**AGRICULTURE CROP PRODUCTION USING  
MACHINE LEARNING**

A Project Report Submitted in partial fulfillment of the requirements  
for the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

*Submitted By*

**G.VALLI**

**(16221A0534)**

**(16221A0546)**

**(16221A0548)**

**(16221A0516)**

**K.SRINU**

**K.SAI VAMSI**

**BINOD.P**

Under the Esteemed Guidance of

**Mrs. B.GANGA BHAVANI**

**M.Tech„**

Associate Professor



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
BONAM VENKATA CHALAMAYYA ENGINEERING COLLEGE  
(AUTONOMOUS)**

(Approved by A.I.C.T.E. New Delhi & Affiliated to J.N.T.U.K. Kakinada)

(Accredited by NAAC with "A" Grade)

ODALAREVU-533210  
2016 - 2020

**BONAM VENKATA CHALAMAYYA ENGINEERING COLLEGE**

**(AUTONOMOUS)**

**ODALAREVU-533210**

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



**CERTIFICATE**

This is to certify that the project work entitled **“AGRICULTURE CROP PRODUCTION USING MACHINE LEARNING”** is being submitted for the partial fulfillment of the requirements for the award of the degree of Bachelor Of Technology in **Computer Science and Engineering** to B V C Engineering College, Odalarevu, is a bonafide work done by **G.VALLI (16221A0534), K.SRINU (16221A0546), K.SAI**

**VAMSI (16221A0548), BINOD.P (16221A0516)** under my guidance during the academic year **2016 - 2020** and it has been found suitable for acceptance according to the requirement of the University.

The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree.

**Project Guide**

**Head of the Department**

**Dr. GUNAMANI JENA**

**M.E, Ph.D**

**Professor**

**Mrs. B.GANGHA BHAVANI**

**M.Tech,**

**Associate Professor**

**EXTERNAL EXAMINER**

**ACKNOWLEDGEMENT**

We profoundly grateful to express our deep sense of gratitude and respect towards our guide **Mrs. B.GANGHA BHAVANI, Associate Professor** of CSE Department, B V C Engineering College, Odalarevu, for her excellent guidance right from selection of the project and her valuable suggestions throughout the project work. Her constant encouragement and support has been the cause of our success, in completing this thesis in the college.

We also grateful to the project Coordinator **Dr. P.DEVABALAN Professor** CSE Department, B V C Engineering College, Odalarevu, for her excellent guidance right from selection of the project and her valuable suggestions throughout the project work.

We really thankful to **Dr. GUNAMANI JENA Professor** and **Head of the Department of Computer Science and Engineering** for providing the laboratory facilities to the fullest extent as and when required and also for giving us the opportunity to carry out the project work in the college.

We also thankful to **Dr. D S V PRASAD, Principal,** B V C Engineering College, Odalarevu, for his support during and till the completion of the project.

We thankful to all the **teaching and non-teaching staff** of Computer Science and Engineering Department, **Management,** B V C Engineering College, Odalarevu, and our **friends** for their direct and indirect help provided to us in completing the project.

Project Associates

**G.VALLI (16221A0534)**

**K.SRINU (16221A0546)**

**K.SAI VAMSI (16221A0548)**

**BINOD.P (16221A0516)**

ABSTRACT

The Agriculture plays a dominant role in the growth of country's economy. Climate and other environmental changes has become a major threat in the agriculture field. Machine Learning (ML) is an essential approach for achieving practical and effective solution for this problem. For better crop yield the artificial Neural Network have demonstrated to be an effective tool for modeling and prediction by using Linear Regression Algorithm.

This paper mainly focused on the techniques and measures taken to improve farming by in calculating the technical knowledge and developments in order to make the agricultural sector more reliable and easy for the farmers by predicting the suitable crop by using Machine Learning techniques. By using the technology, precision agricultural technicians can break down a property in a very precise way so that what's planted on each inch of ground will have the optimal conditions for growth. This means looking at factors such as soil composition or water table levels. By using the existing conditions more efficiently, precision agricultural technicians. Across the globe India is the second largest country having people more than 1.3 Billion. Many people are dependent on the agriculture and it is the main resource.

TABLE OF CONTENTS

[ABSTRACT i](#bookmark9)

[TABLE OF CONTENTS ii](#bookmark13)

CHAPTER 1 1

INTRODUCTION 1

1. [Objective of the Study 1](#bookmark15)
2. [Purpose and Scope 2](#bookmark19)
3. [Existing System 2](#bookmark25)
4. [Proposed System 3](#bookmark29)

CHAPTER 2 4

SYSTEM ANALYSIS 4

1. [Problem Analysis 4](#bookmark40)
2. [Types of Crops 4](#bookmark42)
3. [Feasibility Study 10](#bookmark62)
4. [Economical Feasibility 10](#bookmark65)
5. [Technical Feasibility 11](#bookmark67)
6. Social Feasibility 11
7. [Requirement Specification: 11](#bookmark70)
8. [Functional Requirements 13](#bookmark74)
9. [Non- Functional Requirements 13](#bookmark76)
   1. [System Specification: 15](#bookmark34)
      1. [Hardware Requirements 15](#bookmark36)
      2. [Software Requirements 15](#bookmark38)

[CHAPTER 3 16](#bookmark85)

SYSTEM DESIGN 16

1. [System Architecture 16](#bookmark87)
2. [Data Flow Diagram 17](#bookmark99)

CHAPTER 4 22

IMPLEMENTATION 22

v

1. [Technologies Used: 22](#bookmark109)
2. [Sample Code 38](#bookmark155)

[CHAPTER 5 42](#bookmark157)

TESTING 42

1. [About Testing: 42](#bookmark159)
2. [Types of Testing: 42](#bookmark162)
3. [Unit Testing: 42](#bookmark164)
4. [Integration Testing: 43](#bookmark171)
5. [Acceptance Testing 43](#bookmark173)

[CHAPTER 6 50](#bookmark201)

SCREENS 50

[CHAPTER 7 54](#bookmark205)

CONCLUSION & FUTURE ENHANCEMENT 54

1. [Conclusion: 54](#bookmark207)

[BIBILIOGRAPHY 55](#bookmark213)

**CHAPTER 1**

**INTRODUCTION**

Agriculture is one of the most important occupation practiced in our country. It is the broadest economic sector and plays an important role in overall development of the country. About 60% of the land in the country is used for agriculture in order to suffice the needs of 1.2 billion people. Thus, modernization of agriculture is very important and thus will lead the farmers of our country towards profit.

Data Analytic (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Earlier yield prediction was considering the farmers experience on a particular field and crop. However, as the conditions change day by day very rapidly, farmers are forced to cultivate more and more crops. Being this as the current situation, many of them don‘t have enough knowledge about the new crops and are not completely aware of the benefits they get while farming them. Also, the farm productivity can be increased by understanding and forecasting crop performance in a variety of environmental conditions. Thus, the proposed system takes the location of the user as an input. The processing part also take into consideration two or more datasets i.e. One obtained from weather department, forecasting the weather expected in current year and other data being static data.

This static data is the crop production and data related to demands of various crops obtained from various government websites.The proposed system applies machine learning and prediction algorithm like linear regression,decision tree, random forest to identify the patterns among the data in the process it as per input conditions. This in turn will propose the best feasible crops according to given environmental conditions. Thus, this system will only require the location of the user and it will suggest number of profitable crops providing a choice directly to the farmer about which crop to cultivate . As past year production is also taken into account, the prediction will be more accurate.

* 1. Objective of the Study:

The objective of the study is the prediction of the state crop production using historical annual data based on artificial intelligence methodologies such as linear regression. The extraction procedures/algorithms will produce the output by classification of the data according to the categories using Neuro-Fuzzy. The similar data will be grouped for the accurate and precise information that will predict crop yield more correctly and with perfect figures. The accurate and exact predictions will help in developing in order to make the agricultural sector more reliable and easy for the farmers by predicting the suitable crop by using Machine Learning techniques.

The data for the crop yield prediction is collected from Metrology Department of Kaggle. This is the annual data with all parameters of crops including states, crops, profit, yield. The aim of the proposed study is too effective and efficient in predicting which crop will be cultivated in which state with accuracy and precision.

* 1. Purpose and Scope:

Purpose:

This project mainly focused on the techniques and measures taken to improve farming by in calculating the technical knowledge and developments in order to make the agricultural sector more reliable and easy for the farmers by predicting the suitable crop by using Machine Learning techniques.

Based on the crop yield and profit farmers can cultivate the suitable crops for particular states. This will increase the crop production and beneficial to the farmers and country1 s economy.

Scope:

The analysis of Agriculture crop prediction variations through yield mapping form last 10 years data and predictions.

* 1. Existing System:

In Existing system, agriculture production is still lacking which could assist in lost in profit, soil exhaustion. Neighboring populations began compete for resources, such as good farmland and adequate water supplies , leading to conflict.

Disadvantages:

1. Lack of water resources- agriculture predominantly dependent on good monsoon
2. Paucity of electricity. Lack of farm mechanisation.
3. Avg.operational land holiding are fragmented and also reducing the over the years.
4. Low productivity of agriculture
5. No focused human resource training development initiatives.

1.4 Proposed System:

-Agriculture Crop Prediction II represents a mathematical method called Linear Regression to predict the Crop profit in various states in India. The Linear Regression method is modified in order to obtain the most optimum error percentage by iterating and adding some percentage of error to the input values. This method provides an estimate of annual profit using different crop yield and crop state to predict the profit. The linear regression is applied on the set of data and the coefficients are used to predict the profit based on the crop and yield. Thus, an estimate value of the annual profit at a given place can be predicted accurately.

Advantages :

1. Strengthening the agriculture economy.
2. Increases agriculture awareness.
3. Reduce the soil erosion.
4. Improves the soil structure.

**CHAPTER 2**

**SYSTEM ANALYSIS**

1. Problem Analysis:

Problem analysis is an important activity that takes place when we are building a new system or changing the existing one. Analysis helps to understand the existing system and the requirements necessary for building the new system. If there is no existing system then analysis defines the only requirements. One of the important factors in the system analysis is to understand the system and its problems. A good understanding of the system enables designer to identify and correct problems. Based on the drawbacks of the existing system the system is being planned, so the total definition of the problem has been analyzed

1. Types of Crops:

**Rice:**

Rice is the most important food crop of India covering about one-fourth of the total cropped area and providing food to about half of the Indian population.

This is the staple food of the people living in the eastern and the southern parts of the country, particularly in the areas having over 150 cm annual rainfall. There are about

10,000 varieties of rice in the world out of which about 4,000 are grown in India.



Wheat:

Wheat is a grass widely cultivated for its seed, a cereal grain which is a worldwide staple food. The many species of wheat together make up the genus Triticum, the most widely grown is common wheat. Wheat is grown on more land area than any other food crop (220.4 million hectares, 2014). World trade in wheat is greater than for all other crops combined. In 2017, world production of wheat was 772 million tonnes, with a forecast of 2019 production at 766 million tonnes, making it the second most-produced cereal after maize. Since 1960, world production of wheat and other grain crops has tripled and is expected to grow further through the middle of the 21st century. Global demand for wheat is increasing due to the unique viscoelastic and adhesive properties of gluten proteins, which facilitate the production of processed foods, whose consumption is increasing as a result of the worldwide industrialization process and the westernization of the diet.



**■Iowar ;**

Sorghum popularly known as jowar is the most important food and fodder crop of dry land agriculture. The cereal crop is perennial in nature and possessing corn like leaves and bearing the grain in a compact cluster. Sorghum is the fifth most important cereal crop in the world after wheat, rice, maize and barley. It is found in the arid and semi aridparts of the world, due to its feature of being extremely drought tolerant. The nutritional value of sorghum is same as of that of corn and that is why it is gaining importance as livestock feed. Sorghum is also used for ethanol production, producing grain alcohol, starch production, production of adhesives and paper other than being used as food and feed.

Jowar is one of the major staple food grain crops in India. Out of the total area under jowar cultivation in India, 50% is cultivated in Maharashtra. Whereas out of the total production of Jowar in the nation, 52% is from Maharashtra. Karnataka, Andhra Pradesh, TamilNadu.

One more advantage of this crop is that it can be grown in both Kharif and Rabi season. Also, it can handle and grow on a wide range of soil types starting from fertile to less nutrient soils but an effective output largely depends on soil moisture, resistance and porosity.



Mustard :

Mustard is the second most important and most prominent winter oilseed crop of India. It is grown mainly in the northern plains of India with some cultivated area in the eastern geography as well. It belongs to the group Cruciferae, with several cousin species cultivated. The others crops under the =Rapeseed & Mustard‘ category include Toria, Yellow Sarson, Brown Sarson, GobhiSarson or Canola and Black Mustard or Banarasi Rai. The small brown or yellow seeds contain up to 45 percent oil. The de-oiled cake is used as animal feed.



Sugarcane :

Sugarcane is a tropical, perennial grass that forms lateral shoots at the base to produce multiple stems, typically 3 to 4 m (10 to 13 ft) high and about 5 cm (2 in) in diameter. The stems grow into cane stalk which, when mature, constitutes around 75% of the entire plant. A mature stalk is typically composed of 11-16% fiber, 12-16% soluble sugars, 2-3% non-sugars, and 63-73% water. A sugarcane crop is sensitive to climate, soil type, irrigation, fertilizers, insects, disease control, varieties, and the harvest period. The average yield of cane stalk is 60-70 tonnes per hectare (24-28 long ton/acre; 27- 31 short ton/acre) per year. However, this figure can vary between 30 and 180 tonnes per hectare depending on knowledge and crop management approach used in sugarcane cultivation. Sugarcane is a cash crop, but it is also used as livestock fodder.



Types of Soils:

1. **Alluvial soil:**

Alluvial soil is rich in nutrients and may contain heavy metals. These soils are formed when streams and rivers slow their velocity. The suspended soil particles are too heavy for the decreasing current to carry and are deposited on the riverbed. The finest particles are deposited at the mouth of the river, forming a delta. Alluvial soils vary in mineral content and specific soil characteristics depending on the region and geologic makeup of the area.



The soil is porous because of its loamy (equal proportion of sand and clay) nature.

* The proportion of nitrogen is generally low.
* The proportion of Potash, phosphoric acid and alkalies are adequate
* The proportion of Iron oxide and lime vary within a wide range.

2 . Black soil:

This soil has high water holding capacity. So crops can be grown with less irrigation. It has high buffering and can hold nutrients in comparatively larger amount and for longer duration. This soil is very fertile and has been used traditionally for cotton cultivation leading to its name also.



Characteristics of black soil are-

• clayey texture and are highly fertile

• the structure of these soils is cloddish or occasionally friable

• rich in calcium carbonate, magnesium, potash, and lime but poor in nitrogen and phosphorous

* highly retentive of moisture, extremely compact and tenacious when wet
* contractible and develops deep wide cracks on drying
* calcareous and neutral to mild alkaline in reaction, high

in carbon exchange capacity and low in organic matter

* self-ploughing and comparatively less fertile on the uplands than

on the lowlands.

3 . Red soil:

Red soil is a type of soil that develops in a warm, temperate, moist climate under deciduous or mixed forest, having thin organic and organic-mineral layers overlying a yellowish-brown leached layer resting on an illuvium red layer. Red soils are generally derived from crystalline rock. They are usually poor growing soils, low in nutrients and humus and difficult to be cultivated because of its low water holding capacity.

Characteristics of red soil:



* Seen mainly in low rainfall area.
* Also known as Omnibus group.
* Porous, friable structure.
* Absence of lime, kankar (impure calcium carbonate).
* Deficient in: lime, phosphate, manganese, nitrogen, hummus, and potash.
* Colour: **Red** because of Ferric oxide.
* Texture: Sandy to **clay** and loamy.

1. Feasibility Study:

An analysis of the ability to complete a project successfully, taking into account legal, economic, technological, scheduling. Rather than just diving into a project and hoping for the best, a feasibility study allows project managers to investigate the possible negative and positive outcomes of a project.

* A feasibility study evaluates the project's potential for success.
* Feasibility studies aim to objectively and rationally uncover the strengths and weaknesses of an existing project.

Many Feasibility studies are disillusioning for both users and analysts. First, the study often presupposes that when feasibility document is being prepared, the analyst is in a position to evaluate solutions. Second, most studies tend to over look the confusion inherent in the system development the constraints and the assume attitudes. The feasibility study is to serve as a decision document for a project. Three key considerations involved in the feasibility analysis:

* Economical Feasibility
* Technical Feasibility
* Social Feasibility

1. **Economical Feasibility:**

The purpose of economic feasibility study is to demonstrate the net benefit of a proposed project for accepting or disbursing electronic fimds/benefits, taking into consideration the benefits and costs to the agency. The project -Agriculture crop productionll is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available.

1. Technical Feasibility:

Technical feasibility is concerned with specifying equipment and software that will successfully satisfy the user requirements.The technical needs of the system may vary considerably, but might include the facility to produce outputs in a given time, response time under certain conditions.

The project -Agriculture crop production -is technically feasible as the system developed has high demand on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

1. **Social Feasibility:**

The affect that a proposed project may have on the social system in the project environment is addressed in the social feasibility. It may happen that a particular category of employees may be short or not available as a result of ambient social structure. The influence on the social status of the participants by the project should be evaluated in order to guarantee compatibility. It must be identified that employees in the particular industries may have specific status symbols within the society. The aspect of study is to check the level of acceptance of the system by the user.

The project -Agriculture crop production- includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

1. Requirement Specification:

A **Software Requirements Specification (SRS) -** A requirements specification for a software system - is a complete description of the behavior of a system to be developed. It includes a set of use cases that describe all the interactions the users will have with the software. In addition to use cases, the SRS also contains non-functional requirements. Non­functional requirements are requirements which impose constraints on the design or implementation (such as performance engineering requirements, quality standards, or design constraints).

**System requirements specification:** A structured collection of information that embodies the requirements of a system. A business analyst, sometimes titled system analyst, is responsible for analyzing the business needs of their clients and stake holders to help identify business problems and propose solutions. Within the systems development life cycle domain, the BA typically performs a liaison function between the business side of an enterprise and the information technology department or external service providers. Projects are subject to three sorts of requirements:

* **Business requirements** describe in business terms what must be delivered or accomplished to provide value.
* **Product requirements** describe properties of a system or product (which could be one of several ways to accomplish a set of business requirements.)
* **Process requirements** describe activities performed by the developing organization. For instance, process requirements could specify specific methodologies that must be followed, and constraints that the organization must obey.

Product and process requirements are closely linked. Process requirements often specify the activities that will be performed to satisfy a product requirement. For example, maximum development cost requirement (a process requirement) may be imposed to help achieve a maximum sales price requirement (a product requirement); a requirement that the product be maintainable (a Product requirement) often is addressed by imposing requirements to follow particular development style.

A systems engineering, a requirement can be a description of what a system must do, referred to as a functional requirement. This type of requirement specifies something that the delivered system must be able to do. Another type of requirement specifies something about the system itself, and how well it performs its functions. Such requirements are often called Non-fimctional requirements, or 'performance requirements' or 'quality of service requirements. Examples of such requirements include usability, availability, reliability, supportability, testability and maintainability.

A collection of requirements define the characteristics or features of the desired system. A 'good' list of requirements as far as possible avoids saying how the system should implement the requirements, leaving such decisions to the system designer. Specifying how the system should be implemented is called "implementation bias" or "solution engineering". However, implementation constraints on the solution may validly be expressed by the future owner, for example for required interfaces to external systems; for interoperability with other systems; and for commonality (e.g. of user interfaces) with other owned products. In Software Engineering, the same meanings of requirements apply, except that the focus of interest is the software itself.

1. Functional Requirements:

Functional requirements are very important system requirements in the system design process. These requirements are the technical specifications, system design parameters and guidelines, data manipulation, data processing, and calculation modules etc, of the proposed system. Functional requirements are in contrast to Non-Functional requirements which are descripitive of the parameters of system performance, quality attributes, reliability and security, cost, constraints in design/implementation, etc

* Enter subdivision
* Enter month values
* The predicted value is displayed to the user.

1. Non- Functional Requirements:

Non-functional requirements tend to be stated in terms of constraints on the results of tasks which are given as functional requirements (e.g.., constraints on the speed or efficiency of a given task), a task based functional requirements statement is a useful skeleton upon which to construct a complete requirements statement. That is the approach taken in this work. It can be helpful to think of non-functional requirements as adverbially related to tasks or functional requirements.

Non-functional requirements are often called qualities of a system. Other terms for non-functional requirements are -constraintsII, -quality attributesII, -quality goals II, -quality of service requirementsII and -non-behavioral requirementsII.

Accessibility:

Accessibility is a general term used to describe the degree to which a product, device, service, or environment is available to as many people as possible. Accessibility can be viewed as the -ability to access II and benefit from some system orentity. Accessibility is often used to focus on people with disabilities or special needs and their right of access to entities, often to use of assistive technology.

This project is accessible for understanding of the crop and ability to simulate yield are constantly evolving at the same time, the resources of those using yield are increasing and becoming more sophisticated We are improving confidence in predictions, particularly in profit as result of improved understanding of crop and yield.

Availability:

Availability is the degree to which a system, sub-system, or equipment is in a specified operable and committable state at the start of a mission, when a mission is called for at an unknown, i.e., a random, time. Simply put, availability is the proportion of time a system is in a functioning condition. This is often described as a mission capable rate; it is available in this project because equipments used in this one are operable.

With the growing availability of crop-related big data, researchers have realized that introducing data-driven approaches to profit prediction can achieve considerable success. Recent advancements in machine learning combined with the growing availability of crop- related data has served to dramatically improve the accuracy of profit.

**Scalability:**

Scalability is the ability of a system, network, or process, to handle growing amount of work in a capable manner or its ability to be enlarged to accommodate that growth. For example, it can refer to the capability of a system to increase total throughput under an increased load when recourses (typically hardware) are added. An analogous meaning is implied when the word is used in a commercial context, where scalability of a company implies that the underlying business model offers the potential for economic growth with the company. -Agriculture crop predictionII models in order to emphasize the close liaison between the applications and the hardware used to run these models moreover, the agriculture prediction to provide the ways to turn the applications to achieve the best possible speed up and best utilization of the processing involved.

Portability:

Portability in high-level computer programming is the usability of the same software in different environments. The pre requirement for portability is the generalized abstraction between the application logic and system interfaces. It is portable because it runs on different environment. Agriculture crop production contributes by exploring the Support Vector Machine (SVM), k Nearest Neighbor (KNN), Decision Tree (J48) and Multilayer Perceptron (MLP). The dataset is obtained from a weather forecasting website

and consists of several atmospheric attributes. For effective prediction, pre-processing technique is used which consists of cleaning and normalization processes. Performance of used data mining techniques is analyzed in terms of precision, recall and f-measure with various ratios of training and test data. Robustness: Robustness is the ability of a computer system to cope with errors during execution or the ability of an algorithm to continue to operate despite abnormalities in input , calculations, etc. Formaltechniques, such as fuzz testing, or essential to showing robustness since this type of testing involves invalid or unexpected inputs. Various commercial products perform robustness testing of software systems. Precipitation is characterized by substantial natural variability, including on regional and decadal scales. This relatively large variability poses a grand challenge in assessing the significance of profit. Here we use multiple large ensembles of crops experiments to evaluate whether, on regional scales, yields are distinguishable.

1. System Specification:
2. Hardware Requirements:

Processor

RAM

Secondary Storage

Key Board

Mouse

1. Software Requirements:

Operating System

Scripting languages

Python Packages

Intel Pentium Inside

minimum 1 GB

minimum 160 GB

Standard Windows Keyboard

Two or Three Button Mouse

Windows XP/Windows 7/8/10

Python

Numpy, Pandas, Matplotlib

CHAPTER 3

**SYSTEM DESIGN**

1. System Architecture:

The overall system design consists of following modules:

Data Collection:

The dataset used in this project was an open source dataset from Kagglelnc. It consists of 4000 records with 19 parameters that have the possibility of affecting the property prices. However out of these 19 parameters only 4 were chosen which are bound to affect the annual rainfall. Parameters such as Jan-Feb, Mar-May, are independent varibles. Annual is a dependent variable on several other independent variables.

Preprocessing:

It is a process of transforming the raw, complex data into systematic understandable knowledge. It involves the process of finding out missing and redundant data in the dataset. Entire dataset is checked for NaN and whichever observation consists of NaN will be deleted. Thus, this brings uniformity in the dataset. However in our dataset, there was no missing values found meaning that every record was constituted its corresponding feature values.

Data Classification:

Classification is the problem of identifying to which of a set of categories (subpopulations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

Data regression:

Regression is basically a statistical approach to find the relationship between variables. In machine learning, this is used to predict the outcome of an event based on the relationship between variables obtained from the data-set. Linear regression is one type regression used in Machine Learning.

**Prediction of Output:** Output can be predicted by using Machine Learning algorithms.



I/O

**Inpul  
Footon**

Data is collected and stored in NoSQL / SQL format. That data is divided into two parts i) Training data ii) Testing data. Training data is used for training the model and then that model is tested using testing data. After this, the trained model is used for predicting annual rainfall given feature set.

1. Data Flow Diagram:

Data flow diagrams provide a graphical representation of how information moves between processes in a system. Data flow diagrams follow a hierarchy; that is, a diagram may consist of several layers, each unique to a specific process or data function.

Symbols Used in DFD’s:

|  |  |  |  |
| --- | --- | --- | --- |
| **Symbol** | | | **Meaning** |
| **Process** | | | **Single process: xA. circle is used to represent the entire system.** |
|  | | | **Data flow: ±An arrow is used to represent the flow of data between the process and external entities.** |
|  |  | **External**  **entity** | **External entity: A square or rectangle represents any person or organisation that sends data to or receives data fi'om**  **the system.** |
|  | **Data store** | | **Data store: An open rectangle represents**  **the location where data is stored. It could be a filing cabinet, hard disk.** |

**LEVEL 0 DIAGRAM:**

Machine learning workflow

**Import**

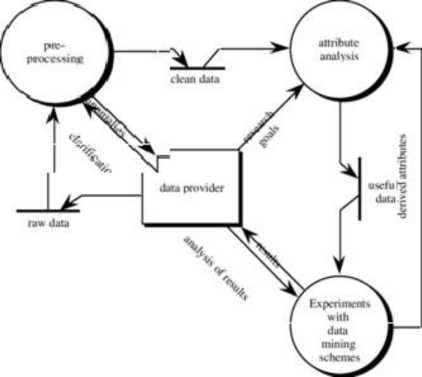
**Process**

**Visualize**

**Model**

**Evaluate**

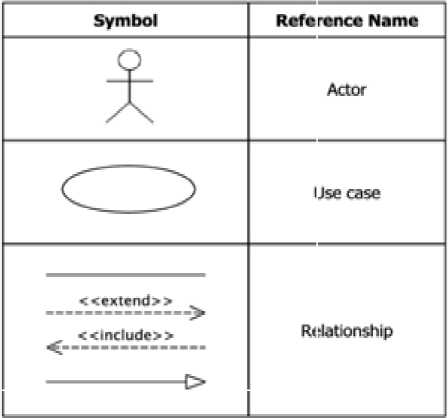
**LEVEL 1 DIAGRAM:**



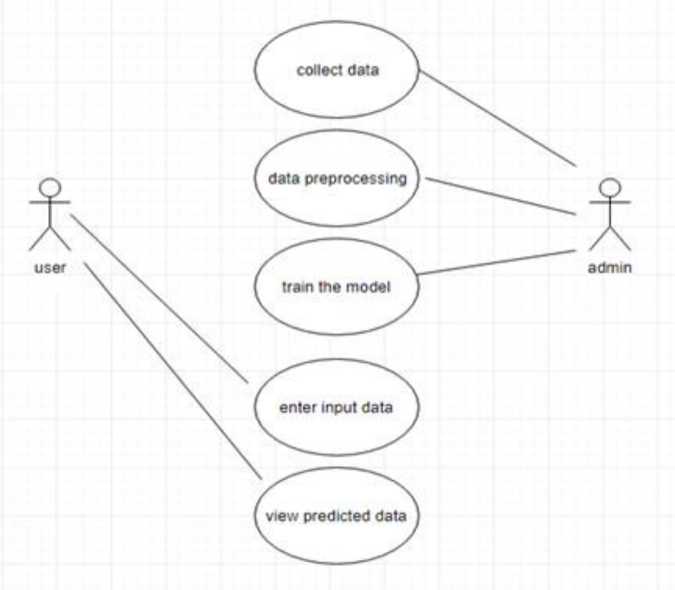
**Usecase Digram:**

A use case diagram is a dynamic or behavior diagram in UML. Use case diagrams model the functionality of a system using actors and use cases. Use cases are a set of actions, services, and functions that the system needs to perform. In this context, a "system" is something being developed or operated, such as a web site. The "actors" are people or entities operating under defined roles within the system.

SYMBOLS:



They provide a good high level analysis from outside the system. Use case diagrams specify how the system interacts with actors without worrying about the details of how that functionality is implemented.



Sequence Diagram:

A sequence diagram shows object interactions arranged in time sequence.

It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario. Sequence diagrams are typically associated with use case realizations in the logical view of the system under development. Sequence diagrams are sometimes called event diagrams or event scenarios.



\*

I

Message CaTTF?

Activation Bar

MessageL^^

Receiver

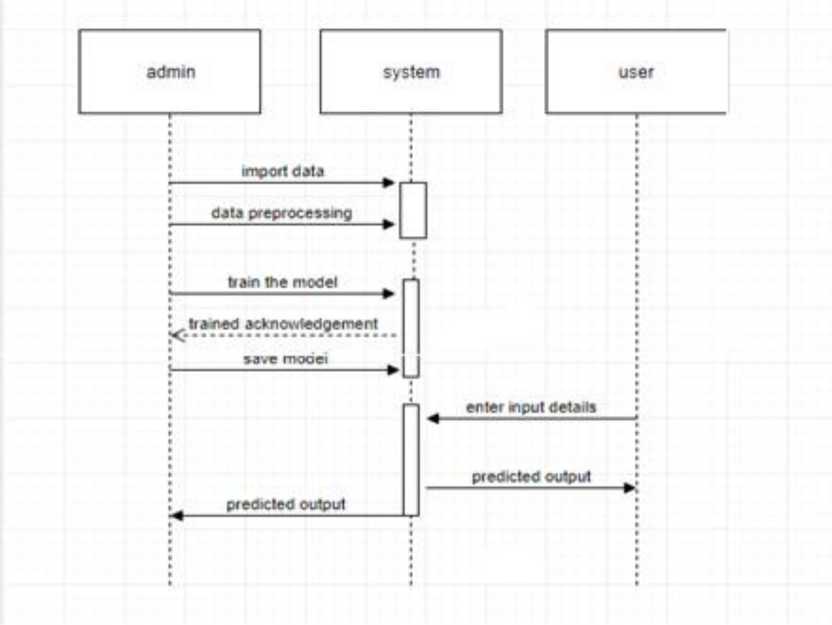
Activation Bar I

Object

Fig . Block diagram for sequence diagram

A sequence diagram shows, as parallel vertical lines (lifelines), different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur. This allows the specification of simple runtime scenarios in a graphical manner.

Fig: Sequence Diagram



**CHAPTER 4**

**IMPLEMENTATION**

1. Technologies Used:

**PYTHON:**

Python is a popular platform used for research and development of production systems. It is a vast language with number of modules, packages and libraries that provides multiple ways of achieving a task.

Python and its libraries like NumPy, SciPy, Scikit-Learn, and Matplotlib are used in data science and data analysis. They are also extensively used for creating scalable machine learning algorithms. Python implements popular machine learning techniques such as Classification, Regression, Recommendation, and Clustering.

Libraries and Packages:

To understand machine learning, you need to have basic knowledge of Python programming. In addition, there are a number of libraries and packages generally used in performing various machine learning tasks as listed below:

* Numpy - is used for its N-dimensional array objects
* Pandas - is a data analysis library that includes data frames
* Matplotlib - is 2D plotting library for creating graphs and plots
* Scikit-learn - the algorithms used for data analysis and data mining tasks
* Seaborn - a data visualization library based on matplotlib

Installation Steps Involved in Machine Learning:

A machine learning project involves the following steps:

* Defining a Problem
* Preparing Data
* Evaluating Algorithms
* Improving Results
* Presenting Results

The best way to get started using Python for machine learning is to work through a project end-to-end and cover the key steps like loading data, summarizing data, evaluating algorithms and making some predictions. This gives you a replicable method that can be used dataset after dataset. You can also add further data and improve the results.

Installation:

You can install software for machine learning in any of the two methods as discussed here:

Method 1:

Download and install Python separately from python.org on various operating systems as explained below:

To install Python after downloading, double click the .exe (for Windows) or .pkg (for Mac) file and follow the instructions on the screen.

For Linux OS, check if Python is already installed by using the following command at the prompt:

$ python --version....

If Python 3.7 or later is not installed, install Python with the distribution's package manager. Note that the command and package name varies.

On Debian derivatives such as Ubuntu, you can use apt:

**$ sudo apt-get install python3**

Now, open the command prompt and run the following command to verify that Python is installed correctly:

$ python3 —version

Python 3.6.2

Similarly, we can download and install necessary libraries like numpy, matplotlib etc. individually using installers like pip. For this purpose, you can use the commands shown here: $pip install numpy

$pip install matplotlib

$pip install pandas

$pip install seaborn

Method 2:

Alternatively, to install Python and other scientific computing and machine learning packages simultaneously, we should install Anaconda distribution. It is a Python implementation for Linux, Windows and OSX, and comprises various machine learning packages like numpy, scikit-learn, and matplotlib. It also includes Jupyter Notebook, an interactive Python environment. We can install Python 2.7 or any 3.x version as per our requirement.

To download the free Anaconda Python distribution from Continuum Analytics, you can do the following:

Visit the official site of Continuum Analytics and its download page. Note that the installation process may take 15-20 minutes as the installer contains Python, associated packages, a code editor, and some other files. Depending on your operating system, choose the installation process as explained here:

For Windows:

Select the Anaconda for Windows section and look in the column with Python 2.7 or 3.x. You can find that there are two versions of the installer, one for 32-bit Windows, and one for 64-bit Windows. Choose the relevant one.

**For Mac OS:** Scroll to the Anaconda for OS X section. Look in the column with Python 2.7 or 3.x. Note that here there is only one version of the installer: the 64-bit version. **For Linux OS:** We select the "Anaconda for Linux" section. Look in the column with Python 2.7 or 3.x.

Note that you have to ensure that Anaconda‘s Python distribution installs into a single directory, and does not affect other Python installations, if any, on your system.

To work with graphs and plots, we will need these Python library packages: **matplotlib and seaborn.**

If you are using Anaconda Python, your system already has numpy, matplotlib, pandas, seaborn, etc. installed. We start the Anaconda Navigator to access either **Jupyter Note book** or Spyder **IDE** of python.

After opening either of them, type the following commands:

import numpy

import matplotlib

Now, we need to check if installation is successful. For this, go to the command line and type in the following command:

$ python

Python 3.6.3 (Anaconda custom (32-bit)| (default, Oct 13 2017,14:21:34) [GCC 7.2.0] on linux

Next, you can import the required libraries and print their versions as shown:

»>import numpy

»>print numpy. version

1.14.2

»> import matplotlib

»> print (matplotlib.\_version\_)

2.1.2

»import pandas

»> print (pandas.\_version\_)

0.22.0

»> import seaborn

»> print (seaborn.\_version\_)

Machine Learning:

Machine learning is a discipline that deals with programming the systems so as to make them automatically learn and improve with experience. Here, learning implies recognizing and understanding the input data and taking informed decisions based on the supplied data. It is very difficult to consider all the decisions based on all possible inputs. To solve this problem, algorithms are developed that build knowledge from a specific data and past experience by applying the principles of statistical science, probability, logic, mathematical optimization, reinforcement learning, and control theory.

Machine learning is associated with the study of the algorithms that enhance the efficiency of the machines/computers automatically through the training and testing of the machine/computers with certainly different variables. The machine learning is among the most favourable and fastest growing areas of computer technology. The computers work efficiently with different algorithms and functions. The machine learning is the training the computer with certainly different algorithms to experience the machine in automatic smart data processing. The machine learning enhances the efficiency and accuracy of the data processing and is used in a wide range of fields. The machine learning is developed with effectual algorithms that utilize a certain set of tools and functions to solve the complex and huge data.

The input to a learning algorithm is training data, representing experience, and the output is any expertise, which usually takes the form of another algorithm that can perform a task. The input data to a machine learning system can be numerical, textual, audio, visual, or multimedia. The corresponding output data of the system can be a floating-point number, for instance, the velocity of a rocket, an integer representing a category or a class, for example, a pigeon or a sunflower from image recognition. In this chapter, we will learn about the training data our programs will access and how learning process is automated and how the success and performance of such machine learning algorithms is evaluated.

Applications of Machine Learning Algorithms the developed machine learning algorithms are used in various applications such as:

* Vision processing
* Language processing
* Forecasting things like stock market trends, weather
* Pattern recognition
* Games
* Data mining
* Expert systems
* Robotics

Steps Involved in Machine Learning:

A machine learning project involves the following steps:

* Defining a Problem
* Preparing Data
* Evaluating Algorithms
* Improving Results
* Presenting Results

The best way to get started using Python for machine learning is to work through a project end-to-end and cover the key steps like loading data, summarizing data, evaluating algorithms and making some predictions. This gives you a replicable method that can be used dataset after dataset. You can also add further data and improve the results.

Data Analysis:

During the last decades, there has been an incredible growth in our capabilities of generating and storing data. In general, there is a competitive edge in being able to properly use the abundance of data that is being collected in industry and society today. Efficient analysis of collected data can provide significant increases in productivity through better business and production process understanding and highly useful applications for e. g. decision support, surveillance and diagnosis.

The purpose of data analysis is to extract answers and useful patterns such as regularities and rules in data. These patterns can then be exploited in making predictions, diagnoses, classifications etc. Typical examples of working industrial and commercial applications are

* Virtual sensors, i.e. an indirect measurement of values computed from values that are easier to access.
* Predictive maintenance and weak point analysis through e. g. maintenance and warranty databases.
* Incremental step-wise diagnosis of equipment such as car engines or process plants.
* Intelligent alarm filtering and prioritisation of information to operators of complex systems.
* Fraud and fault detection in e. g. data communication systems and Business.
* Sales and demand prediction, e. g. in power grids or retail.
* Speed-up through model approximation in control systems, e. g. replacing a slower simulator with a faster learning system approximation.
* Clustering and classification of customers, e.g for targeted pricing and advertising, and identification of churners, i.e. customers likely to change provider.

With all data analysis and machine learning related applications running within industry, government, and homes, it is very hard to argue that the fields have not produced successful real world applications. However, there is still a definite gap between the development of advanced data analysis and machine learning techniques and their deployment in actual applications. There are several reasons for this.

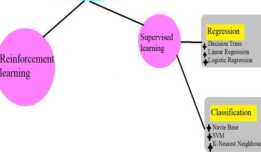
Adapting and applying theoretical machine learning models to practical problems can be very difficult. Although it is often possible to achieve fair performance with a standard model formulation, we usually need a quite high degree of specialisation to achieve good performance and to satisfy constraints on e. g. computational complexity. Even if this is not necessary in certain situations, we usually still have to at least specify some model parameters or structure.

Understanding and preparing data for testing, validation and the actual application can be immensely time consuming. The data analyst trying to understand the data and the problem to be modelled is often not an expert in the application area, making acquisition of expert knowledge an important and time consuming task. Real-world data is also often notoriously dirty. It contains encoding errors (e. g. from errors during manual input) andambiguities, severe levels of noise and outliers, and large numbers of irrelevant or redundant attributes.

The best way to get started using Python for machine learning is to work through a project end-to-end and cover the key steps like loading data, summarizing data, evaluating algorithms and making some predictions. This gives you a replicable method that can be used dataset after dataset. You can also add further data and improve the results.

Broadly, there are 3 types of Machine Learning Algorithms

lachine Learning Algorithms



AnahsistLDA)

Machine

Learning

1. **Supervised Learning:**

This algorithm consists of a target / outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables). Using these set of variables, we generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data.

Supervised learning is commonly used in real world applications, such as face and speech recognition, products or movie recommendations, and sales forecasting. Supervised learning can be further classified into two types:

Regression and Classification.

* Regression trains on and predicts a continuous-valued response, for example predicting real estate prices.
* Classification attempts to find the appropriate class label, such as analyzing positive/negative sentiment, male and female persons, benign and malignant tumors, secure and unsecure loans etc.

**Examples of Supervised Learning:** Regression, Decision Tree, Random Forest, KNN, ANN, Logistic Regression etc.

1. Decision Tree :

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

1. The leaves are the decisions or the final outcomes.
2. The decision nodes are where the data is split.

Classification trees (yes / no types ):

What we‘ve seen above is an example of classification tree, where the outcome was a variable like =fit‘ or =unfit‘. Here the decision variable is Categorical.

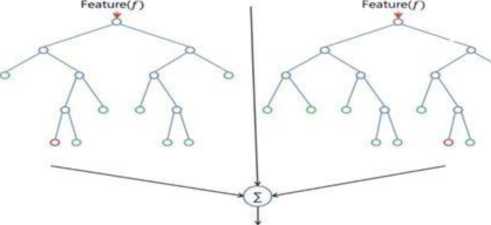
Regression Trees (continuous types ):

Here the decision or the outcome variable is Continuous, e.g. a number like 123 **Algorithm Steps :**

1. Create root node for the tree
2. If all examples are positive, return leaf node =positive‘
3. Else if all examples are negative, return leaf node =negative‘
4. Calculate the entropy of current state H(S)
5. For each attribute, calculate the entropy with respect to the attribute =x‘ denoted by H(S, x)
6. Select the attribute which has maximum value of IG(S, x)
7. Remove the attribute that offers highest IG from the set of attributes
8. Repeat until we run out of all attributes, or the decision tree has all leaf nodes.
9. Random Forest:

Random Forest is a supervised learning algorithm. Like you can already see from its name, it creates a forest and makes it somehow random. The -forestII it builds, is an ensemble of Decision Trees, most of the time trained with the -baggingll method. The general idea of the bagging method is that a combination of learning models increases the overall result. To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is, that it can be used for both classification and regression problems, which form the majority of current machine learning systems. We will talk about random forest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:



The great quality of the random forest algorithm is that it is very easy to measure the relative importance of each feature on the prediction. Sklearn provides a great tool for this, that measures a features importance by looking at how much the tree nodes, which use that feature, reduce impurity across all trees in the forest. It computes this score automatically for each feature after training and scales the results, so that the sum of all importance is equal to 1.

1. Unsupervised Learning:

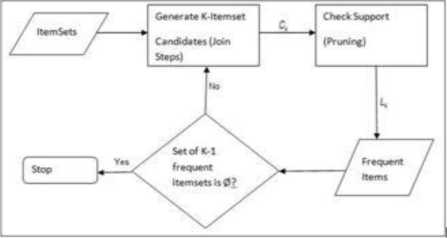
In this algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention.

Unsupervised learning algorithms are extremely powerful tools for analyzing data and for identifying patterns and trends. They are most commonly used for clustering similar input into logical groups. Unsupervised learning algorithms include K-means, Random Forests, and Hierarchical clustering and so on.

**Examples of Unsupervised Learning:** Apriori algorithm, K-means.

1. Apriori Algorithm:

The Apriori algorithm was proposed by Agrawal and Srikant in 1994. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers). Other algorithms are designed for finding association rules in data having no transactions. Each transaction is seen as a set of items (an itemset). Given a threshold, the Apriori algorithm identifies the item sets which are subsets of at least transactions in the database.



Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length from item sets of length .Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

The pseudo code for the algorithm is given below for a transaction database and a support threshold of Usual set theoretic notation is employed, though note that is a multiset is the candidate set for level. At each step, the algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. Accesses a field of the data structure that represents candidate set, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies.

procedureApriori (T, minSupport) / //?' is the database and min-Support is the minimum support

Ck: Candidate itemset of size k

Lk: frequent itemset of size k

Ll= {frequent items/;

forlk= I; Lk !=0: k++)do begin

1= candidates generated from Lk; for each transaction r in database do{

increment the count of all candidates in Gc+i that are contained in t

candidates in with min\_support

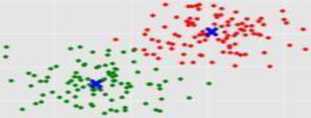
/end

return lAL\*:

K-Means Algorithm:

K-Means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-Means clustering aims to partition =n‘ observations into =k‘ clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into voronoicells. The problem is computationally difficult however, efficient heuristic algorithms converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of gaussian distributions via an iterative refinement approach employed by both k-means and gaussian mixture modeling. They both use cluster centers to model the data; however, k- means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

The algorithm has a loose relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with K- Means due to the name. Applying the 1-nearest neighbor classifier to the cluster centers obtained by K- Means classifies new data into the existing clusters. This is known as nearest centroid classifier or rocchio algorithm.



1. Reinforcement Learning:

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.

**Example of Reinforcement Learning:** Markov Decision Process Similarly, there are four categories of machine learning algorithms as shown below:

* Supervised learning algorithm
* Unsupervised learning algorithm
* Semi-supervised learning algorithm
* Reinforcement learning algorithm

However, the most commonly used ones are supervised and unsupervised learning.

Purpose of Machine Learning:

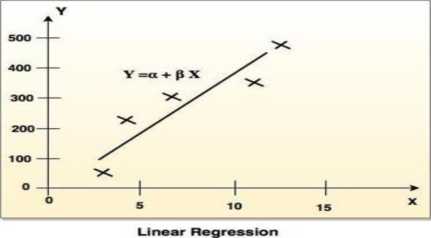
Machine learning can be seen as a branch of Al or Artificial Intelligence, since, the ability to change experience into expertise or to detect patterns in complex data is a mark of human or animal intelligence. As a field of science, machine learning shares common concepts with other disciplines such as statistics, information theory, game theory, and optimization. As a subfield of information technology, its objective is to program machines so that they will learn. However, it is to be seen that, the purpose of machine learning is not building an automated duplication of intelligent behavior, but using the power of computers to complement and supplement human intelligence. For example, machine learning programs can scan and process huge databases detecting patterns that are beyond the scope of human perception.

Algorithm:

**Linear Regression:**

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a \*X + b.

Before knowing what linear regression is, let us get ourselves accustomed to regression. Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.



Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

y = a\_0 + a\_l \* x

* Y - Dependent Variable
* a - Slope
* X - Independent variable

The motive of the linear regression algorithm is to find the best values for a\_0 and a l. Before moving on to the algorithm, let‘s have a look at two important concepts you must know to better understand linear regression.

**Cost Function:**

The cost function helps us to figure out the best possible values for a\_0 and a l which would provide the best fit line for the data points. Since we want the best values for a\_0 and a l, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

We choose the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error (MSE) function. Now, using this MSE function we are going to change the values of a\_0 and a l such that the MSE value settles at the minima. **Gradient Descent:**

The next important concept needed to understand linear regression is gradient descent. Gradient descent is a method of updating a\_0 and a l to reduce the cost function (MSE). The idea is that we start with some values for a\_0 and a l and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.

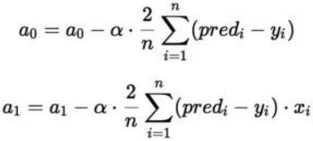


Big learning rate

Small learning rate

Fig. Gradient Descent

You may be wondering how to use gradient descent to update a\_0 and a l. To update a\_0 and a l, we take gradients from the cost function. To find these gradients, we take partial derivatives with respect to a\_0 and a l. Now, to understand how the partial derivatives are found below you would require some calculus but if you don‘t, it is alright. You can take it as it is.



The partial derivates are the gradients and they are used to update the values of a\_0 and a l. Alpha is the learning rate which is a hyper parameter that you must specify. A smaller learning rate could get you closer to the minima but takes more time to reach the minima, a larger learning rate converges sooner but there is a chance that you could overshoot the minima.

The best way to understand linear regression is to relive this experience of childhood. Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights! What do you think the child will do? He/She would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life! The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation above. In the equation, these coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

Algorithm for linear regression:

SUB Regress(x, y, n, al, aO, syx, r2)

sumx = 0: sumxy = 0: st= 0

sumy = 0: sumx2 =0: sr = 0

DO i= 1, n

sumx = sumx + xi

sumy = sumy + yi

sumxy = sumxy + xi\*yi

sumx2 = sumx2 + xi\*xi

**END DO**

xm = sumx/n

ym = sumy/n

al = (n\*sumxy - sumx\*sumy)/(n\*sumx2 - sumx\*sumx) aO = ym - al\*xm

**DO** i= 1, n

st = st + (yi - ym)A2

sr = sr + (yi - al \* xi - aO)A2

**END DO**

syx = (sr/(n - 2))A0.5

r2 = (st - sr)/st

END Regress

1. Sample Code

import os

import pandas as pd

import numpy as np

os.getcwd()

os. chdir('C :\\Users\\HPWDownloads')

os.getcwd()

df = pd. read\_csv(" cost-of-cultivation 1. csv") df

df. state. value\_counts()

import seaborn as sns

import matplotlib.pyplot as pit

colormap=plt. cm. viridis

pit. figure(figsize=( 12,12))

plt.title("person correlation coefficient of features",y=l.05,size=l 5)

sns.heatmap(df.corr(),linewidth=0.1,vmax=1.0,square=True,cmap=colormap,linecolor=" whit e", annot=True)

x=df.iloc[:,:-l].values

x

y=df.iloc[:,2], values

y

x=df. iloc[: ,3 ] .values

x

pit. scatter(x,y,label-'",color="b",s=100)

plt.xlabel("yield")

plt.ylabel("profit")

plt.title("yield vs profit")

plt.legend("yield")

plt.show()

x=df. iloc[: ,3 ] .values

pit. scatter(x,y,label-'",color="b",s=100)

plt.xlabel("yield")

plt.ylabel("profit")

plt.title("yield vs profit")

plt.legend("yield")

YI=pd.DataFrame({ "yield" :x, "Profit" :y})

sns.lmplot(x="yield",y="Profit",markers=["X"],data=YI)

import seaborn as sns

sns. version

sns. scatterplot(data=df,x="yield",y="profit",hue=" state")

df.state.value\_counts().sort\_values(ascending=False)[0:50].plot.bar()

df{"yield"].sort\_values(ascending=False)[0:80].plot.bar()

df{"yield"].sort\_values(ascending=False)[30:80].plot.bar()

sns.boxplot(data=df,x="state",y="cost")

sns.boxplot(data=df,x="state",y="profit")

sns.boxplot(data=df,x="state",y="yield")

sns.boxplot(data=df,x="yield",y="profit")

df.boxplot(column="profit",by="crop")

df.boxplot(column="profit",by="state")

df.boxplot(column="yield",by="state")

df.boxplot(column="cost",by="state")

sns.distplot(df{"yield"])

sns.distplot(df{"cost"])

df\_dummy=pd.get\_dummies(df)

dfdummy

Y=df\_dummy. profit. values

Y

dfdummy. drop( "profit", axis= 1, inplace=T rue)

X=df\_dummy. value s

X

from sklearn. modelselection import train test split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,Y,test\_size=0.2,random\_state=0) from sklearn.preprocessing import StandardScaler

scaler=StandardScaler().fit(X)

scaler

X

X=scaler.transform(X)

X

from sklearn. model selection import train test split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,Y,test\_size=0.2,random\_state=0)

from sklearn. linear model import LinearRegression

linear=LinearRegression()

linear. fit(X,Y)

linear. score(X,Y)

linear. score(x\_train,y\_train)

linear. score(x\_test,y\_test)

from sklearn.ensemble import RandomForestRegressor

RF=RandomForestRegressor(n\_estimators=100,oob\_score=True)

RF=RandomForestRegressor(max\_depth=20)

RF. fit (xtrain, ytrain)

RF.score(X,Y)

from sklearn.tree import DecisionTreeRegressor

DT=DecisionTreeRegressor()

from sklearn.model selection import train test split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(X,Y,test\_size=0.2,random\_state=0)

DT. fit(x\_train,y\_train)

DT.score(X,Y)

y\_pred

y\_test

pit. scatter(y\_test,y\_pred)

profit=pd.DataFrame({ "y\_test" :y\_test,"y\_pred" :y\_pred})

sns.lmplot(x="y\_test",y="y\_pred",markers=["x"],data=profit)

from math import sqrt

from sklearn.metrics import mean squared error

rmse=sqrt(mean\_squared\_error(y\_test,y\_pred))

rmse

import statsmodels.formula.api as sm

regressor\_OLS=sm.OLS(endog=Y,exog=X).fit()

regressor\_OLS.summary()

CHAPTER 5

**TESTING**

1. About Testing:

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

1. Types of Testing:

Types of testings are:

* Unit testing
* Integration testing
* Acceptance testing

1. Unit Testing:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases. Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration.

This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results

**Test Strategy and Approach:** Field testing will be performed manually and functional tests will be written in detail.

Test Objectives:

* All field entries must work properly.
* Pages must be activated from the identified link.
* The entry screen, messages and responses must not be delayed.

Features To Be Tested:

* Verify that the entries are of the correct format

•No duplicate entries should be allowed

* Alllinks should take the user to the correct page.

1. Integration Testing:

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent.Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications, e.g. components in a software system or - one step up - software applications at the company level - interact without error.

**Test Results** : All the test cases mentioned above passed successfully. No defects encountered.

1. Acceptance Testing:

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

**Test Results:** All the test cases mentioned above passed successfully. No defects

encounter.

**Blackbox Testing Techniques for Machine Learning Models:**

The following represents some of the techniques which could be used to perform blackbox testing on Machine Learning models:

Model performance

* Metamorphic testing
* Dual coding
* Coverage guided fuzzing
* Comparison with simplified, linear models
* Testing with different data slices

1. Model Performance;

Testing model performance is about testing the models with the test data/new data sets and comparing the model performance in terms of parameters such as accuracy/recall etc., to that of pre-determined accuracy with the model already built and moved into production. This is the most trivial of different techniques which could be used for blackbox testing.

Metamorphic Testing:

In metamorphic testing, one or more properties are identified that represent the metamorphic relationship between input-output pairs. For example, hypothetically speaking, an ML model is built that predicts the likelihood of a person suffering from a particular disease based on different predictor variables such as age, smoking habit, gender, exercise habits, etc. Based on the detailed analysis, it is derived that given the person is a smoker and a male, the likelihood of the person suffering from the disease increases by 5% with an increase in his age by 3 years. This could be used to perform metamorphic testing as the property, age, represents the metamorphic relationship between inputs and outputs.

In metamorphic testing, the test cases that result in success lead to another set of test cases which could be used for further testing of Machine Learning models. The following represents a sample test plan:

* Given the person is a male and a smoker, determine the likelihood of the person suffering from the disease when his age is 30 years.
* Increase the age by 5 years. The likelihood should increase by more than 5%.
* Increase the age by 10 years. The likelihood should increase by more than 15% but less than 20%.

Test cases such as above can be executed until all results in success or failure at any step. In case, one of the test cases fail, it could result in the logging of a defect which could be dealt with, by data scientists.

Dual Coding:

With dual coding technique, the idea is to build different models based on different algorithms and comparing the prediction from each of these models given a particular input data set. Let's day, a classification model is built with different algorithms such as random forest, SVM, neural network. All of them demonstrate a comparative accuracy of 90% or so with random forest showing the accuracy of 94%. This results in the selection of random forest. However, during testing, the model for quality control checks, all of the above models are preserved and input is fed into all of the models. For inputs where the majority of remaining models other than random forest gives a prediction which does not match with that of the model built with random forest, a bug/defect could be raised in the defect tracking system. These bugs could later be prioritized and dealt with by data scientists.

Coverage Guided Fuzzing:

Coverage guided fuzzing is a technique where data to be fed into the Machine Learning models could be planned appropriately such that all of the features activations get tested. Take for an instance, the models built with neural networks, decision trees, random forest etc. Let's say the model is built using neural networks. The idea is to come up with data sets (test cases) which could result in the activation of each of the neurons present in the neural network. This technique sounds more like a white-box testing. However, the way it becomes part of the blackbox testing is the feedback which is obtained from the model which is then used to guide the further fuzzing and hence, the name — Coverage guided fuzzing. This is a work in progress.

White Box Testing:

**Metrics of Machine Learning:**

* Classification accuracy
* Logarithmic Loss
* Confusion Matrix
* Area under curve
* Fl Score
* Mean Absolute error
* Mean Squared Error

Classification accuracy:

Classification Accuracy is what we usually mean, when we use the term accuracy.

It is the ratio of number of correct predictions to the total number of input samples.

**A *umber of Correct predictions***

*Accuracy = - —*

***Total number of predictions made***

It works well only if there are equal number of samples belonging to eachclass.For example, consider that there are 98% samples of class A and 2% samples of class B in our training set. Then our model can easily get 98% training accuracy by simply predicting every training sample belonging to class A.

When the same model is tested on a test set with 60% samples of class A and 40% samples of class B, then the test accuracy would drop down to 60%.Classification Accuracy is great, but gives us the false sense of achieving high accuracy.

The real problem arises, when the cost of misclassification of the minor class samples are very high. If we deal with a rare but fatal disease, the cost of failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to more tests.

Logarithmic Loss:

Log Loss is the most important classification metric based on probabilities. It's hard to interpret raw log-loss values, but log-loss is still a good metric for comparing models. For any given problem, a lower log-loss value means better predictions. Log Loss is a slight twist on something called the Likelihood Function. In fact, Log Loss is -1 \* the log of the likelihood function. So, we will start by understanding the likelihood function.

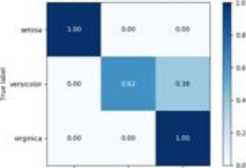
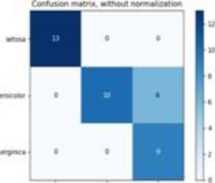
This is the loss function used in (multinomial) logistic regression and extensions of it such as neural networks, defined as the negative log-likelihood of the true labels given a probabilistic classifiers predictions. The log loss is only defined for two or more labels. For a single sample with true label yt in {0,1} and estimated probability yp that

Yt = 1, the log loss is -log P(yt|yp) = -(yt log(yp) + (1 - yt) log(l - yp))

Confusion Matrix:

Confusion matrix usage to evaluate the quality of the output of a classifier on the iris data set. The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

The figures show the confusion matrix with and without normalization by class support size (number of elements in each class). This kind of normalization can be interesting in case of class imbalance to have a more visual interpretation of which class is being misclassified.Here the results are not as good as they could be as our choice for the regularization parameter C was not the best.



Area under curve:

The area under a curve between two points can be found by doing a definite integral between the two points. To find the area under the curve y = f(x) between x = a and x = b, integrate y = f(x) between the limits of a and b. Areas under the x-axis will come out negative and areas above the x-axis will be positive.

Fl Score;

In statistical analysis of binary classification, the F1 score is a measure of a test's accuracy. It considers both the precision ***p*** and the recall ***r*** of the test to compute the score: ***p*** is the number of correct positive results divided by the number of all positive results returned by the classifier, and ***r*** is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as the number of all relevant samples (all samples that should have been identified as positive). The Fl score is the harmonic average of the precision and recall, where an Fl score reaches its best value at 1 (perfect precision and recall) and worst at 0.

\_ 1 \ -1

**recall 4- precision 1 \ precision • recall**

**2 / precision T recall**

Mean Absolute Error:

In statistics, mean absolute error **(MAE)** is a measure of difference between two continuous variables. Assume ***X*** and F are variables of paired observations that express the same phenomenon. Examples of ***Y*** versus ***X*** include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. Consider a scatter plot of ***n*** points, where point ***i*** has coordinates (xz, ***yf)...*** Mean Absolute Error (MAE) is the average vertical distance between each point and the identity line. MAE is also the average horizontal distance between each point and the identity line.

The Mean Absolute Error is given by:



Mean Squared Error:

In statistics, the **mean squared error (MSE)** or **mean squared deviation (MSD)** of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors—that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better.

MSE = ^£(7,-Ki)2.

AGRICULTURE CROP PRODUCTION USING MACHINE LEARNING

CHAPTER 6

**SCREENS**

0.27

0.27

0.27

**u**

EL

0.27

-0.90

k 0.75

ko.60

k 0.45

k 0.30

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| state | Rice | Jo war | Bajra | Maize | Ragi | Wheat | Barley | Gram | Tur |
| Andhra  Pradesh | 2921 185587 | 1054 114954 | 906 734010 | 3328 871137 | 1202 798161 | 792.552509 | 0.000000 | 1139 872298 | 441.485179 |
| Arunachal  Pradesh | 1190.479328 | 0.000000 | 0.000000 | 1365 528823 | 0.000000 | 1641.024943 | 0.000000 | 0 000000 | 1000.000000 |
| Assam | 1449 000000 | 0.000000 | 0.000000 | 710.000000 | 0.000000 | 1150.000000 | 0.000000 | 519.000000 | 707.000000 |
| Bihar | 1300 15D989 | 929 180615 | 933 165970 | 2281.541439 | 794.861011 | 1876.420716 | 1124 001590 | 938.555833 | 1183.194370 |
| Chattisgarh | 1177 000000 | 843 000000 | 0.000000 | 1567 000000 | 263.000000 | 996.000000 | 888.000000 | 728.000000 | 471.000000 |
| Goa | 2652.000000 | 0.000000 | 0.000000 | 0.000000 | 1000.000000 | 0.000000 | 0.000000 | 0 000000 | 0.000000 |
| Gujarat | 1610 000000 | 901 000000 | 1152 000000 | 1300 000000 | 868.000000 | 2451.000000 | 0.000000 | 739.000000 | 796.000000 |
| Haryana | 2894.000000 | 296.000000 | 1313.000000 | 2228.000000 | 0.000000 | 3979.000000 | 2735.000000 | 725.000000 | 988.000000 |
| Himachal  Pradesh | 1447 000000 | 0.000000 | 0.000000 | 2251 000000 | 1104 000000 | 1482.000000 | 1207 000000 | 901.000000 | 0.000000 |
| Jammu & Kashmir | 1960.000000 | 589 000000 | 571.000000 | 1535 000000 | 0.000000 | 1543.000000 | 631.000000 | 0 000000 | 0.000000 |
| Jharkhand | 1413.000000 | 988.000000 | 1253.000000 | 1465.000000 | 632.000000 | 1682.000000 | 922.000000 | 886.000000 | 860.000000 |
| Karnataka | 2561.392871 | 845 264458 | 626.179447 | 2654594052 | 1492.443191 | 736.637569 | 0.000000 | 506.726437 | 489.397884 |
| Kerala | 2197.000000 | 490 000000 | 0.000000 | 0.000000 | 1070.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| Madhya  Pradesh | 862.000000 | 985 000000 | 1244.000000 | 1525 000000 | 351.000000 | 1630.000000 | 1228.000000 | 855 000000 | 754 000000 |

**i i i**

cost profit yield

1000000 I

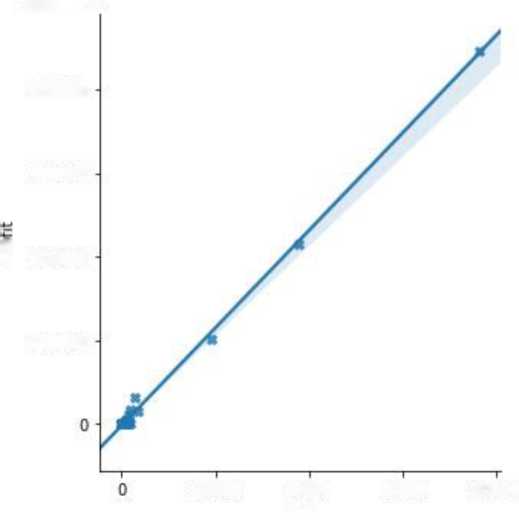
800000

600000

**o**

£ 400000

200000



20000 40000

yield

60000 80000

27500

25000

22500

20000 ■i—■

S 17500

15000

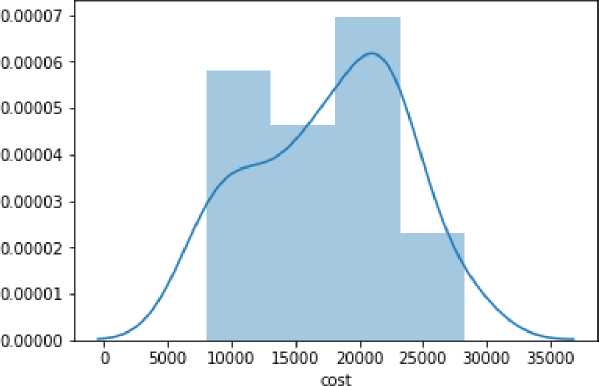
12500

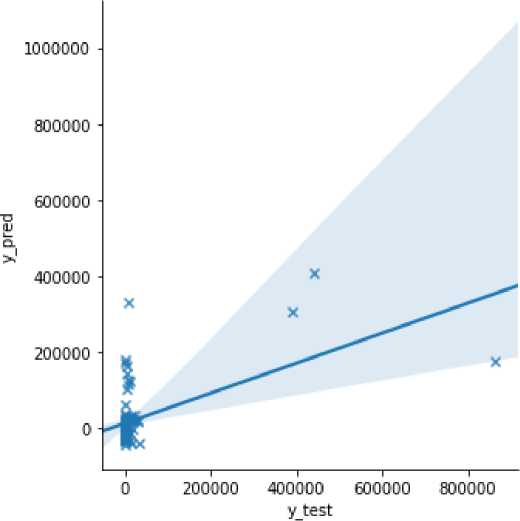
10000

7500

ap ar assam

state





OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | y | **R-squared:** | 0.93S |
| **Model:** | OLS | **Adj. R-squared:** | 0.903 |
| **Method:** | Least Squares | **F-statistic:** | 26.02 |
| **Date:** | Tue. 25 Feb 2020 | **Prob (F-statistic):** | 2.606-14 |
| **Time:** | 11:27:34 | **Log-Likelihood:** | -606.10 |
| **No. Observations:** | 51 | **AIC:** | 1250. |
| **Df Residuals:** | 32 | **BIC:** | 1287. |
| **1)1 Model:** | 19 |  |  |
| **Covariance Type:** | nonrobust |  |  |
| **coef** | **std err t** | **P>|t| [0.025** | 0.975] |
| **x1 -478.4824** | 3452.334 -0.139 | 0.891 -7510.656 | 6553.692 |

CHAPTER 7

**CONCLUSION & FUTURE ENHANCEMENT**

1. Conclusion:

Agriculture is one the most significant natural phenomenon that is not only important for the human beings only but the living beings. Now a days most of the people don‘t have an enough knowledge on farming because of that , most of the people getting loss in agriculture crop production. However, to overcome these issues the agriculture predictions can be done by using Machine Learning. The study aimed at building a predicting system using linear regression, Decision tree and Random forest that could predict agriculture crop accurately and efficiently with minimum error. This makes the farmers to take the right decision for right crop such that the agriculture sector will be developed by innovative ideas.

1. Future Enhancement:

This project describes crop yield prediction ability of the algorithm in future we can determine the efficient algorithm based on their accuracy metrics that will helps to choose an efficient algorithm for crop yield prediction.

BIBILIOGRAPHY

REFERENCES:

1. Data Mining - Concepts and Techniques - JIAWEI HAN & MICHELINE

KAMBER Harcourt India.2nd Ed 2013.

1. The Data Mining Techniques - ARUN K PUJARI, University Press.
2. Data Warehousing in the Real World - SAM ANAHORY & DENNIS MURRAY. Pearson Edn Asia.
3. DW - Data Warehousing Fundamentals - PAULRAJ PONNAIAH WILEY STUDENT EDITION.
4. The Data Warehouse Life cycle Tool kit - RALPH KIMBALL WILEY STUDENT EDITION.
5. The book Knowledge Discovery in Databases, edited by Piatetsky-Shapiro and Frawley [P-SF91], is an early collection of research papers on knowledge discovery from data
6. Pressman, Roger S (2009). Software Engineering: A Practitioner's Approach (7th Ed.). Boston, Mass: McGraw-Hill. ISBN 978-0073375977.
7. Sommerville, Ian (2010) [2010], Software Engineering (9th Ed.). Harlow, England: Pearson Education. ISBN 978-0137035151.
8. Jalote, Pankai (2005) [1991], An Integrated Approach to Software Engineering (3rd ed.). Springer. ISBN 978-0-387-20881-7..
9. R. Kohavi and F. Provost, "Glossary of terms," Machine Learning, vol. 30, no.2- 3, pp. 271-274, 1998.
10. Automate the Boring Stuff with Python by Al Sweigart.

***Internet References:***

1. <https://www.slideshare.net/rhspcte/software-engineering-ebook-roger-s-pressman>
2. <https://nptel.ac.in/downloads/106105087/>
3. <http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann->

Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-

Data-

1. Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf
2. <https://www.kaggle.com/nasirmeh/prediction-of-rainfall>
3. <https://stackabuse.com/linear-regression-in-python-with-scikit-learn/>
4. <https://www.tutorialride.com/data-mining/regression-in-data-mining.htm>