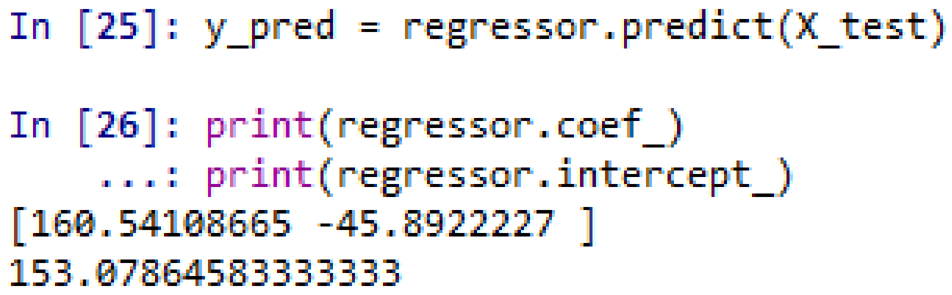
We have done the regression using the sklearn package and with the help of the train set of values.

Regression Execution



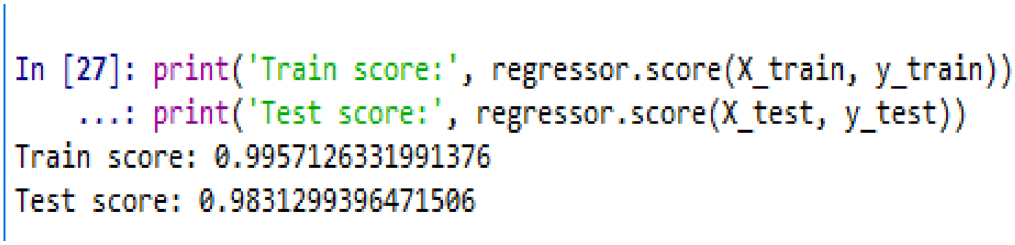
```
In [23]: from sklearn.linear_model import LinearRegression

In [24]: regressor = LinearRegression()
    ...: regressor.fit(X_train, y_train)
Out[24]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

Printing the Co-efficient and the y-intercept value



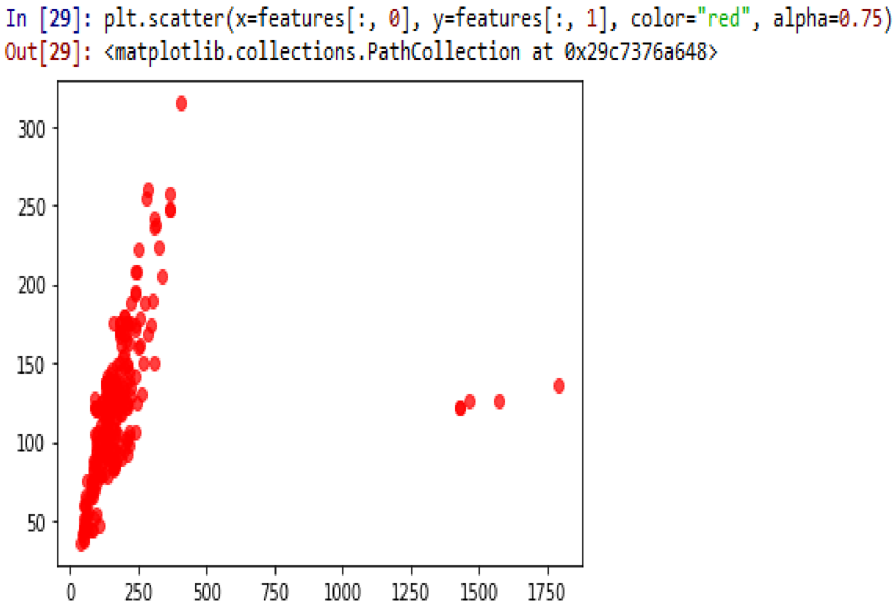
```
In [25]: y_pred = regressor.predict(X_test)

In [26]: print(regressor.coef_)
    ...: print(regressor.intercept_)
[160.54108665 -45.8922227 ]
153.07864583333333
```

Obtaining the test and the train score



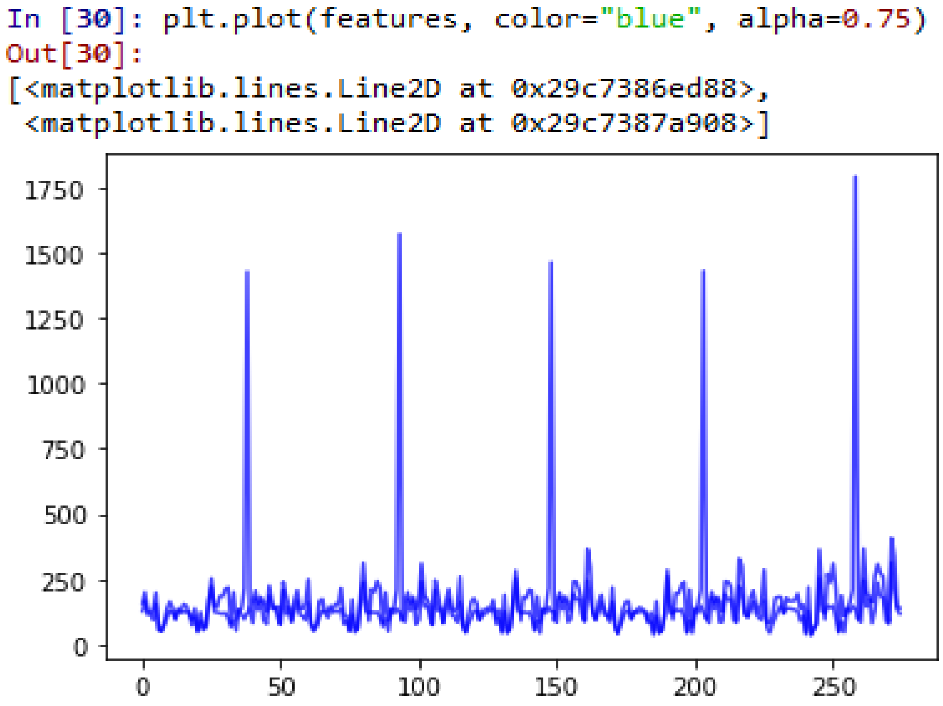
```
In [27]: print('Train score:', regressor.score(X_train, y_train))
    ...: print('Test score:', regressor.score(X_test, y_test))
Train score: 0.9957126331991376
Test score: 0.9831299396471506
```

Scatter plot



```
In [29]: plt.scatter(x=features[:, 0], y=features[:, 1], color="red", alpha=0.75)
Out[29]: <matplotlib.collections.PathCollection at 0x29c7376a648>
```

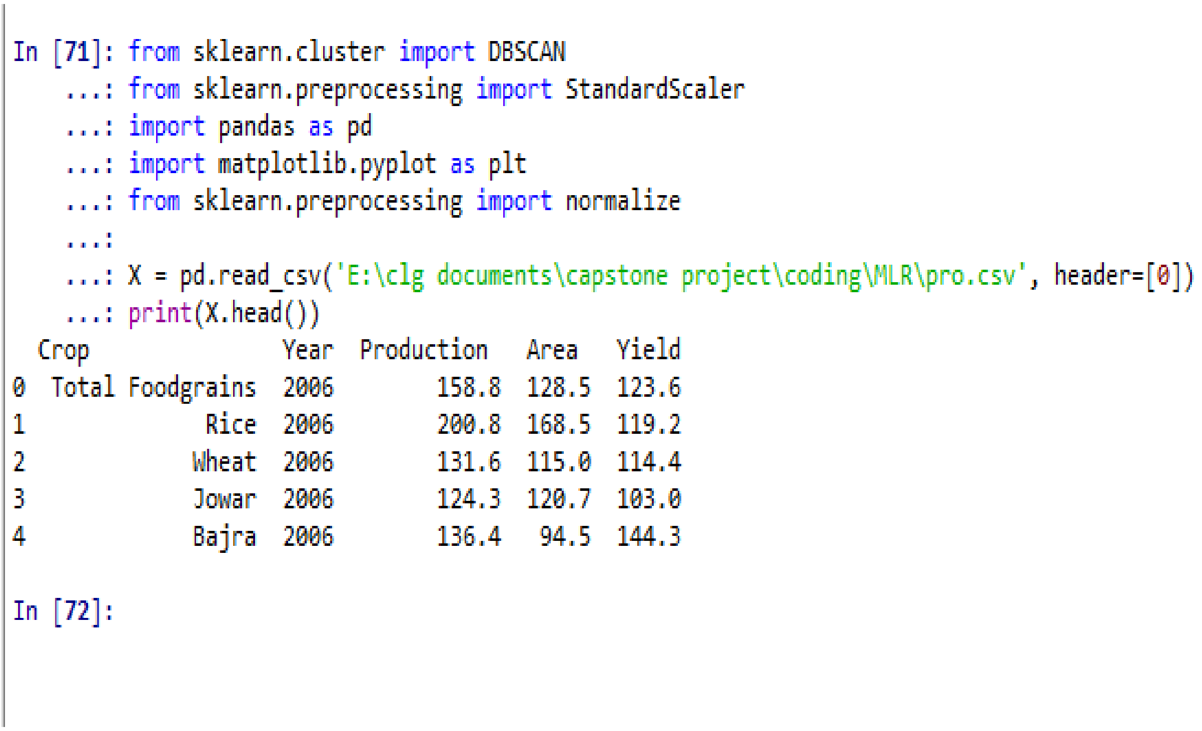
General plot



```
In [30]: plt.plot(features, color="blue", alpha=0.75)
Out[30]:
[<matplotlib.lines.Line2D at 0x29c7386ed88>,
 <matplotlib.lines.Line2D at 0x29c7387a908>]
```

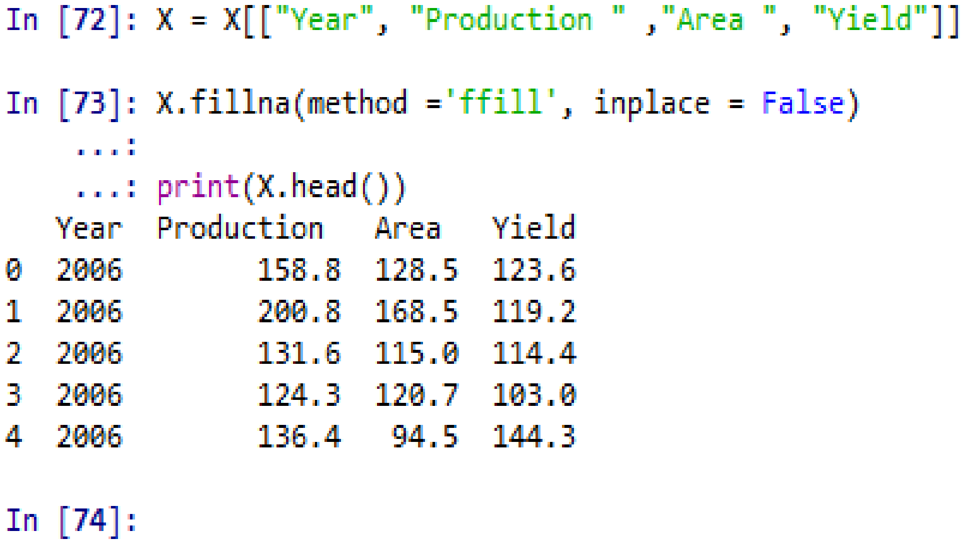
DBSCAN

Importing the necessary packages and the dataset with the appropriate constraints and a sample of the 5 rows have been printed.



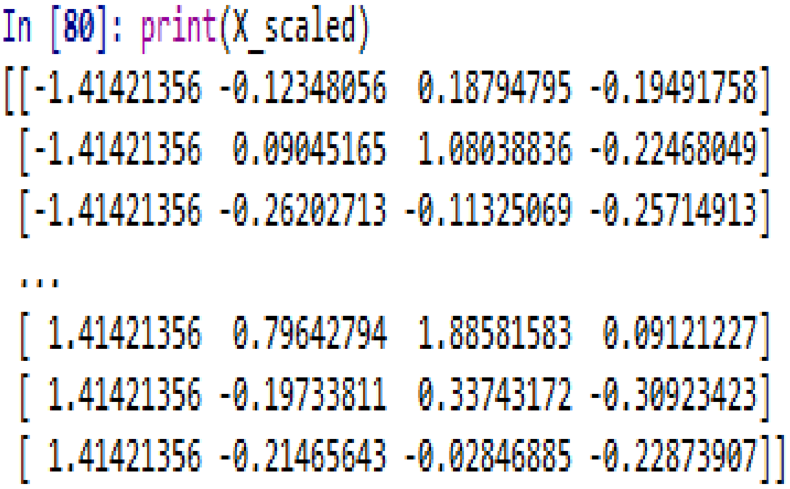
```
In [71]: from sklearn.cluster import DBSCAN
    ...: from sklearn.preprocessing import StandardScaler
    ...: import pandas as pd
    ...: import matplotlib.pyplot as plt
    ...: from sklearn.preprocessing import normalize
    ...:
    ...: X = pd.read_csv('E:\clg documents\capstone project\coding\MLR\pro.csv', header=[0])
    ...: print(X.head())
  Crop              Year  Production   Area  Yield
0  Total Foodgrains  2006       158.8  128.5  123.6
1              Rice  2006       200.8  168.5  119.2
2             Wheat  2006       131.6  115.0  114.4
3             Jowar  2006       124.3  120.7  103.0
4             Bajra  2006       136.4   94.5  144.3

In [72]:
```

Dropping the inappropriate columns from the dataset and filling the missing values in the dataset.



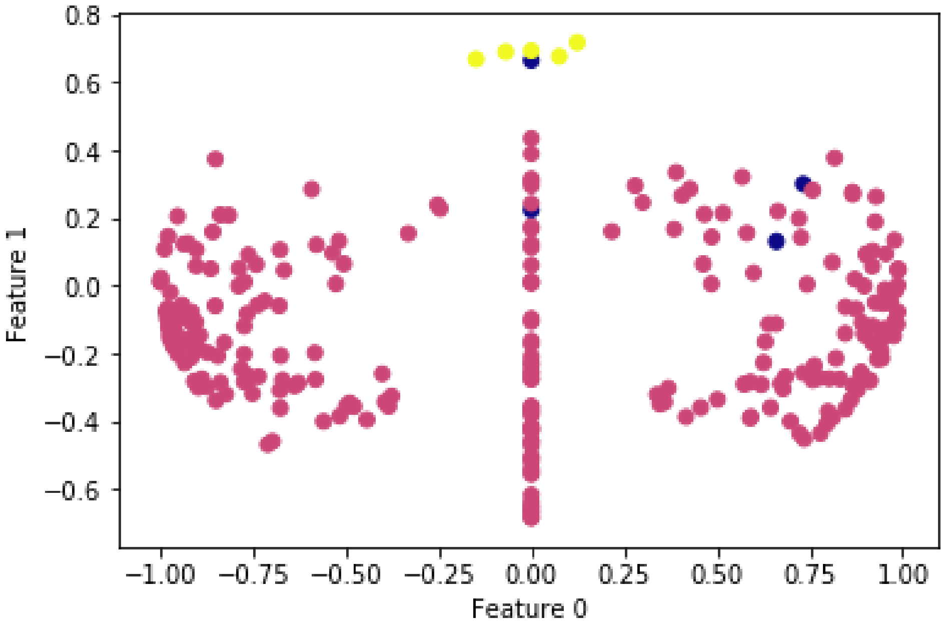
```
In [72]: X = X[["Year", "Production " ,"Area ", "Yield"]]

In [73]: X.fillna(method ='ffill', inplace = False)
    ...:
    ...: print(X.head())
   Year  Production   Area   Yield
0  2006       158.8  128.5  123.6
1  2006       200.8  168.5  119.2
2  2006       131.6  115.0  114.4
3  2006       124.3  120.7  103.0
4  2006       136.4   94.5  144.3

In [74]:
```

Scaling the dataset in order to bring all the attributes to a comparable level.

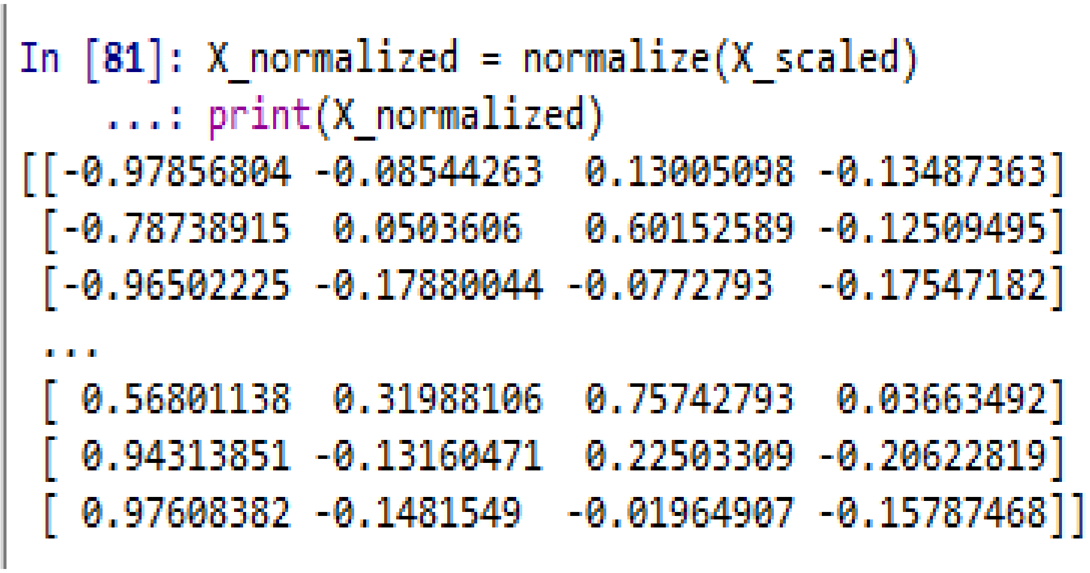


```
In [80]: print(X_scaled)
[[-1.41421356 -0.12348056  0.18794795 -0.19491758]
 [-1.41421356  0.09045165  1.08038836 -0.22468049]
 [-1.41421356 -0.26202713 -0.11325069 -0.25714913]
 ...
 [ 1.41421356  0.79642794  1.88581583  0.09121227]
 [ 1.41421356 -0.19733811  0.33743172 -0.30923423]
 [ 1.41421356 -0.21465643 -0.02846885 -0.22873907]]
```

Raw plot before normalizing

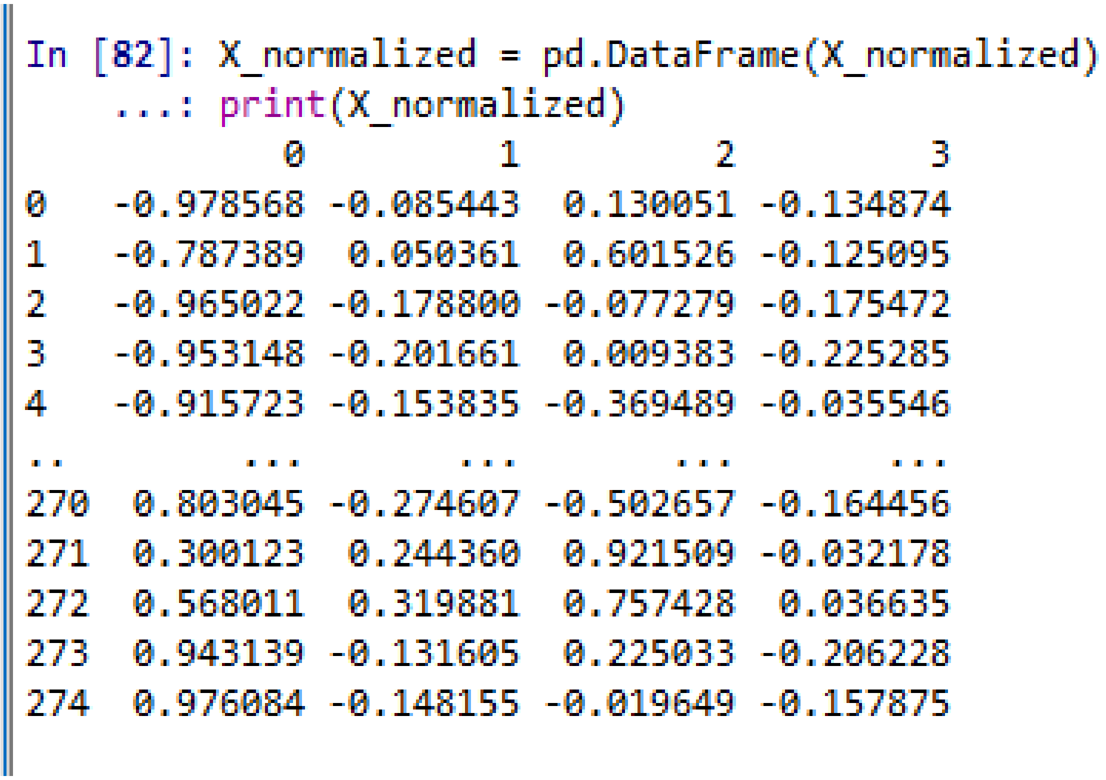


The value is normalized to be followed with a Gaussian Distribution.



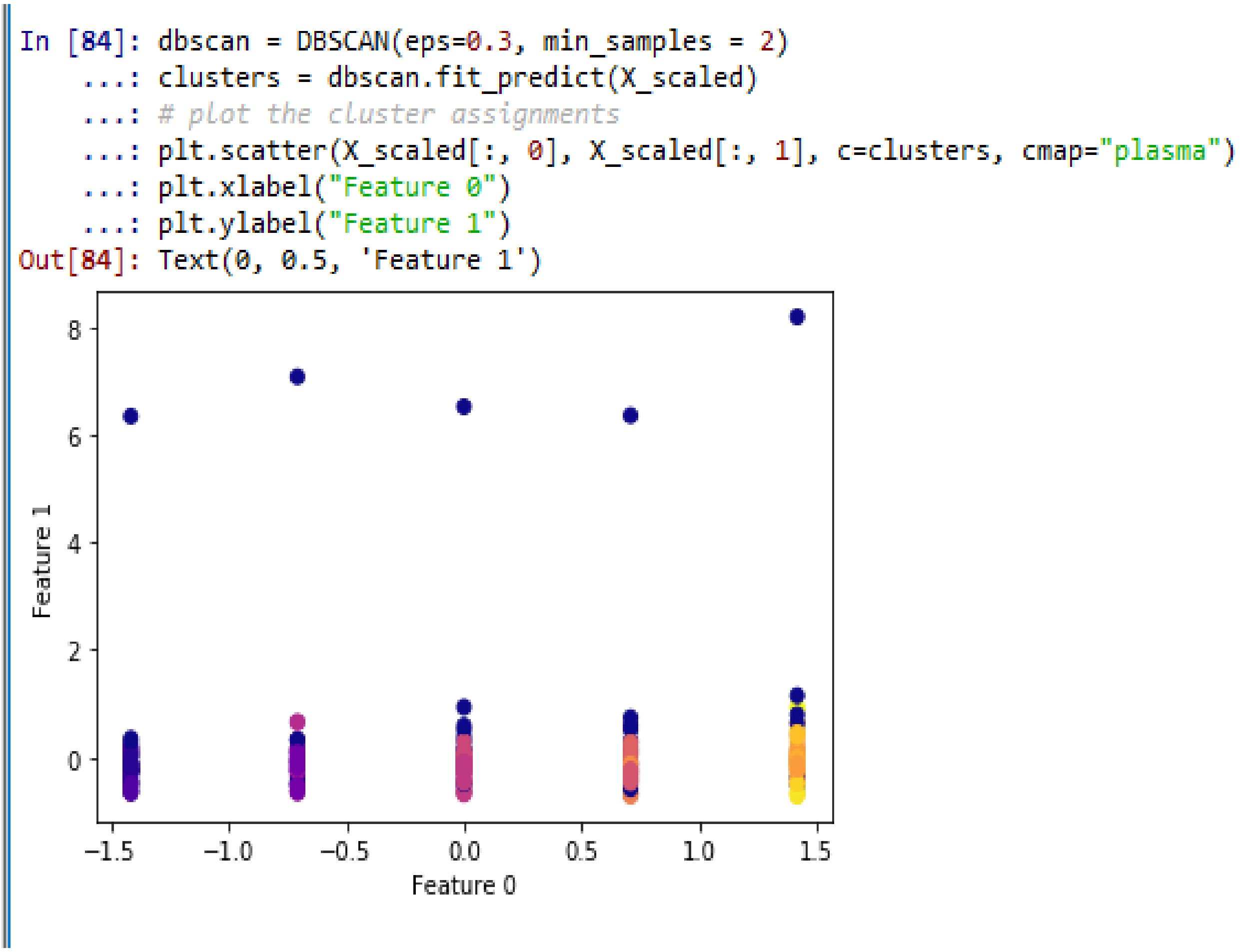
```
In [81]: X_normalized = normalize(X_scaled)
    ...: print(X_normalized)
[[-0.97856804 -0.08544263  0.13005098 -0.13487363]
 [-0.78738915  0.0503606   0.60152589 -0.12509495]
 [-0.96502225 -0.17880044 -0.0772793  -0.17547182]
 ...
 [ 0.56801138  0.31988106  0.75742793  0.03663492]
 [ 0.94313851 -0.13160471  0.22503309 -0.20622819]
 [ 0.97608382 -0.1481549  -0.01964907 -0.15787468]]
```

Values after converting it into the pandas data frame for the accurate fields and the columns structured back into the same.



```
In [82]: X_normalized = pd.DataFrame(X_normalized)
    ...: print(X_normalized)
            0         1         2         3
0   -0.978568 -0.085443  0.130051 -0.134874
1   -0.787389  0.050361  0.601526 -0.125095
2   -0.965022 -0.178800 -0.077279 -0.175472
3   -0.953148 -0.201661  0.009383 -0.225285
4   -0.915723 -0.153835 -0.369489 -0.035546
..        ...       ...       ...       ...
270  0.803045 -0.274607 -0.502657 -0.164456
271  0.300123  0.244360  0.921509 -0.032178
272  0.568011  0.319881  0.757428  0.036635
273  0.943139 -0.131605  0.225033 -0.206228
274  0.976084 -0.148155 -0.019649 -0.157875
```

Here the values is been Plotted after the Pre-processing is done to obtain a complete Clustering pattern from the dataset collected.



```
In [84]: dbscan = DBSCAN(eps=0.3, min_samples = 2)
    ...: clusters = dbscan.fit_predict(X_scaled)
    ...: # plot the cluster assignments
    ...: plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=clusters, cmap="plasma")
    ...: plt.xlabel("Feature 0")
    ...: plt.ylabel("Feature 1")
Out[84]: Text(0, 0.5, 'Feature 1')
```

CHAPTER 7

RESULT & DISCUSSION

CONCLUSION

Agriculture is the most significant application area mainly in the arising countries like India. Data mining is used for large data in agriculture and extraction of knowledge is big challenge. The crop yield prediction is still remaining as a challenging issue for farmers, one can make the use of these techniques in agricultural field that will creates conditions for mankind satisfactory decisions and with that achieving challenging improvement. The present study demonstrated the potential use of data mining and machine learning techniques in predicting the crop yield based on the geographic input parameters. The developed webpage is user friendly and the accuracy of predictions is above 95 per cent in all the crops and districts selected in the study indicating higher accuracy of prediction. The user friendly web page developed for predicting crop yield can be used by any user their choice of crop by providing agricultural data of that place. Crop data on respective crops of various district is collected from the web resource is for exploratory data analysis.

Different data excavation techniques are actualized on the gathered data to survey the simplest execution yielding strategy. The foremost recent work has been done using data mining technique called DBSCAN to procure the clusters of yield for year, area and production. The DBSCAN algorithm forms clusters that are denser and noise free, the clusters are often to be of any shape. Machine learning technique called Multiple Linear Regression also utilized to fetch accurate prediction of the crop yield. Through Multiple Linear regression we are able to get a yield that is more optimal and accurate. As the resultant it has been analyzed that Multiple Linear Regression gives the optimal prediction quality over DBSCAN. This helps the farmers and agribusiness to arrive at future decisions.

FUTURE WORK

This research reports crop yield prediction power of the algorithm. This research work can be enhancing to the next level. In future we can decide the effective algorithm based on their accuse metrics that will assist to select an efficient algorithm for crop yield prediction. Further research is to construct an integrated technique for selecting the most reserve regression technique for each crop dataset.

We can build a recommender system of agriculture production and distribution for farmer. This current system works for structured dataset. In future we can implement data independent system also. It implies format of data whatever, our system should figure out with same effectiveness. There is no necessity for the data structuring, data pre processing and constraints for the minimal group of records. The system should work best for the lower limit of the records.

Along with machine learning algorithms for prediction, it is planned to study the impact of big data techniques in the prediction of crop yield. A conceptual approach is proposed for the same. The proposed approach is being implemented.

Hence by using the soil, weather and market prices, we can build a model, having a workflow as shown above, that can provide accurate predictions about the crop yield suitable for a particular region. These are performed by using Artificial Neural Network (ANN).

Finally, it is necessary to point out that this work deals only with the predictive accuracy of the Multiple Linear Regression and the skilful representation of the data points with the efficient Clustering algorithm. Machine learning techniques are complex, and several factors are related with their performance measuring. Further research will be dedicated to compare the characteristics of ML algorithms and their compatibility with agricultural planning.

By enforcing machine learning to sensor data, farm management schemes are developing into real artificial intelligence organisations, providing richer recommendations and penetrations for the subsequent conclusions and actions with the ultimate background of production betterment. For this scope, in the future, it is expected that the usage of ML models will be even more widespread, allowing for the possibility of integrated and applicable tools. At the moment, all of the approaches regard individual approaches and solutions and are not adequately connected with the decision-making process, as seen in other application domains. This integration of automated data recording, data analysis, ML implementation, and decision-making or support will provide practical tools that come in line with the so-called knowledge-based agriculture for increasing production levels and bio-products quality.

As of now, we are foreseeing just the yield for particular products. We can include the forecast for venture required for various yields. Components like manure, pesticide, arrangement of the ranch for sowing, cultivate gear and bore wells assume a vital part in choosing which yield to develop. We can likewise propose compost supplement needs of the dirt if the agriculturist gives the dirt examination comes about. The recommendations for intercropping and money harvests can likewise be added to expand profitability. A few vegetation files like Normalized Difference Vegetation Index (NDVI), Vegetation Condition Index (VCI) and Temperature Condition Index (TCI) can be utilized to distinguish draft conditions and a few other climate impacts on the yield of the harvest.

In future, one can do investigation to understand that how these techniques can be used with large and complex agriculture datasets and can also be used for crop yield prediction spatially by making the use of GIS technologies.

In this paper, comparison of various data mining & machine learning techniques are made for the smaller datasets and found that higher the accuracy higher will be rate for crop yield prediction.
By making use of large datasets, one can improve the results.

CHAPTER 8

SUMMARY

ORGANIZATION OF THE THESIS

The thesis is structured into nine chapters. This section provides the essence of the chapters that provide quick overview of the research work carried out.

Chapter 1

Introduction

In this chapter an Introduction is been given about the research field. A brief description about agriculture and its importance, agriculture and the role of agro in Indian economy. The importance of crop yield in the agricultural processes. The description about the factors that determine the crop yield. The data mining and the correspondence of data mining in agriculture and the crop yield prediction. The machine learning techniques and its usage in the crop yield prediction. The motivation and objective tells about the need for the yield prediction and the exact goal of the field work. The review of literature is made on data mining techniques and algorithms and machine learning techniques and others with the paper author and their techniques used and the pros and cons of those techniques utilised. The Background area describes the measures taken and the applications of the agriculture from the previous techniques and research. The problem statement formulates the challenges faced by the farmers and the retreats in the algorithms used.

Chapter 2

Project Description and Goals

The Existing System elaborates the surviving models in this crop yield prediction field with its advantages and disadvantages and a brief description about the existing models and techniques used on their research. The objective of the paper is to propose a well – defined model for the crop yield prediction. It also describes the goals of the crop yield prediction. Scope explains the area of interest of this research and the domain where the obtained results could be applied and utilized.

Chapter 3

Technical Specification

Here the identification of the Requirements specification, which includes the hardware and the software demands for the project. The memory and the security requirements have also been specified.

Chapter 4

Design Approach and Details

In this chapter we present an elaborated conception of the project or so called the system design. In this chapter execution of the above mentioned two algorithms is done. The modules have been split and described. The description of the technologies used on our implementation is been done. The Multiple linear regression and the DBSCAN is been carried out using the gathered crop dataset. Kharif, Rabi and Zaid crop data sets across various years were subjected to hybrid data mining techniques after 12 preprocessing of the raw data. This chapter commences with the explanation of these algorithms and followed by its procedure and the steps involved in them. Later a code is generated and they have been executed in the console. A few sample test cases is been given to evaluate their accurate working.

Chapter 5

Schedule, Tasks and Milestones

Here we have mentioned about the Schedules taken in our project. The factors that affect our schedules and tasks and the management of our time. The tasks we carried out using the blueprint and the time allotted for each of these sections and the chapters.

Chapter 6

Project Demonstration

In this chapter we give sample screenshots of the above executed procedure and with the clear and brief description of each and every step of execution of the MLR and DBSCAN algorithm.

Chapter 7

Result & Discussion

In this chapter, analysis of crop yield using the various factors such as seasons, production, year is considered in different regions. The results are obtained from the hybrid data mining technique. Results and analysis of a hybrid approach for analysis of dynamic changes in spatial data were done. A statistical evaluation of the algorithm (DBSCAN) in clustering task is also carried out. The study shows that the hybrid algorithm is found to be a potential method for clustering tasks as its performance of accuracy is maximal. The MLR gives the accuracy rate to the actual data using the various values such as intercepts and coefficients. It further moves to the future word to be done in the paper to improve the performance and the accuracy rate in comparison with the rest of all the algorithms to fetch the best of all the other.

Chapter 9

References

The reference for the project with the various paper and the research work is been given. It also gives the year includes the authors and the Journal title with the techniques utilised.

CHAPTER 9

REFERENCES

[1] Bhuvana, Dr.C.Yamini (2015), ‘Survey on Classification Algorithms in Data mining.’ International Conference on Recent Advances in Engineering Science and Management.

[2] D Ramesh, B Vishnu Vardhan, “Analysis of Crop Yield Prediction using Data Mining Techniques”, International Journal of Research in Engineering and Technology (IJRET),Vol.4, 2015.

[3] DakshayiniPatil, Dr. M .S Shirdhonkar, “Rice Crop Yield Prediction using Data Mining Techniques: An Overview”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 5, ISSN: 2277 128X ,2017.

[4] D Ramesh , B Vishnu Vardhan. “Data Mining Techniques and Applications to Agricultural Yield Data”. International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 9, September 2013,pp.3477-3480.

[5] D. Diepeveen and L. Armstrong, “Identifying key crop performance traits using data mining” World Conference on Agriculture, Information and IT, 2008.

[6] Ye, Nong; Data Mining: Theories, Algorithms, and Examples, CRC Press, 2013.

[7] Mohammad Motiur Rahman, Naheena Haq and Rashedur M Rahman “Comparative Study of Forecasting Models on Clustered Region of Bangladesh to Predict Rice Yield”, 17th. IEEE International Conference on Computer and Information Technology (ICCIT), Dhaka, 2014.

[8] Tng Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms", Proceedings of the twenty-first international conference on Machine Learning.

[9] Rossana MC, L. D. (2013). A Prediction Model Framework for Crop Yield Prediction. Asia Pacific Industrial Engineering and Management Society Conference Proceedings Cebu, Philippines, 185.

[10] J. Liu, C. E. Goering, Lei Tian, 2001. “A neural network for setting target corn yields”. Transactions of the American Society of Agricultural Engineers44 (3):705-713.

[11] SamuelA.L,”Some Studies in Machine Learning Using the Game of Checkers”. IBM J.Res. Dev. 1959, 44,206-226.

[12] Aakash G Ratkul, Gangadhar Akalwadi, Vinay N Patil, Kavi Mahesh“Farmers adviser system” 2016 IEEE International conference on cloud computing in emerging markets.

[13] Yan Xiaozhen, Xie Hong, Wang Tong, “A multiple linear regression data predicting method using correlation analysis for wireless sensor networks”, 2011 Cross Strait Quad-regional Radio Science and Wireless Technology Conference.

[14] Aakunuri Manjula and Dr. G. Narsimha, “XCYPF: A flexible and extensible framework for agricultural crop yield prediction”, IEEE Sponsered 9th ISCO 2015.

[15] Ronald Cody, Ed. D., Robert wood Johnson Medical school, Piscataway, NJ, “Data cleaning 101”.

[16] "A survey on pre-processing and post processing techniques in data mining "Tomar, Divya, and Sonali Agarwal International Journal of Database Theory & Application 7.4 (2014)”.

[17] Enhancing data analysis with noise removal." Xiong, Hui, et al, IEEE Transactions on Knowledge and Data Engineering 18.3 (2006): 304- 319.

[18] Data clustering: a review by Jain A, Murty MN, Flynn PJ. ACM Comput Surv. 1999.

[19] MotiurRahman M, Haq N, Rahman RM. Application of data mining tools for rice yield prediction on clustered regions of Bangladesh. IEEE. 2014;2014:8–13.

TEAM NUMBER – 1

Name: KAMESH R

Register Number: 17BCA0091

Date of Birth: 20/10/1999

Blood Group: B+

Phone Number: 6380252775

Email: kmshravi@gmail.com

Address: No 74, Jayaram Chetty Street,
Vellore fort, Vellore - 632004.

TEAM NUMBER – 2

Name: JAGAN T

Register Number: 17BCA0035

Date of Birth: 14/03/1999

Blood Group: O+

Phone Number: 7868055109

Email: jagant14399@gmail.com

Address: 550(3)/3, CHITTOOR MAIN ROAD
PUDUR MEDU,
TIRUVALLUR DISTRICT -631302
TAMIL NADU, INDIA.

TEAM NUMBER – 3

Name: Mohammed Arham Rayyan J

Register Number:17BCA0068

Date of Birth: 04/05/2000

Blood Group: B+

Phone Number: +91 99942 10572

Email: mohammedarhamrayyanj@gmail.com

Address: 11 Mullah mohammed kasim street, khaderpet, vaniyambadi. -635751.