APPLICATION OF MACHINE LEARNING MODELLING FOR PREDICTION OF SOIL MOISTURE

Dissertation submitted to National Institute of Technology, Rourkela

In partial fulfillment for the award of the degree

Master of Technology

In

Civil Engineering (Water Resource Engineering)

By:

Shubham Vijay Papal

(Roll number: 716CE4014)

Under the Guidance of:

Prof. Sanat Nalini Sahoo

May, 2021

Department of Civil Engineering

National Institute of Technology Rourkela

Department of Civil Engineering

National Institute of Technology Rourkela

May 31, 2021

Certificate of Examination

Roll number: 716CE4014

Name: Shubham Vijay Papal

Title of dissertation: *Application of machine learning modelling for prediction of soil moisture.*”

We the below signed, after checking the dissertation mentioned above and the official record book(s) of the student, hereby state our approval of the dissertation submitted in partial fulfillment of the requirements of the degree of Master of Technology in water resources engineering at national institute of technology Rourkela. We are satisfied with the volume, quality, correctness, and originality of the work.

Department of Civil Engineering
National Institute of Technology Rourkela

Prof. Sanat Nalini Sahoo

May 31, 2021

Supervisor’s Certificate

this is to clarify that the work presented in the dissertation entitled “*Application of machine learning modelling for prediction of soil moisture*” submitted by *Shubham Vijay Papal*, roll number 716CE4014, is a record of original research carried out by him under our supervision and guidance in partial fulfillment of requirements of the degree of *Int. B.tech/M.Tech* in *Technology* in *Civil Engineering*. Neither this dissertation nor any part of it has been submitted earlier for any degree or diploma to any institute or university in India or abroad.

Prof. Sanat Nalini Sahoo

Declaration of Originality

I, Shubham Vijay Papal, roll number 716CE44014 hereby declare that this dissertation entitled "*Application of machine learning modelling for prediction of soil moisture*" presents my original work carried out as a *Int. B.tech/M.Tech* degree student of NIT Rourkela and, to the best of my knowledge, contains no material previously published or written by another person, nor any material presented by me for the award of any degree or diploma of NIT Rourkela or any other institution. Any contribution made to this research by others, with whom I have worked at NIT Rourkela or elsewhere, is explicitly acknowledged under the sections "Reference" or "Bibliography". I have also submitted my original research records to the scrutiny committee for evaluation of my dissertation.

I am fully aware that in case of any non-compliance detected in future, the senate of NIT Rourkela may withdraw the degree awarded to me on the basis of the present dissertation.

May 31, 2021

Shubham Vijay Papal

NIT Rourkela

Acknowledgement

The final year project has helped me to learn both theoretically and practically some important concepts of Engineering Hydrology which will be useful for me in future. This would not be possible without the people involved in the learning process. I wish to express my sincere gratitude to my supervisor Prof. Sanat Nalini Sahoo, Assistant Professor of Civil Engineering department, National Institute of technology, Rourkela for guiding me to this interesting research work. I thank her for constantly motivating me through his valuable counsel as well as her excellent tips to build my research and writing skills. I would also like to thank all the faculty members of the department of civil engineering for their constant support and motivation. I would also like to thank my lab colleagues Nathiajay Chandra and Sovan Sankalp for their valuable time to time help. Finally, I would like to thank my parents and brother for their blessings and support they have provided throughout the tenure of the project.

May31, 2021 | ***Shubham Vijay Papal***

NIT Rourkela | Roll Number: 716CE4014

ABSTRACT:

In recent years, due to a combination of drought and over-pumping surface and groundwater in many agricultural areas of India, water supplies have grown increasingly stressed. Farmers who used to grow millet, sorghum, and other grains have switched to sugarcane and paddy which are thirsty crops. Water intensive cultivations in drought-prone locations is a surefire way to run out of water. Despite this, the area under sugarcane cultivation has risen dramatically. Crop type, soil features, duration of excess water or flooding, initial soil moisture and nitrogen status of the soil before floods, crop stage, air and soil temperature, and other factors all influence the impact of excess water on crop growth and yield.

Adopting correct irrigation management practices can help to mitigate the negative effects of overwatering and maintain a balance between crop water needs and available water. Water loss, increased energy demand for pumping, and nitrogen and other micronutrient leaching are all consequences of excessive irrigation. Over-irrigation increases fertilizer costs for crops, as well as nitrogen expenses and nitrogen losses to groundwater. Groundwater contamination can be caused by poor irrigation and fertilization management.

Although there are no records to show when an irrigator first sampled the soil to determine whether it was time to irrigate, we can presume that this practice is nearly as old as irrigation. The approach of irrigation scheduling based on ‘soil moisture’ assessment is perhaps the oldest in use today. A proposed method of soil moisture prediction, uses novel machine learning techniques such as linear regression, K nearest neighbors, Support vector machine and Random forest regression. Model uses various parameters such as maximum temperature, minimum temperature, rainfall, evapotranspiration and historical data of soil moisture. Usage of machine learning techniques saves time. These techniques yield high coefficient of determination (R^2), a low mean squared error (MSE), and a low mean absolute error (MAE), according to numerical data.

Contents

List of Figures:

List of Tables:

Notations:

η_c	Efficiency of water conveyance.
η_a	Efficiency of water application.
η_s	Efficiency of water storage
η_u	Efficiency of water use.
C_u	Consumptive use.
w	Normal vector.
ϵ	Insensitive loss function.
ξ, ξ^*	Slack variables.
K	Kernel function
P	Probability Function.
SM	Soil Moisture
ET	Evapotranspiration.
T_{max}	Maximum Temperature.
T_{min}	Minimum Temperature
MAE	Mean Absolute Error
RMSE	Root Mean Square Error.
LR	Linear Regression
SVM	Support Vector Machine.
DTR	Decision Tree Regression.
KNN	K Nearest Neighbors.

CHAPTER-1:
INTRODUCTION

1.1 Overview:

Modern Humans evolved in Africa about two thousand years ago. In the next thousand years human migration reached distant mountains of Siberia and Alaska. Until then, the world's population was most likely lesser than a million people. Only with the advent of farming did the human population begin to rise exponentially. During 1 AD, discovery of massive fertile lands in the basins of Indus and Yangtze rivers led humans to grow population up to 170 million. Even today, The Vast farmlands of northern India and East China hold one third of human population. This has made possible due to availability of irrigation water and vastly spread fertile lands in river basins.

Agriculture has a long history in India, dating back to the Indus Valley Civilization. While irrigation channels were developed in Indus valley civilization during 4500 CE. As a result of this innovation, the Indus civilization developed in size and prosperity, leading to more planned communities with drainage and sewers. India is the world's second-largest producer of agricultural products. Agriculture employed more than half of the Indian workforce in 2018 and generated 17–18% of the country's GDP. India leads the globe in net cropped area, followed by the United States and China. Agriculture, however, is India's most populous economic sector and plays a key part in the country's entire socio-economic fabric.

The Invention of steam Engine in late 18th century has changed the economics of water forever. In postindustrial era, water has not been limited to agricultural outputs but it is also a secret ingredient in Computer chips and electricity consumption. Metropolitan areas have led an agglomeration of population like never before. This obviously led a question on irrigation water consumption which has relatively low output. India has 18% of the world's population and 4% of the world's pure water, with 80% of it being utilized for agriculture. India is now seeing a decline in accessible water resources, which has ramifications for the country's agriculture industry. Water scarcity is affecting several parts of the country. If the country's water consumption efficiency does not improve, it may face water shortage in the next one to two decades. It is critical that the agricultural sector helps to preventing the situation from worsening by maximizing the use of existing technology and resources to improve water efficiency. It is necessary to explore improving policies, tactics, and regulatory measures to reduce water usage. Water consumers in the agriculture sector can be educated and orientated to convert to more water-efficient production techniques, which can assist the country combat water constraint. Best practices enforcement can aid current policymakers and planners in improving governance structures and better understanding important indicators that can aid in data-driven decision-making.

1.2 Irrigation:

Throughout its life cycle, every crop requires a set amount of water at regular intervals. There is no need for irrigation water if the natural rain is adequate and timely to meet both of these requirements. However, in a tropical nation like India, natural rainfall is either insufficient or does not fall on a regular basis, as crops require. Because the amount and frequency of rainfall vary across the nation, certain crops may require irrigation in one section of the nation but not in another. The term "arid region" refers to an area where irrigation is required for cultivation. Semi-arid regions are those where substandard crops may be grown without irrigation.

1.2.1 Duty and Delta of crop:

Throughout its growing stage, each crop requires a specific amount of water at a specific interval of time. The amount of water required varies based on the climate, soil, and crop variety. This entire quantity of water required by the crop for complete development (maturity) can be represented in hectare-meters (Acre-ft), million cubic meters (million cubic-ft), or simply as the depth to which water would stand above the surface of the irrigated area if the complete quantity given stood above the surface without percolation or evaporation. The delta is the total depth of water (in cm) that a crop needs to reach maturity (A).

The link between the volume of water and the area of the crop that matures is known as the 'duty' of water. It may be defined as the number of hectares of land irrigated for full crop growth with a 1 m3 /sec water supply. Throughout the crop's whole base period (B), on a continuous basis. For example, if water flows at a rate of one cubic meter per second for k days and matures 100 hectares, the duty of water for that crop will be defined as 100 hectares per cumec to the base of k days. The letter D is commonly used to indicate a responsibility. Duty of irrigation water depends upon the following factors:

I. Type of crop: Varying crops demand different amounts of water, and so the responsibilities vary. In comparison to a crop that requires less water, a crop that requires more water would have less thriving acreage for the same amount of water. As a result, duty will be lower for crops that require more water and vice versa.

II. Climate and Season: As previously noted, duty includes water lost through evaporation and percolation. The number of losses will vary depending on the season. As a result, duty changes from season to season and even within the same season. The statistics for responsibilities that we use in general are their average levels during the full crop season.

III. Useful Rainfall: If part of the rain that falls directly on irrigated land is valuable for crop growth, then irrigation water will be used considerably less to mature the crop. The more the beneficial rainfall, the lower the irrigation water needs, and hence the greater the irrigation water duty.

IV. Type of soil: The water lost owing to percolation will be more if the permeability of the soil under the irrigated crop is high, and hence the obligation will be lower. As a result, the water obligation is lower on sandy soils with higher permeability.

V. Efficiency of cultivation method. If irrigation water is used wisely, the water duty will improve since the same amount of water will be able to irrigate a larger area with the same amount of water. As a result, cultivators should be thoroughly trained and informed on how to use irrigation water efficiently.

1.2.2 Irrigation Efficiencies:

The ratio of water output to water intake is called efficiency, and it's commonly stated as a percentage. Input minus output equals losses, thus if losses are higher, output will be lower, and efficiency will be lower. As a result, efficiency and losses are inversely proportional. Because water is wasted in irrigation through numerous processes, many types of irrigation efficiency exist, as shown below.

I. Efficiency of water conveyance: It is the proportion of water supplied to the fields from the canal's outflow point to the water entering the channel at its beginning point. η_c might be used to represent it. It takes into account transportation or transit losses.

II. Efficiency of water application: It's the ratio of the amount of water held in the root zone of the plants to the amount of water provided to the field. n_a might be used to symbolize it. It's also known as agricultural efficiency since it considers the amount of water lost on the farm.

III. Efficiency of water storage: It's the proportion of water held in the root zone during irrigation to the amount of water required in the root zone before irrigation (i.e., field capacity of existing moisture content). n_s might be used to symbolize it.

IV. Efficiency of water use: It's the proportion of beneficially utilized water, including leaching water, to water given. It might be symbolized by the letter η_u

1.2.3 Evapotranspiration:

The entire quantity of water utilized by the plant through transpiration and evaporation from nearby soils or from plant leaves in any given time may be characterized as consumptive usage for a certain period. Consumptive usage (Cu) levels may change for various crops, as well as for the same crop at various periods and locations. In reality, the consumption of a given crop at a given location might change during the day, the month, and the crop cycle. Values of daily or monthly consumptive consumption are normally calculated for a certain crop and at a certain location. The irrigation need of the crop is then calculated using the values of monthly consumptive usage across the full crop period.

1.2.4 Effective Rainfall:

Effective rainfall is rain that falls throughout a crop's growing season and is sufficient to fulfil the crop's evapotranspiration requirements. It excludes water lost through deep percolation under the root zone and water lost through surface runoff.

1.2.5 Factors affecting Consumptive Use (Evapotranspiration):

Several variables influence the rate of evapotranspiration at any given point on the Earth's surface.

I. Energy Availability: The higher the rate of evapotranspiration, the more energy is accessible. To turn 1 gramme of liquid water into a gas, 600 calories of heat energy are required.

II. Humidity: humidity gradient affects the rate of evapotranspiration. Drier air increases both the pace and quantity of water vapor entering the atmosphere.

III. Wind Speed: The speed of the wind just above the surface has positive impact on evapotranspiration rate. Winds have an impact on evapotranspiration because they transfer heat energy into an area.

IV. Water Availability: If there is no water available, evapotranspiration will not occur.

V. Physical attributes of vegetation: Vegetative cover, plant height, leaf area index, and leaf shape, as well as the reflectance of plant surfaces, may all impact evapotranspiration rate.

VI. Soil Characteristics: Heat capacity, soil chemistry, and albedo are all factors that can influence evapotranspiration.

1.2.6 Soil Moisture-irrigation Relationship:

Ground water is the water below the water table, while soil moisture is the water above the water table. The soil zone, also known as the root zone, extends down from the ground surface and is defined as the depth of overburden pierced by the roots of plants, as illustrated in figure 3. This zone is the most significant from an irrigation standpoint since it is from this zone that the plants get their water. When water falls on the ground, a portion of it is absorbed in the root zone, while the remainder flows downward under gravity and is referred to as gravity water.

I. Field Capacity: When all the gravity water has flowed down to the water table following a rain or irrigation water application, a certain quantity of water is held on the surfaces of soil grains via molecular attraction and loose chemical interactions (i.e., adsorption). The field capacity refers to the amount of water that cannot be easily drained by gravity. The water content of soil after free drainage has occurred for a sufficient amount of time is therefore the field capacity.

$$Field\ Capacity = \frac{wt.of\ water\ retained\ in\ certain\ volume\ of\ water}{wt.of\ the\ same\ volume\ of\ dry\ soil.} \qquad (1)$$

.

II. Readily available moisture: It is the fraction of the available moisture that the plants can most easily absorb, which is around 75 to 80 percent of the available moisture.

III. Soil moisture deficiency: Field moisture deficiency, also known as soil moisture deficit, is the amount of water necessary to bring the soil moisture content of a particular soil to the field capacity.

IV. Equivalent Moisture: The field capacity is the amount of water retained by a saturated soil after gravity has acted on it. Similarly, comparable moisture is the amount of water retained by a saturated soil after being spun for 30 minutes at 1000 times the gravitational force. As a result, it is slightly less than or equal to the field capacity.

1.2.7 Estimating irrigation depth and frequency using the soil moisture-regime concept:

Plants eat water or soil moisture through their roots. As a result, it's critical that enough moisture is accessible in the soil from the surface to the root zone depth. As previously stated, the root zone soil moisture can vary between field capacity (upper limit) and wilting point moisture content (lower limit), as illustrated in Fig. 1 It is also clear from the preceding discussion that soil moisture should not be allowed to become reduced to the point of wilting, since this would result in a significant drop in crop production. The optimal degree of soil moisture depletion in the root zone that may be tolerated without a drop in crop output must be determined via experimentation for each crop and soil. Irrigation water should be delivered as soon as the moisture content reaches this optimal level (setting irrigation frequency), and the quantity should be just enough to bring the moisture content up to the field capacity, with application losses taken into account. After the fresh irrigation dosage is delivered, the plants will begin to use the water, and soil moisture will begin to diminish. As soon as the soil moisture reaches the ideal level, as illustrated in Fig. 2, it will be recouped by a fresh dosage of irrigation.

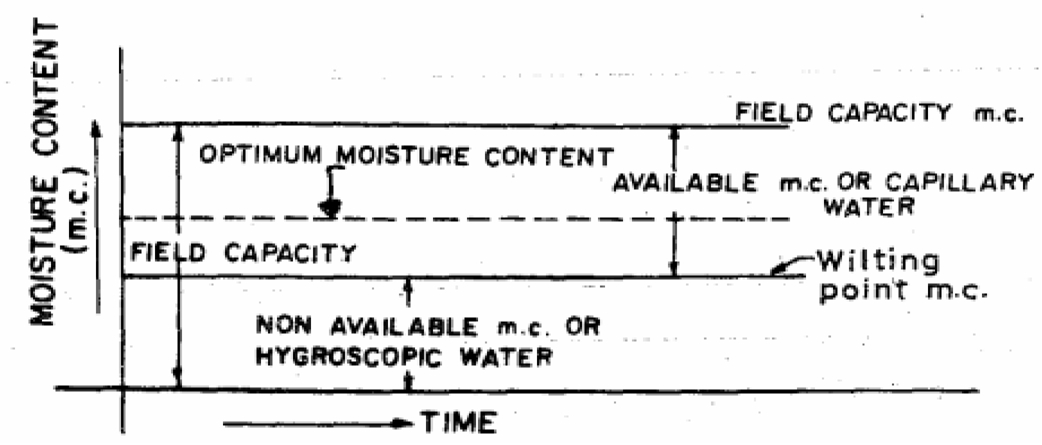


Figure. 1. Initial Soil moisture Variation

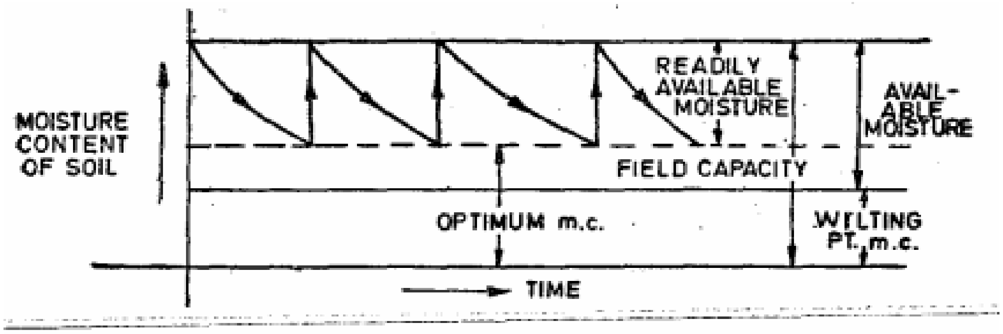


Figure. 2 Soil Moisture Variation at ideal Stage

1.3 Irrigation in India:

Agriculture employs more than 50% of India's population, either directly or indirectly. Rice, wheat, sugarcane, tea, cotton, groundnut, jute, coffee, rubber, and garden crops (such as coconuts, oranges, and other citrus fruits) are among India's most important crops. Different soil types are required for different sorts of crops. For example, high retentive soil (40 percent clay) is ideal for growing crops such as sugarcane, rice, and other water-intensive crops. Light sandy soil (2 to 8% clay) is ideal for crops like gramme, fodder, and other low-water crops.

The areas with the highest proportion of irrigation include fertile alluvial plains with perennial rivers and drinkable groundwater, as well as those with less than 125 cm of annual precipitation. The Kashmir Valley, large parts of Punjab (Northern India) and Haryana, the Ganga-Yamuna Doab of Uttar Pradesh (Northern India), the Western part of the South Bihar (Eastern India) Plain, Birbhum, West Bengal (Eastern India), Lakhimpur, Assam (Northeastern, the Godavari Krishna Deltas and Chengalpattu district), Tamil Nadu have the highest intensity of irrigation (Southern India).

Irrigation now utilizes roughly 84 percent of the available water. The industrial and household sectors, respectively, require around 12 and 4% of total available water. With irrigation expected to continue to be the primary water consumer, “per drop more crop” is a must. There has been a major shift in irrigation sources throughout time. Canal’s percentage of net irrigated land has decreased from 39.8% in 1950-51 to 23.6 percent in 2012-13. During the same time span, the percentage of groundwater sources has risen from 28.7% to a stunning 62.4 percent. This increase reflects the stability and better irrigation efficiency of groundwater irrigation, which is 70–80% compared to 25–45% for canal irrigation [4].

Flooding is one of the common irrigation techniques used in India. Flooding or over-irrigation can generate wet soil conditions, which can harm crops, diminish yields, and lead to groundwater pollution. Adopting effective irrigation management practices can help to mitigate the negative effects. Overwatering can result in nitrogen leaching and runoff. Excess water, according to research, can increase weed pressure and produce a disease-friendly environment.

1.4 Application of Machine Learning Techniques and Internet of Things (IOT) in the field Water Resources and Irrigation:

The Internet of Things (IoT) is a popular method of linking objects and collecting data. IoT frameworks are used to handle and interact with data and information via the Internet of Things. Users may register their sensors, generate data streams, and process data in the system. IoT may be used in a variety of agricultural approaches. Smart Cities, Smart Environment, Smart Water, Smart Metering, Security and Emergency, Industrial Control, Smart Agriculture, Home Automation, and e-Health are examples of IoT applications. The ‘Internet of Things' is built on a device that is capable of analyzing and communicating sensed data to the user.

According to UN projections, water shortages would directly affect roughly 20% of the world's population by 2025, and will have an indirect impact on the rest of the population, economy, and ecosystems. Smart water systems based on a combination of Internet of Things, big data, and AI technologies can help prevent these predictions from occurring and repair the harm that has already been done due to injudicious use of water resources

1.4.1 Goals of Smart Systems in Water Resources:

The basic goal of smart water management is to use and recycle water resources in a fair and sustainable manner. Water is becoming an ever more valuable resource due to rising population, environmental concerns, and strain on the food and agriculture sectors. Water management technologies and activities aim to achieve the following goals in this regard.

I. Reduce the amount of water wasted in high-volume industries including manufacturing, agriculture, and power generation. Precision farming, smart irrigation, and real-time water metering are examples of high-tech methods that may be implemented.

II. Improve water quality and protect it from contamination from chemical waste and natural pollutants like acidification. It needs a sensor technology for real-time monitoring and control to enhance and maintain water quality.

III. Water systems such as water collectors, treatment facilities, distribution mains, and wastewater recycling facilities may all be made more efficient. Systems can maintain crucial parameters like water pressure, temperature, and flow in sight, conduct predictive maintenance, and reduce breakage and downtime by using IoT and data analytics for asset management.

IV. Smart water management systems with leak and moisture sensors can be used to control leaks. Leakage control is critical to keep water supplies and budgets safe, since about $3 billion is spent each year to repair the damage caused by leaks.

V. Practice consumption monitoring to optimize and maintain water resource usage under control at many levels in the home, industry, country, or the entire planet.

1.4.2 Field Application of Sensor Technology:

IOT-based remote sensing collects data from sensors installed along farms, such as weather stations, and transfers it to an analytical tool for examination. Sensors are devices that detect abnormalities. Farmers may keep an eye on their crops using an analytical dashboard and take action depending on what they learn.

I. Crop Monitoring:

Sensors positioned throughout the farms track changes in light, humidity, temperature, shape, and growth of the crops. Any irregularity discovered by sensors is investigated, and the farmer is alerted. As a result, remote sensing can aid in disease prevention and crop monitoring.

II. Weather Conditions:

The data acquired by sensors in terms of humidity, temperature, moisture, precipitation, and dew determines the weather pattern in farms, allowing for crop cultivation that is optimal for the crop.

III. Soil Quality:

The examination of soil quality aids in assessing nutritional value and drier parts of farms, as well as soil drainage capacity and acidity, allowing the quantity of water required for irrigation to be adjusted and the most effective form of cultivation to be chosen.

1.4.3 Field Application of Machine Learning:

I. Soil Management: Soil is a varied natural resource for agricultural professionals, with complicated processes and nebulous mechanisms. Its temperature alone can provide information into the impact of climate change on regional output. To understand the dynamics of ecosystems and the impact in agriculture, machine learning algorithms investigate evaporation processes, soil moisture, and temperature.

II. Water Management: Agriculture's water management has an influence on the hydrological, climatological, and agronomic balance. So far, the most developed ML-based applications are related to daily, weekly, or monthly evapotranspiration estimation, which allows for more efficient irrigation system use, and daily dew point temperature prediction, which aids in identifying predicted weather events and estimating evapotranspiration and evaporation.

III. Yield Predictions: Yield prediction, which includes yield mapping and estimating, crop supply and demand matching, and crop management, is one of the most significant and popular issues in precision agriculture. To make the most of the produce for farmers and the people, state-of-the-art systems have gone far beyond basic prediction based on historical data, including computer vision technology to offer data on the fly and thorough multidimensional analysis of crops, weather, and economic situations.

IV. Disease Detection: The most extensively employed practice in pest and disease management, both in open-air and greenhouse situations, is to evenly spray insecticides across the cropping area. To be effective, this method necessitates the use of large volumes of pesticides, which comes with a tremendous financial and environmental cost. ML is employed as part of a broader precision agriculture strategy, in which agrochemicals are applied at specific times, locations, and to specific plants.

V. Crop Quality: Crop quality traits can be accurately detected and classified, which helps raise product prices and minimize waste. Machines, in comparison to human specialists, can employ seemingly useless data and linkages to disclose and discover new attributes that have a role in the overall quality of crops.

1.5 Objectives:

The objectives of the project are listed below:

1. To apply the machine learning techniques for prediction of soil moisture. Hence, to test the Support vector machine, Random Forest regression and multiple linear regression for obtained datasets.
2. To study the data structure of Hydrological and Metrological datasets obtained from various sources. Hence, understanding the interdependencies of parameters and choosing the appropriate parameters to improve the yield of soil moisture.

3. To improve the accuracy of models by data cleaning and data transformation. Hence to choose the best data transforming techniques to yield the better predictions of soil moisture.
4. To choose the best Machine Learning method for respective dataset by using fitness function calculated by using various error formulae.

1.6 Organization of thesis:

This research project consists of six chapters. Basic introduction is given in first chapter, second chapter contains literature review, subject background and methodology is provided in chapter three, fourth chapter is comprised of results and analysis, chapter five consist of conclusion and scope for the future work, and finally references are given in chapter 6.

Starting with a brief overview of the study, chapter 1 contains a summary of basic concepts of irrigation, relationship between consumptive use and soil moisture, modern methods of irrigation and application of machine learning in water resources. Objectives of the study are also given in chapter 1.

Previous works of many prominent researchers, scientists, and investigators on the present topic are provided in chapter 2 and it highlights the research related to Sensors based irrigation, Machine learning in water resources and trend analysis of hydrological data.

Chapter 3 comprise of methodology that is mathematics behind all four machine learning techniques applied in this research. It also includes the short descriptions of error analysis and fitness criteria, data transformation and feature extraction.

In chapter 4, all graphical and numerical results have been included. Result section has been divided into three parts which are Data Structure, training and cross validation and testing. All simulations are done by using Python programming language.

Chapter 5 includes the conclusion and scope for the future work part. And in chapter 6 the references for writing this thesis are provided.

CHAPTER-2: LITERATURE REVIEW

2.1. Sensor Based Irrigation:

G.W. Kite, P. Droogers, (1999) have introduced the modern techniques to measure the evapotranspiration. Evaporation and transpiration are crucial parts of the hydrological cycle that can't be observed directly. The residual in the water balance equation has traditionally been used to calculate real evapotranspiration. Researchers have recently begun to estimate area-specific actual evapotranspiration using scintillometers, remotely sensed data, and hydrological models. Authors have compared various datasets and locations in turkey to examine these techniques and to decide whether they suit the conditions   Recommendations have been given based on capabilities and limitations of each method. Planning and construction of irrigations schemes, Allocation and efficiency water are some the major areas covered. Impact of climate change will more likely to have huge impact of agriculture and irrigation planning. The best method to study this impact is through field data and not through remote sensing. The satellite and FAO-24 methodologies have the most variability, according to the findings. The FAO-56, models, and field methodologies all demonstrate a higher level of consistency.

Russeil J. Quaus et al., (2002) Despite the fact that air temperature and the number of days with precipitation appear to be more important drivers of actual water usage, ET and precipitation should be the governing elements in irrigation**.** The study's goal was to assess the technology's efficacy and dependability in terms of water saving. In Boulder, Colorado, in 1997, granular matrix soil moisture sensors were utilized to regulate urban landscape irrigation. Soil moisture sensors are one example of this technology. They may be hard-wired into a clock-driven irrigation system to prevent or enable planned watering based on changes in the soil's moisture status. Homeowners and landscaping users have continued to be hesitant to utilize soil moisture sensors since they were difficult to operate, had poor longevity, and so on. The paper outlines a straightforward approach for homeowners and landscaping contractors to track the sensor's performance in relation to ET over the course of a season using historical climatic records or easily available current weather conditions. The findings of this study show that the GMSs are both water-saving and cost-effective to run.

Michael D. Dukes et al., (2003) On average, the humid region gets more precipitation each year than it loses to evapotranspiration (ET). Short drought episodes, on the other hand, have been demonstrated to impair yields and quality of some vegetables. At the 15 and 30 cm depths, sensor-based irrigation treatments resulted in considerably higher soil volumetric moisture levels. The findings suggest that high-frequency irrigation events based on soil moisture sensor management can sustain agricultural yields while lowering irrigation water usage; nevertheless, more study is needed to replicate the first year's findings.

K. P. Sudheer et al., (2003) The potential of artificial neural networks (ANN) in calculating real crop evapotranspiration from restricted meteorological data is investigated in this research. The model predictions are compared to lysimeter ET measurements. The study's findings clearly show that the ANN approach is capable of predicting the ET. According to the findings, using an ANN technique, crop ET may be calculated from air temperature. However, because the current study only looked at a particular crop over a short period of time, further research with different crops and weather conditions may be needed to confirm these findings.

A. Narayanamoorthy, (2003) Improving present water usage efficiency is one of the most important prerequisites for the long-term usage of irrigation water**.** Due to massive losses in distribution and evaporation, water utilization efficiency under the flood technique of irrigation, which is widely used in India, is extremely poor**.** The drip irrigation system, which was just recently introduced into Indian agriculture, has shown to be an efficient technology for enhancing water usage efficiency. The findings of experimental station data suggest that DMI saves between 12 and 84 percent of water per hectare for various crops while also enhancing crop production. For the country as a whole, DMI's core and net potential areas are assessed to be 51.42 million hectares (mha) and 21.27 mha, respectively. The entire amount of water saved by utilizing DMI's net potential area is expected to be around 11.271 million ha m. An additional irrigated area of 11.22 mha under FMI or 24.12 mha under DMI may be generated by conserving water.

Bernard Cardenas-Lailhacar et al., (2005) Soil moisture sensors (SMSs)-based irrigation system automation has the ability to deliver maximum water usage efficiency by keeping soil moisture at optimum levels. Depending on the soil moisture state, these SMSs may cause scheduled watering cycles to be disrupted. Other comparisons were made between plots with and without a rain sensor. There was also a non-irrigated treatment. There were no significant variations in turfgrass quality across treatments, as shown by the high quality in non-irrigated plots. Treatment without a rain sensor was utilized 45 percent of the time.

S.L. Davis et al., (2009) As compared to a time-based treatment without a rain sensor, the ET controllers saved 43 percent of water and were nearly twice as efficient at lowering irrigation when compared to a rain sensor alone. The net irrigation demand for the region was used to construct time-based treatments, which resulted in less water being applied than if the irrigation was scheduled without utilizing historical ET and effective rainfall. The RTIME treatment yielded savings comparable to ET controllers. The water-saving potential of these controllers in landscapes will be determined by the irrigator's behaviors and preferences. To demonstrate water savings, these controllers must be examined in ‘‘real world" settings.

Sam Geerts and Dirk Raes, (2009) In arid places, deficit irrigation (DI) has been thoroughly studied as a profitable and long-term production method. This method tries to enhance water productivity and stabilize rather than maximize yields by restricting water applications to drought-sensitive plant phases. A specific level of seasonal moisture must be ensured at all times. Drought tolerance varies greatly by genotype and phenological stage, therefore DI requires accurate information of crop response to drought stress.

Richard G. Allen et al., (2011) Multiple factors govern the accuracy of measurement of evapotranspiration. ET measurement accuracy necessitates well-calibrated and well-maintained devices, as well as, in most instances, a solid understanding of the physics of turbulence and heat and radiation transport that drive the measurement. To compensate for the spatial non-uniformity of vegetation in most vegetation systems, numerous sensor sites within the measurement region are required. To create information on ET, ET as a residual must gather reliable and representative measurements of net radiation and soil heat flux density. While remote sensing techniques do not ‘measure' ET but rather derive it from energy balance or vegetation indices, these approaches, particularly those based on thermally driven energy balance, are highly strong in terms of geographic coverage and measurement of spatial variation in ET.

Sharon Dabach et al., (2011) The adjustment of operational parameters such as irrigation threshold and irrigation volume is required to improve the sustainability of irrigation systems. The link between irrigation schedule, irrigation threshold, and irrigation volume has been investigated. The findings reveal that HYDRUS 2D/3D irrigation event and matric head forecasts are in good agreement with experimental data, and that the code may be used to optimize irrigation thresholds and water quantities applied in an irrigation episode to improve water efficiency.

Shahbaz Mushtaq and Mahnoosh Moghaddasi, (2011) System managers and irrigators are being compelled to examine deficit-irrigation solutions as worries about climate change and environmental water demand grow. The potential of deficit irrigation was assessed using a non-linear optimization model that included endogenous crop production and profit factors**.** Deficit irrigation optimization might achieve both environmental flows and net return targets, improve total water usage efficiency, and so provide an effective adaptive response to climate change.

G. L. Grabow et al., (2013) There are two primary irrigation control technologies that are commercially available. One is based on ET estimations, while the other is based on data from a soil-moisture sensor (SMS). Although the ET treatment supplied adequate watering, it had the lowest irrigation efficiency. When compared to timer-based treatments, SMS treatments saved 39 percent on average in SMS1 treatments and 24 percent in SMS2 treatments, but ET treatments used 11 percent more water on average than timer-based treatments.

2.2 Data driven Modelling in Hydrology and Irrigation:

S. K. Jain, et al., (1999) The use of artificial neural networks (ANNs) for reservoir inflow prediction and operation. The ANN was shown to forecast high flows better, whereas the autoregressive integrated moving average model predicted low flows better. Through linear and nonlinear regression as well as the ANN, the best release was linked to storage, inflow, and demand. The ANN is a potent tool for input-output mapping and may be utilized well for reservoir inflow forecasting and operation, according to intercomparison.

Glenn De’ath and Katharina E. Fabricius, (1999) For the examination of complicated ecological data, classification and regression trees are suitable. Physical and geographical characteristics were equally significant predictors of abundances when used separately and lost little when used together. In this vast reef complex, where information on the physical environment is frequently unavailable, spatial variables serve as good surrogates for physical variables. Trees discover patterns that linear models can't identify.

Yang Qiu et al., (1999) Spatiotemporal prediction of soil moisture content using multiple-linear regression. MLR yields the lowest values in mean absolute error of prediction (MAE) and Akaike information criteria, this model was either the most exact or the most economical in predicting soil moisture content in space and time (AIC). The spatial-GM model and the land use-GM model have similar performance and cost-benefit ratios. The spatiotemporal-SM model's improved resilience over the other two models is particularly noticeable in the prediction of soil moisture content at 0–5 cm, and declines as soil depth increases.

Hongli Jiang and William R. Cotton, (2004) A feasibility of estimation of soil moisture using artificial neural network**.** The ANN model is a viable option for estimating soil moisture. Daily precipitation, normalized difference vegetation index (NDVI), skin temperature, and soil moisture profiles are used to calibrate (train) and validate (test) the ANN model. The ANN technique to soil moisture assessment has the benefit of being able to generate estimates with resolution comparable to remotely sensed IR data and having the possibility for global coverage.

Mohammad Sajjad Khan, et al., (2006) Potential of the support vector machine SVM in long-term prediction of lake water levels. SVM performed well and was shown to be competitive with MLP and SAR models. Based on root-mean square error and correlation coefficient performance criteria, the SVM model beats the two other models for 3- to 12-month-ahead prediction.

M. Kashif Gill et al., (2006) Soil Moisture Prediction using support vector machines. SVM forecasts for the next four and seven days are generated using soil moisture and weather data. In terms of soil moisture predictions, SVM models outperformed ANN models.

S. S. Zanetti et al., (2006) Using Artificial Neural Networks with Minimum Climatological Data to Estimate Evapotranspiration. The ANNs (multilayer perceptron type) were trained to estimate ET as a function of maximum and lowest air temperatures, extraterrestrial radiation, and daylight hours, the last two of which had previously been estimated as a function of the local latitude. When using only the highest and minimum air temperatures in the testing phase of ANN, it is feasible to predict ET.

M. Kashif Gill, Mariush W. Kemblowski et al., (2006) combining two state-of-the-art techniques, support vector machines (SVMs) and ensemble Kalman filter (EnKF) for prediction of soil moisture. It has been demonstrated that the SVM-EnKF coupling improves soil moisture prediction.

Sharad K Jain et al., (2008) Physical interpretation of models for estimating evapotranspiration using ANN. The values of numerous meteorological variables must be known in order to estimate evapotranspiration (ET). An ANN can reliably estimate ET even with minimal meteorological factors. The reasons for the ANN's capability in estimating the ET successfully from limited climatic data were explained by a study of the ANN-ET models produced.

Yanbing Qi et al., (2008) Artificial neural network models were used to predict consumptive use under various soil moisture and salinity conditions. Corn seedlings ET is susceptible to soil salt stress at all phases of development, albeit the salinity threshold at which the impact is felt and the magnitude of the impact varied for each stage, with the booting and tasseling stages being the most resilient. Actual crop evapotranspiration can be evaluated by considering explicitly water and salinity stresses.

Shaoe Yang et al., (2009) To forecast soil moisture and nitrate nitrogen content, a Support Vector Machine based on Time Series was used. SVM-TS is capable of predicting soil moisture. In the 0-30cm soil layer, there is no discernible variation between anticipated and actual values for NO3—N concentration. In addition, SVM-TS may be used to forecast nitrogen concentration.

Sajjad Ahmad et al., (2009) Machine learning approach for prediction of soil moisture by using remote sensing data. SVM has been used to forecast a quantity in the future based on prior data training. To estimate soil water content (SM), researchers employed remote sensing data from the Tropical Rainfall Measuring Mission (TRMM), as well as the Normalized Difference Vegetation Index (NDVI) from the Advanced Very High-Resolution Radiometer (AVHRR). The SVM model is able to capture the variability in recorded soil moisture. The SVM model is able to capture the variability in soil moisture measurements. The SVM model outperforms the ANN and MLR models in estimating soil moisture.

Irena Hajnsek et al., (2009) The whole vegetation-growing season for three crop kinds was studied for the first-time using SAR data and ground measurements collected during the AgriSAR initiative. Even when compared to highly dissimilar fields such as corn, rape, or wheat, the findings produced from the five decompositions are often not significantly different from one another. The estimation performance of the dihedral as well as the surface derived soil-moisture component across the whole growth season showed a substantial range.

Noel D. Evora and Paulin Coulibaly, (2009) Remote sensing applications in hydrology by using data-driven modelling. In remote sensing applications, ANNs have gained popularity as useful inverse models that can extract physical features of interest, such as precipitation, from remote sensing measurements acquired from radars or satellites. The extraction of precipitation and snow water equivalent (SWE) from remote sensing data is the subject of this research.

Luca Pasolli et al., (2011) Estimating Soil Moisture with the Support Vector Regression Technique. Because it delivered higher estimation accuracies and was more stable and resilient to outliers and noise in the data, SVR is a viable alternative to the more standard MLP NN regression approach.

Muhammad Sulaiman et al., (2011) using optimal steepness coefficient to improve the water level forecasting performance in an ANN. When compared to the conventional Steepness Coefficient, the optimum Steepness Coefficient enhanced the predicting accuracy of the ANN data training and data validation. Importantly, using the appropriate SC enhanced the performance of ANN data training greatly.

Adebayo J. Adeloye et al., (2011) Modeling of evapotranspiration in a reference crop using neural computing. Estimation of reference crop evapotranspiration is an important factor in order to calculate crop water requirements. SOM-based estimates were also shown to be much better than those calculated using standard empirical ET approaches, which are advised when the complete complement of input data required to drive the Penman-Monteith model is lacking.

Raorane A.A. et al., (2012) Yield estimation of agriculture by using data mining techniques. Three major techniques which are Support Vector Machines, Decision Trees and Artificial Neural Network have been studied for data mining. Wine Fermentation process, Datasets of soil and plants pollution in atmosphere can be analyzed by these methods.

D Ramesh, et al., (2013) Data Mining Techniques and Applications to Agricultural yield data.

K-Means, K-Nearest Neighbor (KNN), Artificial Neural Networks (ANN), and Support Vector Machines are among the data mining techniques used (SVM). Data Mining techniques can be used to tackle yield prediction problems. Biclustering methods can be used to extract useful information from collections of data connected to agriculture. The K-Means technique is capable of clustering the samples.

Sujay Raghavendra et al., (2014) Support vector machines are the most significant advancement in the field of hydrological parameter analysis. after artificial neural networks and fuzzy networks. Several SVM variations, such as epsilon-SVR, nu-SVR, and LS-SVM, provide greater flexibility for modelling unique data-oriented challenges. The kernel of an SVM model determines a lot of its properties. The kernel of an SVM model determines a lot of its properties.

Bushra Zaman and Mac McKee (2014) Using Multivariate Relevance Vector Machines, Spatio-Temporal Prediction of Root Zone Soil Moisture. Four days ahead of time, root zone soil moisture at one- and two-meter depths is projected. For all four days, the MVRVM gives good results. Bootstrapping is a technique for detecting over- or under-fitting as well as uncertainty in model estimations.

Daniela Angileri et al., (2014) Detecting trend using seasonal data. The number of days and years utilized for averaging are two tuning factors in the MASH. MASH is capable of handling seasonal and interannual variability in time series. Significant hydrological patterns may or may not have an impact on water resources.

Ch. Suryanarayana et al., (2014) Groundwater level prediction using an integrated wavelet-support vector machine. Combining multiple methodologies to create a hybrid tool to increase prediction accuracy has been a frequent approach in recent years for a variety of applications. When compared to SVR, ANN, and the standard Auto Regressive Integrated Moving Average, the WA-SVR model delivers superior accuracy in predicting groundwater levels in the research region.

Dr.P. Isakki and R. Sujatha (2016) It is vital to make agriculture more efficient by foreseeing and enhancing yields based on prior agricultural data. The crop output can be improved and the farmer's income level will be boosted by using these approaches.

R.M. Bhagat et al., (2016) Tea Yield Prediction Using Data Mining Techniques in the Face of Climate Change. Multiple Linear Regression is well known technique in field of data mining. Planters/farmers might utilize the methodology to forecast future crop output and, if the estimates fall short of expectations and economic viability, use alternate adaptive steps to maximize yield.

Walison B. Alves et al., (2017) forecasting reference evapotranspiration by artificial neural network. ET is the most active variable in the hydrological cycle and the key component of water balance in agricultural systems. The Penman-Monteith technique was used to estimate ET, and the feed-forward Multi-Layer Perceptron was used to forecast it. The accuracy, precision, and trend of ANN performance are evaluated. Using artificial neural networks and merely air temperature data as an input variable, ET could be reliably approximated with a day to spare at any time of the year.

Gaurav Shukla et al., (2017) implementation and assessment of a random forest classifier as a soil spatial predictive model. A single decision tree may be susceptible to noise, but when numerous decision trees are combined, the influence of noise is reduced, resulting in more accurate outcomes. A single decision tree may be susceptible to noise, but when numerous decision trees are combined, the influence of noise is reduced, resulting in more accurate outcomes. RF and CEB demonstrate the bootstrap-strong ensembling mapping capacity across a single tree model.

Yangbing Qi et al., (2018) Utilization of Artificial Neural Network Model to predict consumptive use under different soil moisture content and soil salinity conditions. soil moisture content, total salt content, plant height, leaf area index, and crop reference evapotranspiration are among the five inputs of artificial neural networks models. corn seedlings at all stages of development, ET is susceptible to soil salt stress. straightforward method of determining real crop evapotranspiration while explicitly incorporating water and salinity stressors.

Engin Pekel, (2019) Estimation of soil moisture using decision tree regression. The data from SM estimation may be handled using decision tree regression, which meets the fitness condition.

Odi Nurdiawan et al., (2020) Comparison of the K-Nearest Neighbor algorithm and the decision tree on moisture classification. K-Nearest Neighbor is the most accurate algorithm for classification of moist soil.

H. K. Karthikeya et al., (2020) Prediction of Agricultural Crops using KNN Algorithm. For the prediction model and crop yield prediction, the K-NN algorithm is employed, and the accuracy is attained. In the realm of agricultural production, the use of machine learning algorithms has a promising future.

Tripti Dimri, et al., (2020) Time series analysis of climate variables using seasonal ARIMA approach. The predicted data corresponds to the data's trend. Extreme rainfall occurrences and temperature results, on the other hand, show over-predictions. Pattern and trend data may be used as a forecasting tool for the creation of improved water management strategies in the region.

CHAPTER-3
METHODOLOGY

3.1. Study Area:

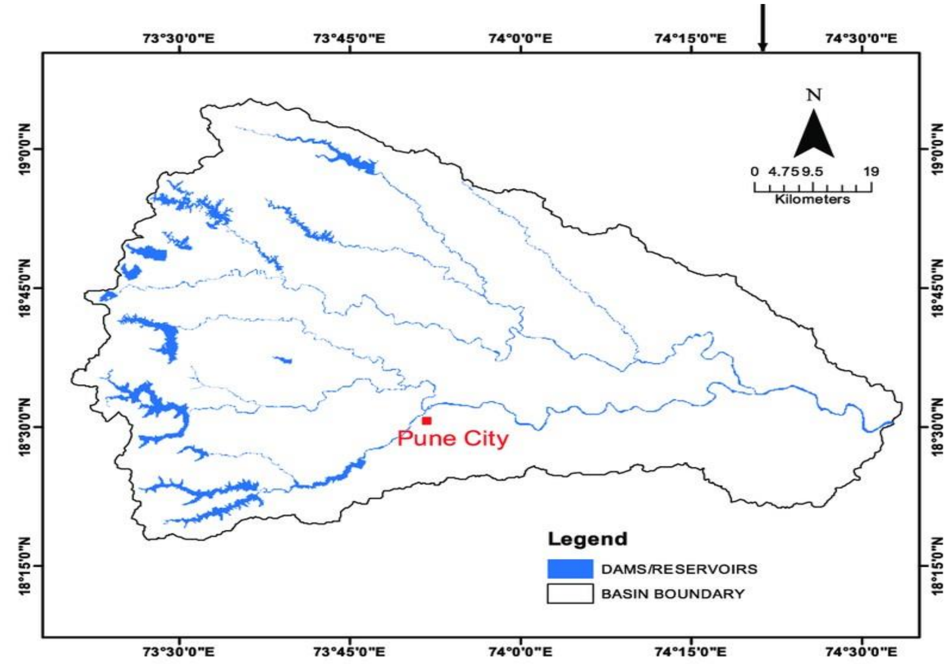


Fig. 3 The Upper Bhima Basin

- In Western Maharashtra, the Upper Bhima Basin is a sub-basin of the Krishna River Basin. With a flow of 69.8 km^3 yr^{-1} and a drainage area of roughly 260000 km^2, the river has a flow of 69.8 km^3 yr^{-1}. The Upper Bhima sub-basin is one of the Krishna's twelve sub-basins, situated almost completely inside the Indian state of Maharashtra.
- The Upper Bhima's environment is extraordinarily complex in terms of both space and time. The basin receives an annual rainfall total of 872 mm^{-1}. During the monsoon season, which lasts from June to October, 80–90% of the annual rainfall falls in sporadic showers.
- Agriculture, which accounts for around 70% of the total land area in the Bhima Basin, is the primary source of water usage. The soils are primarily vertisols, which are typical of the region's geology and climate. Natural vegetation types include grassland, savanna, and grassy forest, which contain a high content of the expanding clay montmorillonite. The principal crops farmed in this area are sugarcane, sorghum, wheat, corn, millet, peanuts, fodder grass, and a variety of other horticulture crops.

- The basaltic aquifers of the Upper Bhima River basin are intensively exploited for small-scale agriculture in western India, but they are under rising demand and uncertainties regarding climate change implications. (Dipak R. Samal, et.al. 2014)

3.2 Data Collection:

The dataset was compiled from a variety of sources. In the initial dataset, there are five inputs and one output. Maximum Daily Temperature, Minimum Daily Temperature, Daily Rainfall from two distinct models, Daily Evapotranspiration, and Daily Soil Moisture are all included in this dataset. It was compiled using information from the IMD website (Pai D.S, et.al. 2014) and the Water Resource Information System (WRIS India) (A. K. Srivastava. Et.al. 2009). For testing purposes, only remote sensing data is used. The data was first generated in a gridded format before being averaged over the Bhima Upper Basin. The dataset was split into two portions for the training and testing stages in this study. The basic dataset has 880 instances with five inputs and one output. There are 880 instances in the original dataset, each with four inputs and one output. For the training and testing stages, respectively, 80% and 20% of the dataset are preserved at random.

3.3 Machine Learning Algorithms:

There are four Machine Learning algorithms have been used for the prediction of Soil Moisture.

3.3.1 Linear Regression:

Linear Regression is the one the most fundamental and the basic ways to model the variables by assuming linear relationship between the variables. It is the oldest subtype of supervised machine learning. It uses least squares method to minimize the error between predictions and original values. Linear regression is commonly used for Predictions, Forecasting and curve fitting. Simple linear regression is used when there is only one explanatory variable, multiple linear regression is used when there are more than one. Multiple linear regression is a specific example of generic linear models with just one dependent variable, and it is an extension of basic linear regression with more than one independent variable. For multiple linear regression, the fundamental model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} + \epsilon_i \tag{2}$$

For each observation, $i = 1, \dots, n$.

Polynomial Regression is another major subtype of linear regression which uses non-linear fit for regression analysis. When the response variable is non-linear, i.e., the scatter plot shows a non-linear or curved shape, polynomial regression is applied (Gatignon, Hubert, 2010). Polynomial Regression can be used with a single regressor variable (Simple Polynomial Regression) or many regressor variables known as Multiple Polynomial Regression (David G. Kleinbaum. 2013) The formula for a Second Order Multiple Polynomial Regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \tag{3}$$

Here,

β_1, β_2 are called as linear effect parameters.

β_{11}, β_{22} are called as quadratic effect parameters.

β_{12} is called as interaction parameter.

Our initial dataset consists 5 input parameters and 1 output parameters. Modelling 5 inputs in multiple polynomial equation cause higher complexities and can be very lengthy to produce. Such, form of equations can be shown into matrix form as below,

$$Y = \beta X + \varepsilon \tag{4}$$

Two major issues with Multiple polynomial regression are Multicollinearity and terms of higher degree. When there are multiple regression variables, there's a good chance they'll be interdependent on one another. In such cases, the regression equation computed does not properly fit the original graph due to the relationship among variables. While Higher degree terms do not make a significant contribution to the equation. This term must be ignored in order to avoid complexity in the solution, but this process can be lengthy and tedious because there is no other option than trial and error.

3.3.2 Support Vector Machine:

In 1992, Corinna Cortes and Vladimir Vapnik put forward a novel data mining technique which creates a set of hyperplanes in an infinite dimensional space. Hence, Vapnik's theory is used to formulate the SVM equations (C. Cortes, V. Vapnik 1995). SVM can be utilized for both regression and classification problems. SVM has two broad classifier which are soft-margin hyperplane and hard-margin classifier. Hard margin classifier is used if training data is linearly separable and soft margin classifier is used when data is non-linear in nature.

Here, training dataset can be assumed as $\{(x_1,y_1),\ldots,(x_l,y_l)\}$ where x_i and y_i belongs to R^n

x_i represents the input parameter while y_i represents output parameter. The following equation can be solved to find the regression equation.

Minimize $$\frac{1}{2}||w||^2 + C\sum_{i=0}^{l}(\xi + \xi^*) \qquad (5)$$

Subject to $$\begin{cases} y_i - b - \langle w, x_i \rangle \leq \epsilon_i + \xi \\ b + \langle w, x_i \rangle - y_i \leq \epsilon_i + \xi^* \\ \xi_1, \xi_i^* \geq 0 \ i = 1, \ldots, l \end{cases} \qquad (6)$$

w denotes a normal vector, b denotes a scalar quantity, C denotes a regularization constant, ϵ is the insensitive loss function, and the slack variables. ξ, ξ^* denote the size of the excess deviation, Upper and lower deviations, respectively.

There are three stages in solving this equation. Primal formulation is turned into dual formulation by suitable manipulations. The dual problem then becomes an objective function for quadratic programming. For nonlinear functions the generalized equation has been reviewed by Shawe-Taylor and sun. In this generalized equation input parameter x_i has been replaced by $\phi(i)$.

$$f(x) = b + \sum_{i=1}^{l}(a_i - a_i^*)k(x_i, x) \qquad (7)$$

$$k(x_i, x) = \langle \phi(x_i), \phi(x) \rangle \qquad (8)$$

$k(x_i, x)$ is a kernel function. A kernel function is a method for taking data as input and transforming it into the format needed for processing. The term "kernel" is used because the

Support Vector Machine uses a set of mathematical functions to provide a window to manipulate data. Radial basis function, Polynomial, linear and sigmoid are some of the well-known kernel functions used for the project. For example, of radial basis function can be written as

$$k(x_i, x) = \exp\ (\gamma(\|x - x_i\|^2) \tag{9}$$

Selection of kernel and other parameters such as $C, \gamma\ and\ \epsilon$ have huge effect on accuracy.

3.3.3 Decision Tree Regression:

Regression and classification Machine-learning methods for constructing prediction models from data are known as trees. The models are created by partitioning the data space in a recursive manner and fitting a simple prediction model to each partition. A decision tree is a data structure in the form of a tree with an arbitrary number of nodes and branches at each node. An internal node is a node with outgoing edges. Leaves are nodes that are not connected to each other. An internal node concerning a specific function divides the instance used for regression or classification into two or more groups. In the training stage, the values of the input variable consider a specific function (John Shawe Taylor and Sun. 2011).

Here, training dataset can be assumed as $\{(x_1,y_1),\ldots,(x_l,y_l)\}$ where x_i and y_i belongs to R^n. x_i represents the input parameter while y_i represents output parameter. The following equation can be solved to find the regression equation. Let df is a feature variable and tr be the threshold value. Let t and $\gamma = (\ df, tr_t)$ be a node and candidate split respectively.

$$Q_1(\gamma) = x, y|x_{df \leq tr_t} \tag{10}$$

The left side of the decision tree, Q_1 , is found by splitting the data into γ candidate splits, as shown in the above equation.

$$Q_r(\gamma) = x, y|x_{df > tr_t} \tag{11}$$

In above equation Q_r is the right side of the decision tree. Which is found by splitting the data in γ candidate splits.

The quality of the estimates may be influenced by a number of factors in decision tree regression. The fitness function, tree depth, split sample, leaf sample, and feature number are only a few of these factors. 'Decision trees' provide you the freedom to deal with a wide range of answer kinds, including categorical and numeric data and Structure of tree can be visualized. In this project we have five input parameters and one output parameter which creates a highly complex tree. One of the most significant disadvantages of decision tree regression is that it might be unreliable, since slight changes in the data might result in a different tree.

3.3.4 K-Nearest Neighbors:

In 1951, Joseph Hodges and Evelyn Fix came with regression technique of non-parametric classification which an item is categorized by a majority vote of its neighbors, with the item allocated to the most common class among its k closest neighbors (Wei-Yin Loh. 2011). This work later expanded for regression. The function itself is approximated locally in k-NN classification, and all computation is postponed until the function is evaluated. Because this method depends on distance for classification, normalizing the training data can greatly increase its performance if the features represent various physical units or arrive in wildly different sizes.

Here, training dataset can be assumed as $\{(x_1,y_1),\ldots,(x_l,y_l)\}$ where x_i and y_i belongs to R^n. x_i represents the input parameter while y_i represents output parameter. Now, Euclidian Distance function is used to simply calculate the distance between various points.

$$distance((x,y),(a,b)) = \sqrt{(x - a)^2 + (y - b)^2} \qquad (12)$$

The procedure will locate the k-nearest neighbors of the data point for a given value of K, and then assign the class to the data point based on which class contains the most data points out of all classes of the K neighbors. The ideal value for k is determined by the data; in general, bigger values of k lessen the influence of noise on classification while making class borders less apparent. Various heuristic strategies, such as cross-validation, can be used to find an appropriate k. The input x is assigned to the class with the highest probability once the distance is computed.

$$P(y = j|X = x) = \frac{1}{K}\sum_{i \epsilon A} I(y^{(i)} = j) \qquad (13)$$

Above simple explanation K-NN holds valid for three dimensional problems but in project, input dataset is multi-dimensional. A distinct phenomenon occurs in large dimensions: the ratio between the nearest and furthest points approaches 1, implying that the points become evenly far from each other. As a result, we've turned our attention to Approximate Nearest Neighbor Search, which, in exchange for some precision, speeds up the procedure for higher-dimensional data. With a high likelihood, it leads to a decent approximation of the precise nearest neighbor.

3.4 Data Transformation:

Accuracy of Machine Learning Algorithms depend on various factors and quality of dataset provided is the most important one. Dataset often contains anomalies, missing points and significant outliers. Besides this when different parameters have different units and different range of values, errors tend to be maximized. Machine Learning algorithms are inherently bad at processing data in different range. In such case, it automatically assumes the higher weightage for higher values even though it has little impact on output. To avoid this problem, 'Feature Scaling' is done. Aim of 'Feature Scaling' is to align all the parameters in nearest range possible without disturbing their original structure. There are two major types of feature scaling which are Standardization and Normalization.

Normalization is a technique for transforming features such that they are on the same scale. The new point is computed as follows:

$$X_n = \frac{X - X_{minimum}}{X_{maximum} - X_{minimum}} \tag{14}$$

Another scaling strategy is standardization, in which the values are centered around the mean with a single standard deviation. As a result, the attribute's mean becomes zero, and the resulting distribution has a unit standard deviation.

$$X_n = \frac{X - \mu}{\sigma} \tag{15}$$

Where μ is mean of a selected row and σ is standard deviation. To yield better prediction results, standardization has been used in the project.

3.5 Feature Extraction:

Feature extraction is another major technique is used to improve the accuracy of the final yield. There is no common path for feature extraction but it varies with the type of dataset, objective of prediction and No. of dimensions. The common goal of feature extraction is to reduce the unwanted rows and add new features which can improve accuracy. In this project, all features are directly or indirectly depending on weather of the location. This gives rise to significant chance for seasonality. There are various graphs and mathematical functions to understand the effect of seasonality on the final yield. Besides this simple correlation values can be used to determine if certain feature contributes to the output or prediction of soil Moisture. Moving Averages over Shifting Horizon (MASH) is an effective tool used for trend analysis. Moving Averages with shifting horizon is nothing but calculation of simple moving Averages (SMA) with horizon shifting in forward direction in order to analyses trend.

Simple Moving Averages (SMA) is simple unweighted calculation average of last few datapoints and then moving in either forward or backward direction.

$$SMA_t = \frac{x_{n-t+1}+x_{n-k+2}+\ldots+x_n}{t} \tag{16}$$

$$SMA_t = \frac{1}{t}\sum_{i=n-t+1}^{n} p_i \tag{17}$$

Generally, moving averages are calculated for 2 to 4 steps forward or backward direction.

3.6 Model Evaluation:

Model Evaluation is a final step in the Machine Learning computations. There are various ways of model evaluation and it depend on type of machine learning algorithm, dimensions of datasets etc. Generally, machine learning algorithms are used for either classification or regression or both. For the classification problems, 'confusion matrix' is commonly used for evaluation. Confusion matrix is 2×2 matrix which contains True Positives, True Negatives, False Positives and False Negatives. Performance matrices such as precision, recall and specificity are also used to evaluate the accuracy of classification results.

Scope of our project is limited to regression analysis. Regression results are evaluated in completely different way. R squared value, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) are commonly used for evaluation

1. R squared Value: R-squared (R^2) is a statistical metric that quantifies the percent of variation explained by an independent variable or variables in a regression model for a dependent variable. There are several steps involved in calculating R-squared. This frequently involves using data points from dependent and independent variables to identify the best fit line, which is commonly done using a regression model. Then subtraction of the actual values from the projected values and square the results. This produces a list of squared errors, which is then added together to equal the unexplained variance. Total Variance can be calculated by subtracting total average value from each of given values.

   $$R^2 = 1 - \frac{Unexplained\ Variation}{Total\ Variation} \tag{18}$$

2. Mean Absolute Error (MAE): Mean Absolute error is the one of the accurate and simplest measure of accuracy. First step is to subtract actual value from predicted value. This would produce list of absolute errors. Mean of these errors by applying mod will give mean absolute error.

   $$MAE = \frac{\sum_{t=1}^{k} |y_t - x_t|}{k} \tag{19}$$

3. Root Mean Square Error (RMSE): Root mean square error is calculated by taking square root of Mean Square error. Square error is calculated by taking square of absolute difference between predicted value and original value. Average of theses squared errors will give mean square error.

   $$RMSE = \sqrt{\frac{\sum_{t=1}^{k} \left|y_{t-x_t}\right|^2}{k}} \tag{20}$$

CHAPTER-4: RESULTS AND ANALYSIS

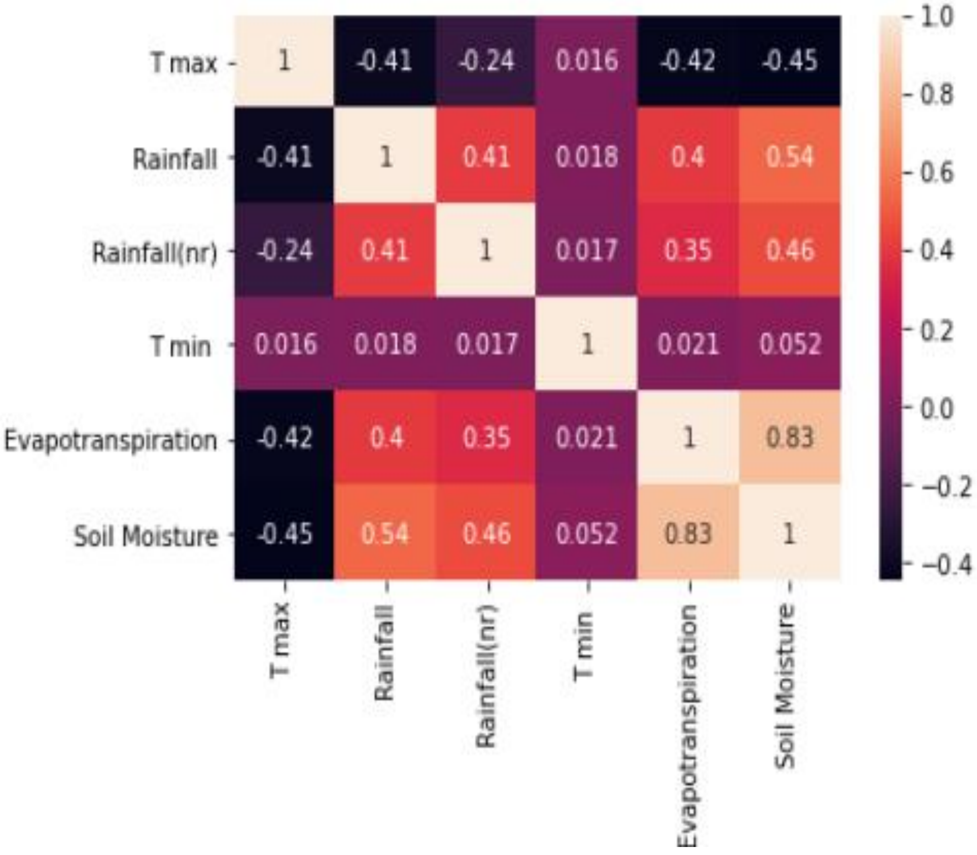
4.1 Data Structure:

Section provides a brief analysis all datasets that has been studied and considered for experiments. These simulations have been performed in Python programming language which operated in 64-bit Windows 10 operating system and 8 GB installed RAM. System has 2.5GHz processor.

There altogether five inputs and an output in initial dataset. Inputs consists of two datasets of rainfall, two datasets of temperature and one dataset of evapotranspiration.

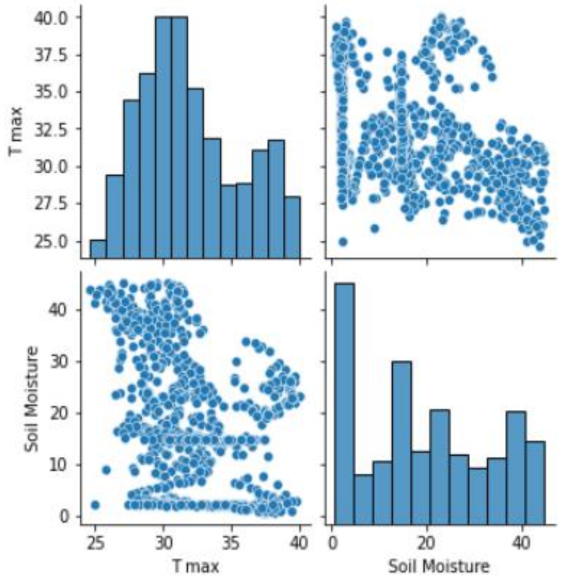
I. Correlation Heatmap:

Fig. 4 Correlation Heatmap

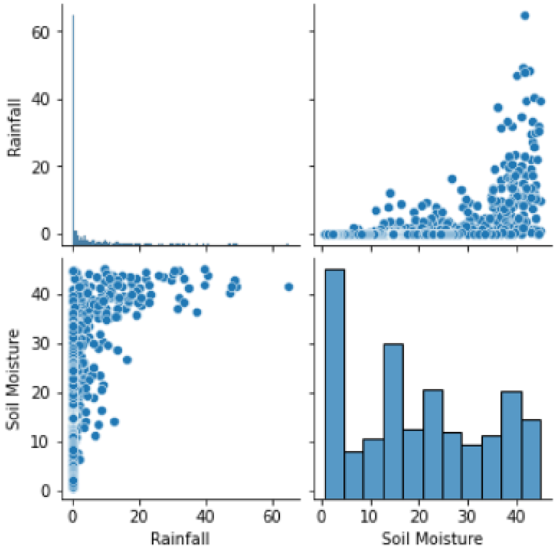


II. Distribution Plots:

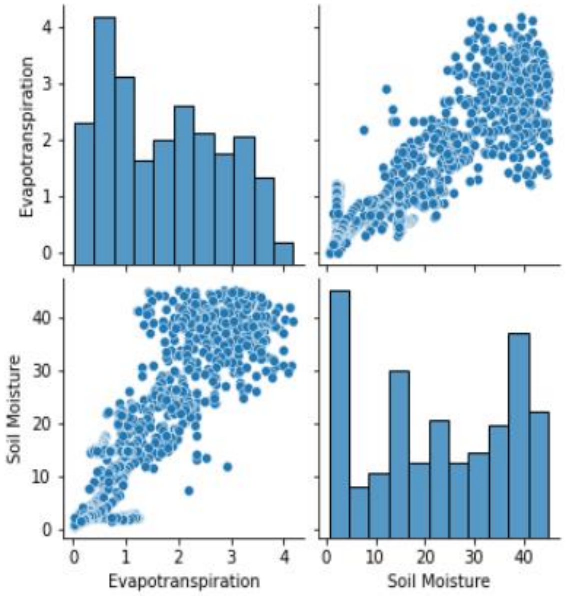
1) Soil Moisture Vs. $T_{max.}$ (Fig. 5)



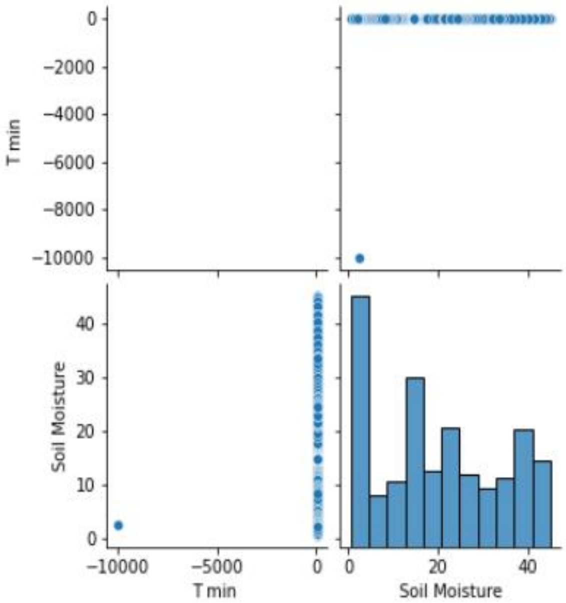
2) Soil Moisture Vs. Rainfall (Fig. 6)



3) Soil Moisture Vs. Evapotranspiration (Fig. 7)



4) Soil Moisture Vs. T_{min} (Fig. 8)



- In above figures, First Quarter contains the Histogram of the parameter tested against Soil Moisture. Second and Third Quarter contains variation of parameters w.r.t soil moisture and Variation of Soil Moisture w.r.t to parameter. In Fourth Quarter, it is histogram of soil Moisture Data.
- In Fig.5, It visible from histogram that Maximum temperature is approximately normally distributed. Variation of soil Moisture w.r.t max. temperature is clearly complex and hap hazardous. Since correlation between Soil Moisture and Max. temperature is -0.45, which shows a moderate impact of temperature on soil moisture. Relation between Maximum temperature and Soil Moisture has very high significance compared to relation between Minimum Temperature and Soil Moisture.
- In fig. 6, it hard to visualize the histogram of rainfall data because there is very high no. of zero rainfall days. Though Soil moisture and Rainfall have moderately high correlation value (0.54), it is hard to visualize with scatter plot since there is clear split between non-rainy days and rainy days. There is little difference between corr. (Max. Temp, Soil moisture) and corr. (Rainfall, Soil Moisture) but that is mostly because of higher frequency of non-rainy days. In order to obtain results with the higher accuracy, Seasonality needs to be understood very clearly.
- In Fig. 7, it is clearly visible from histogram that evapotranspiration has uniform distribution over time. Soil Moisture has linear dependance on evapotranspiration. Evapotranspiration has high effect on soil moisture but vice versa isn’t true. Correlation between evapotranspiration and soil moisture is very high that is 0.83.
- From Penman’s equation, evapotranspiration has huge dependance on solar radiations, Vegetation, saturation Vapor pressure and wind Velocity. This examination of datasets shows very high correlation between soil moisture and evapotranspiration that is 0.83. It implies high weightage of evapotranspiration for prediction of soil moisture as they both are dependent on Wind Velocity, Vegetation, radiation and Saturated Vapor pressure.
- From fig. 8, minimum temperature has negligible effects on variation of soil moisture. In order to obtain more accurate results, it is necessary to exclude Minimum temperature from the input data.

- From the histograms of soil moistures, it is clearly visible a partial periodicity in the data. This implies a strong relation of soil moisture with rainfall when during monsoon season. This can be better understood by feature extraction.

III. Feature Extraction and Data Transformations:

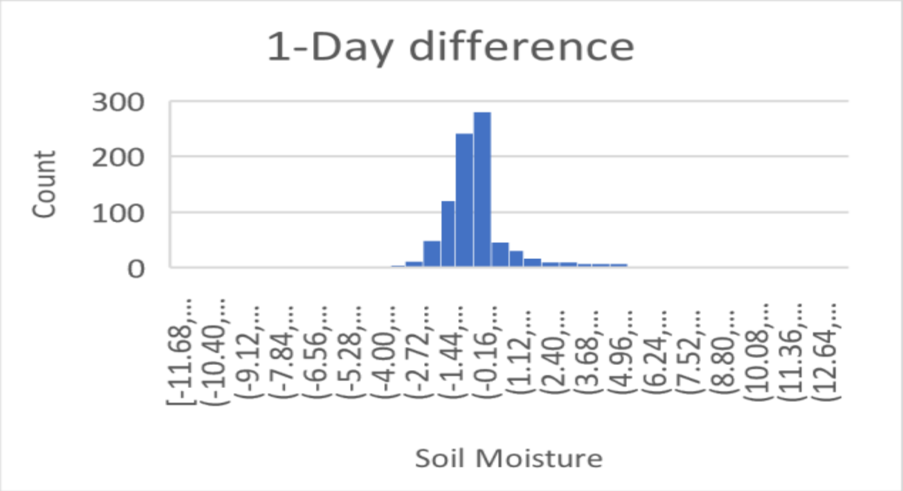


Fig. 9. 1-Day soil moisture Difference Histogram

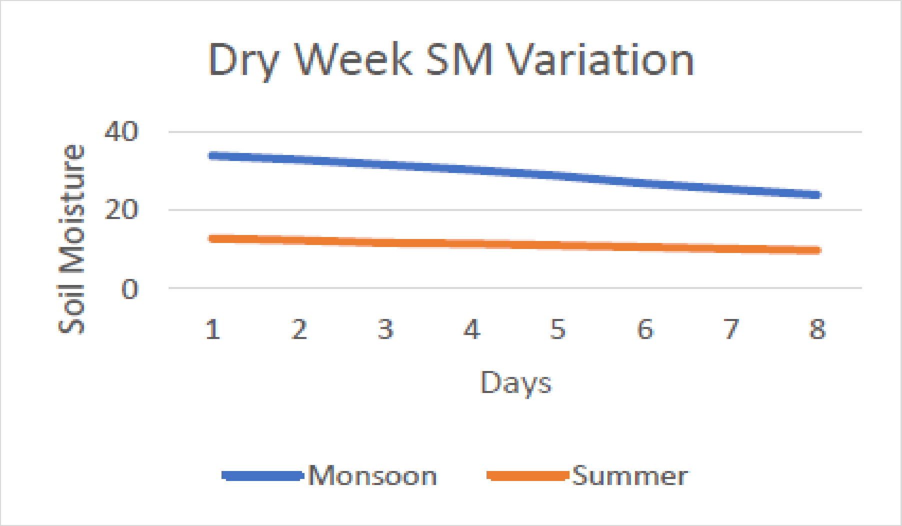


Fig. 10 Dry week Soil moisture Variation.

Comparison of Original Soil Moisture and 2-Days (previous) Moving Averages:

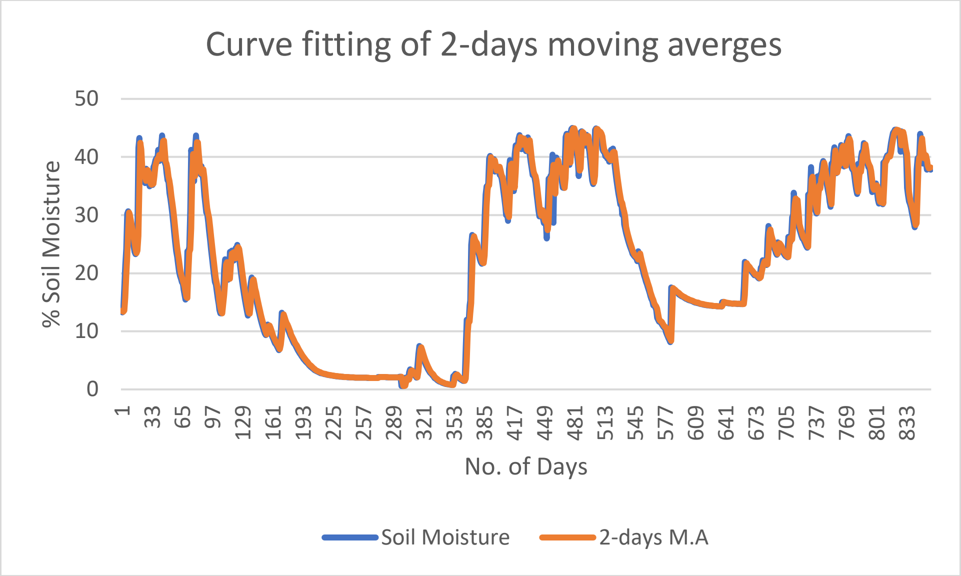


Fig. 11 Curve Fitting of 2-days MA

- In fig. 9, histogram shows difference between soil moisture of previous day and the same day. It is normally distributed hence maximum days have difference near zeroes. In fig. 10, difference between slopes for variation of soil moisture shows, higher variation days are days with higher rainfall. This effect is called seasonality. Fig. 11 clearly shows a previous 2-days Moving averages fit the data with extremely high accuracy and almost negligible outliers.
- Thus, Data has been modified with omitting the 2 columns which contains rainfall from NRSC model and minimum temperature. To improve the accuracy, an additional row with previous 2-days moving averages has been added to the dataset. Also, at the end to compensate weightage of all parameter's standardization has been done.

4.2 Training and Cross Validation:

To avoid overfitting, the training data is subjected to cross-validation. The training dataset is divided into ten subgroups for cross-validation which is known as 10-fold cross validation. One of the 10 subsets is used in the validation method, while the others are trained on the fitness function. In the training step, experiments are repeated ten times for each maintained subgroup.

4.2.1 Linear Regression:

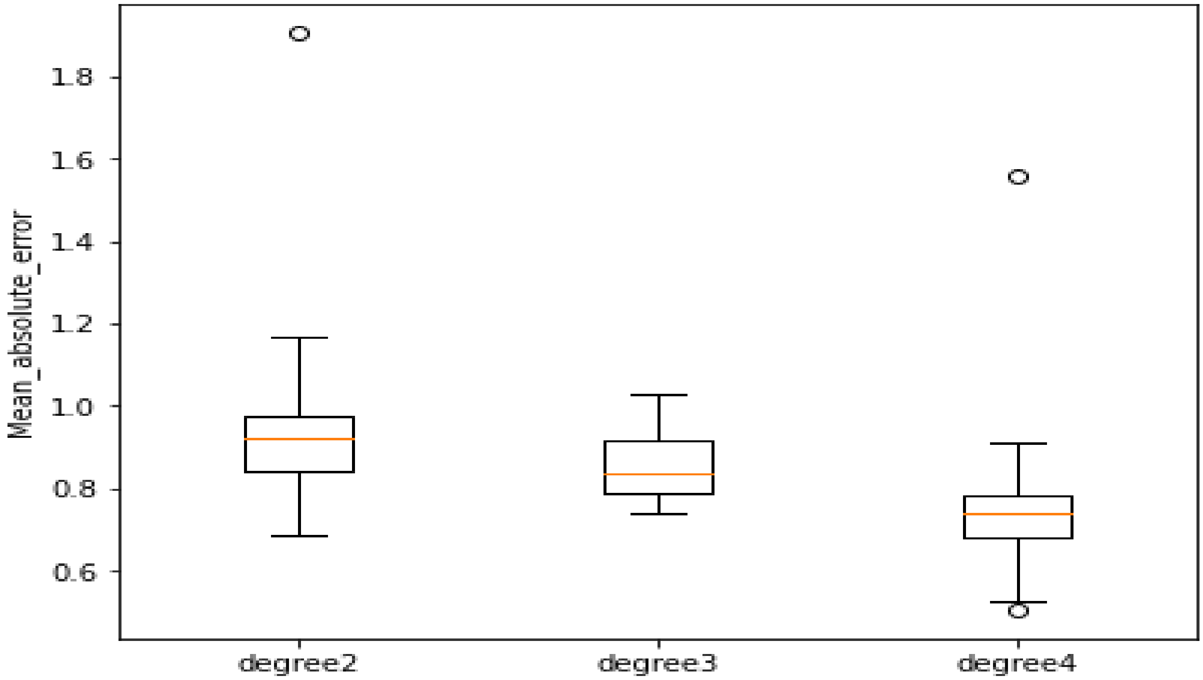


Fig. 12 Boxplot for MAE in linear regression

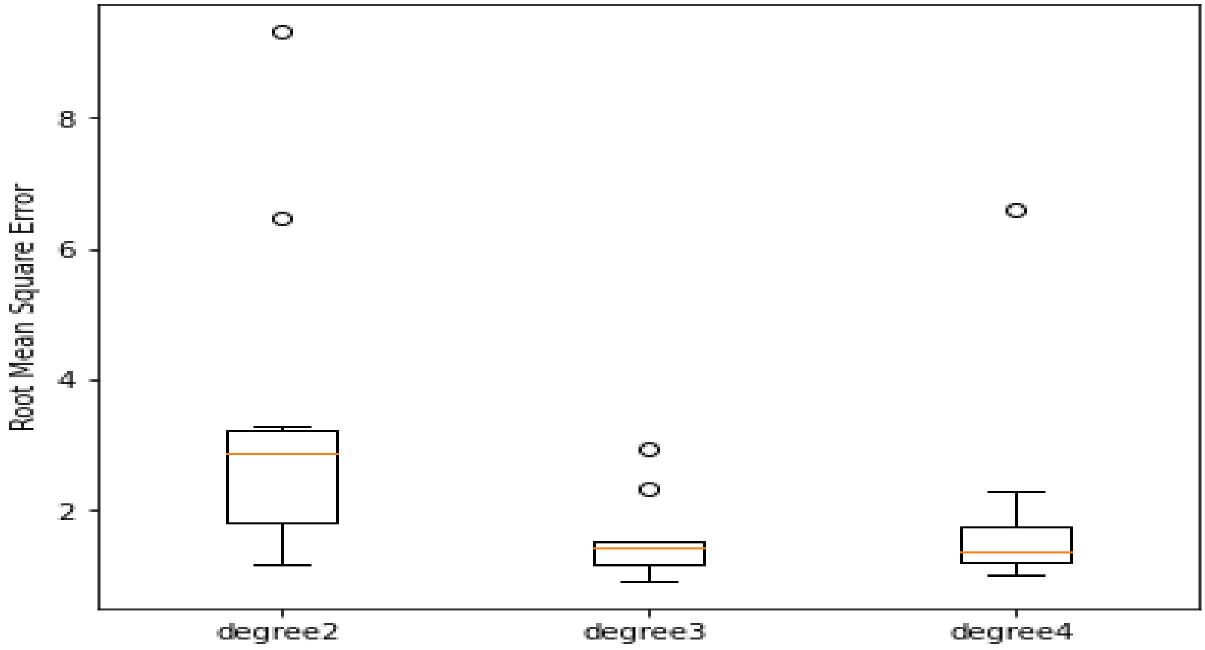
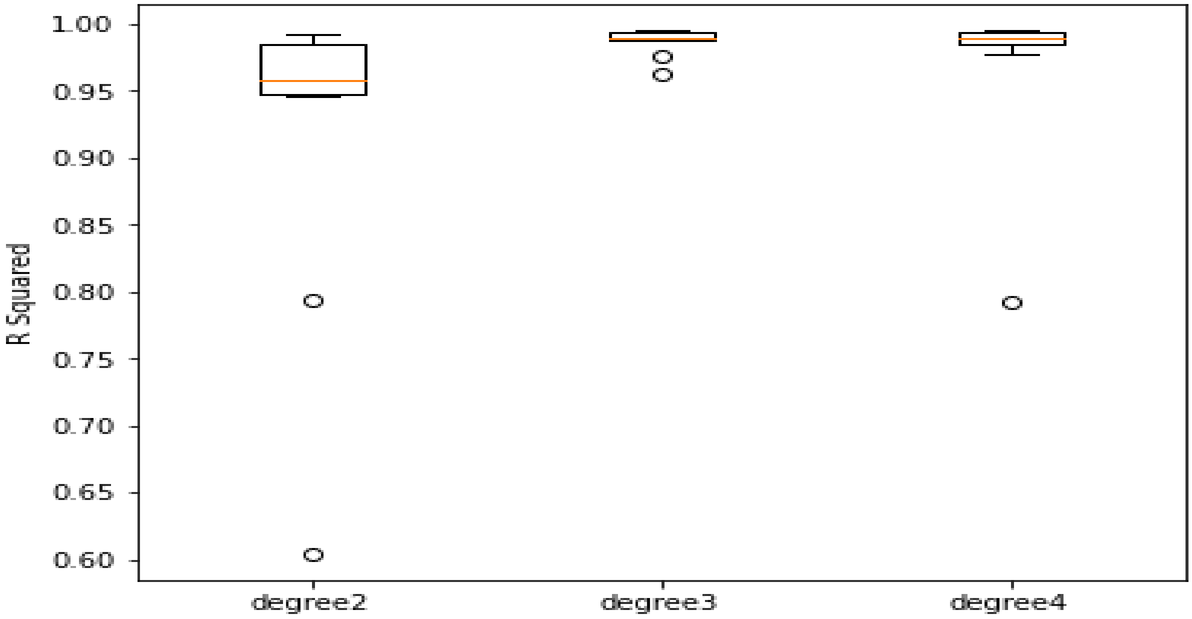


Fig. 13. Boxplot for RMSE in linear regression



Fig, 14 Boxplot for R squared value in linear regression

- In fig. 12, a ‘Box and Whisker’ plot shows a range in which ‘Mean Absolute Error’ varies for all 3 degrees of equation. 4th degree polynomial has lower median value than others but it has one significant outlier. ‘Mean Absolute Errors’ are 1.001, 0.856 and 0.792 for 2nd, 3rd and 4th degree equations respectively. Value of variances are 0.106, 0.008 and 0.0786 for 2nd, 3rd and 4th degree equations respectively. As degree increasing, average value of MAE decreasing but this may cause overfitting. Though, MAE value for 3rd degree is slightly higher than 4th degree, variance is low. Hence MAE is closely spread and probability of significant outliers is low. Thus 3rd degree equation is well suited for datasets in order avoid high bias and high variance.
- In fig. 13, a ‘Box and Whisker’ plot shows a range in which ‘Root Mean Square Error’ varies for all 3 degrees of equation. 4th degree polynomial has lower median value than others but it has one significant outlier. 1 and 2 outliers observed in RMSE values for 2nd and 3rd degree equations respectively. Average value of ‘Root Mean Square Error’ are 3.439, 1.951 and 1.455 for 2nd, 3rd and 4th degree equations respectively. Value of variances are 5.836, 0.345 and 2.544 for 2nd, 3rd and 4th degree equations respectively. As degree increasing, average value of RMSE decreasing but this may cause overfitting. Variance over RMSE values for 4th degree is much lower than 3rd degree and 2nd degree equation. Hence RMSE is closely spread and probability of significant outliers is low.

Thus 4th degree equation is well suited for datasets in order avoid high bias and high variance.

- In fig. 14, a ‘Box and Whisker’ plot shows a range in which ‘R Squared’ value varies for all 3 degrees of equation. R Squared value varies between 0 to 1 and higher value that is close to 1, indicates good fit. All 3 degrees of equations, have R Squared Median value close to 1 but 4th degree polynomial has slightly higher R Squared Value. Average value of ‘R Squared’ are 0.916, 0.987 and 0.969 for 2nd, 3rd and 4th degree equations respectively. Value of variances are 0.014, 0.00009 and 0.0035 for 2nd, 3rd and 4th degree equations respectively. Thus 4th degree equation is well suited for datasets in order avoid high bias and high variance

4.2.2 Support Vector Machines:

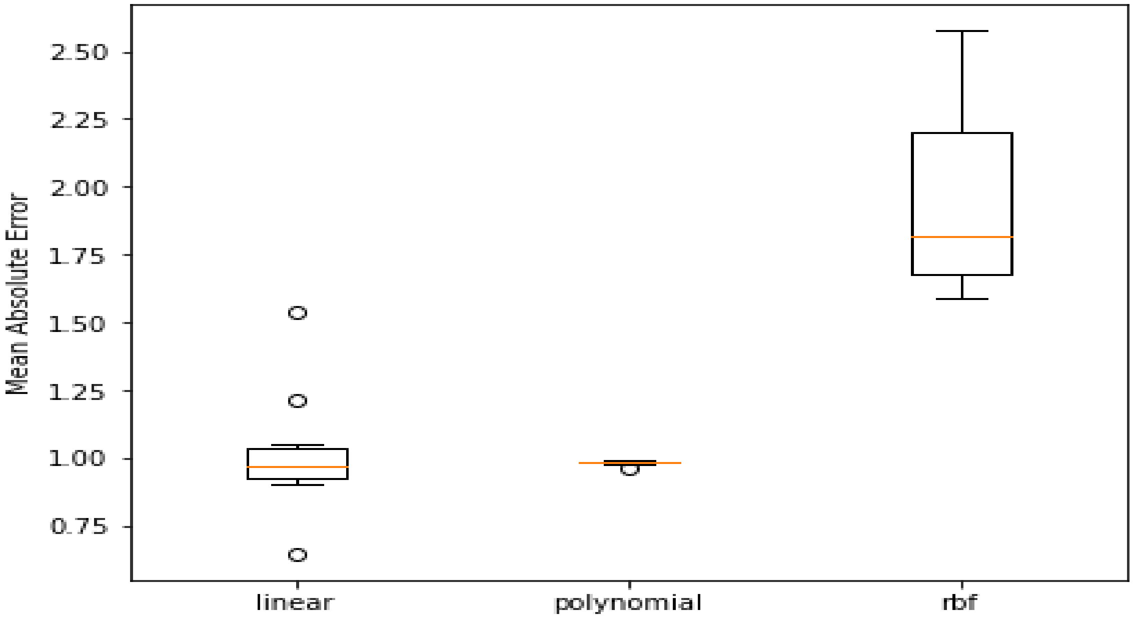


Fig 15 Boxplot for MAE in SVM

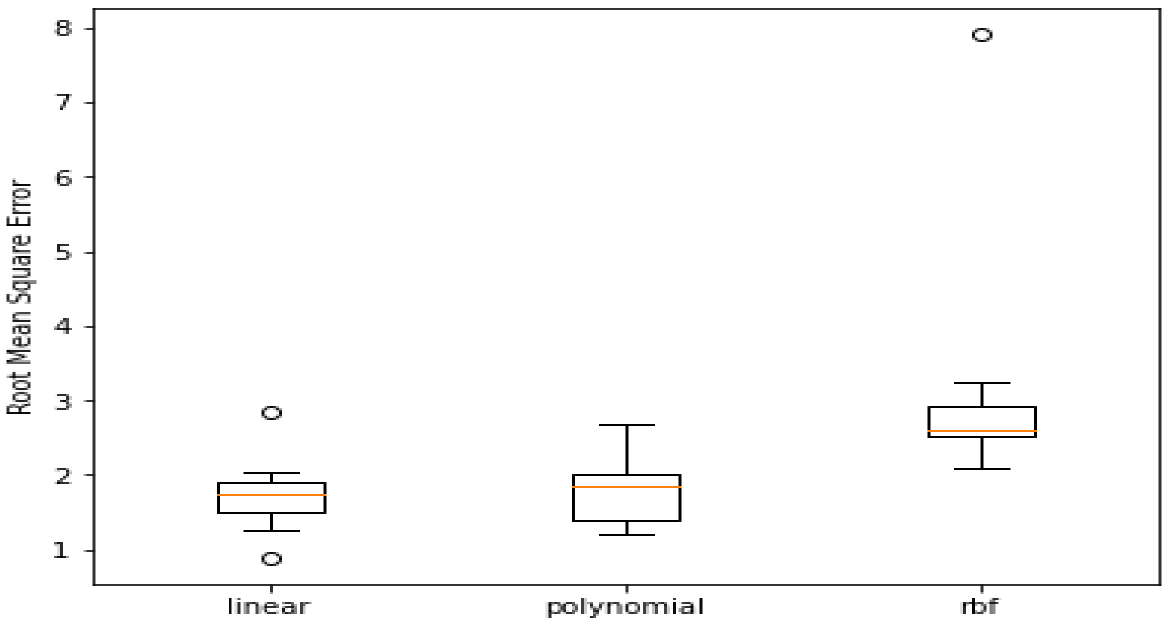


Fig 16 Boxplot for RMSE in SVM

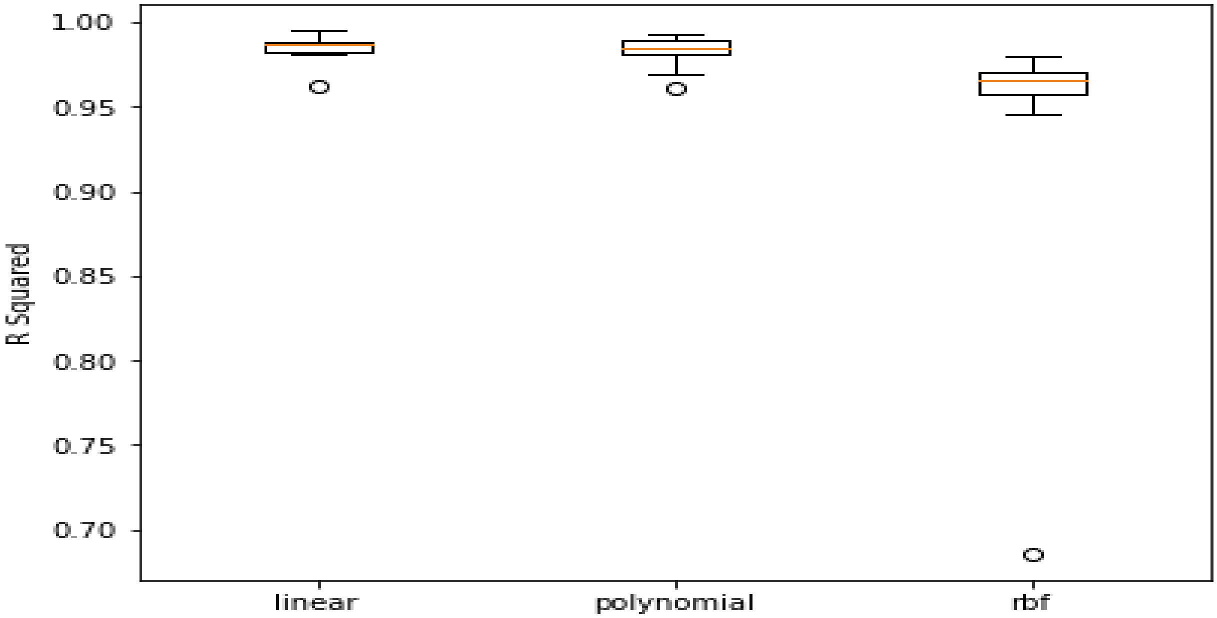


Fig. 17 Boxplot for R squared value in SVM

- In fig. 15, a ‘Box and Whisker’ plot shows a range in which ‘Mean Absolute Error’ varies for all 3 kernels. Linear kernel yields lower median value than others but it has three significant outliers. Average ‘Mean Absolute Errors’ are 1.217, 0.986 and 1.641 for linear, polynomial and rbf kernels respectively. Value of variances are 0.047, 7.513 and

0.041 for linear, polynomial and rbf kernels respectively. Though, MAE value for rbf kernel is slightly higher than linear kernel, variance is low. Hence MAE is closely spread and probability of significant outliers is low. Thus, linear kernel is well suited for datasets in order avoid high bias.

- In fig. 16, a ‘Box and Whisker’ plot shows a range in which ‘Root Mean Square Error’ varies for all 3 kernels. Linear kernel yields lower median value than others but it has two significant outliers. RMSEs are 1.728, 1.819 and 3.137 for linear, polynomial and rbf kernels respectively. Value of variances are 0.243, 0.211 and 2.63 for linear, polynomial and rbf kernels respectively. RMSE value and variance for linear kernel is lower than ‘rbf’ kernel and polynomial kernel. Thus, linear kernel is well suited for datasets in order avoid high bias.
- In fig. 17, a ‘Box and Whisker’ plot shows a range in which ‘R squared’ values vary for all 3 kernels. Linear kernel yields higher median value than others but it has one significant outlier. ‘R squared’ values are 0.984, 0.982 and 0.936 for linear, polynomial and rbf kernels respectively. Value of variances are 0.00007, 0.0009 and 0.00701 for linear, polynomial and rbf kernels respectively. ‘R squared’ value for linear kernel is higher than linear kernel and polynomial kernel and variance is low. Thus, linear kernel is well suited for datasets in order avoid high bias.

4.2.3 Decision Tree Regression:

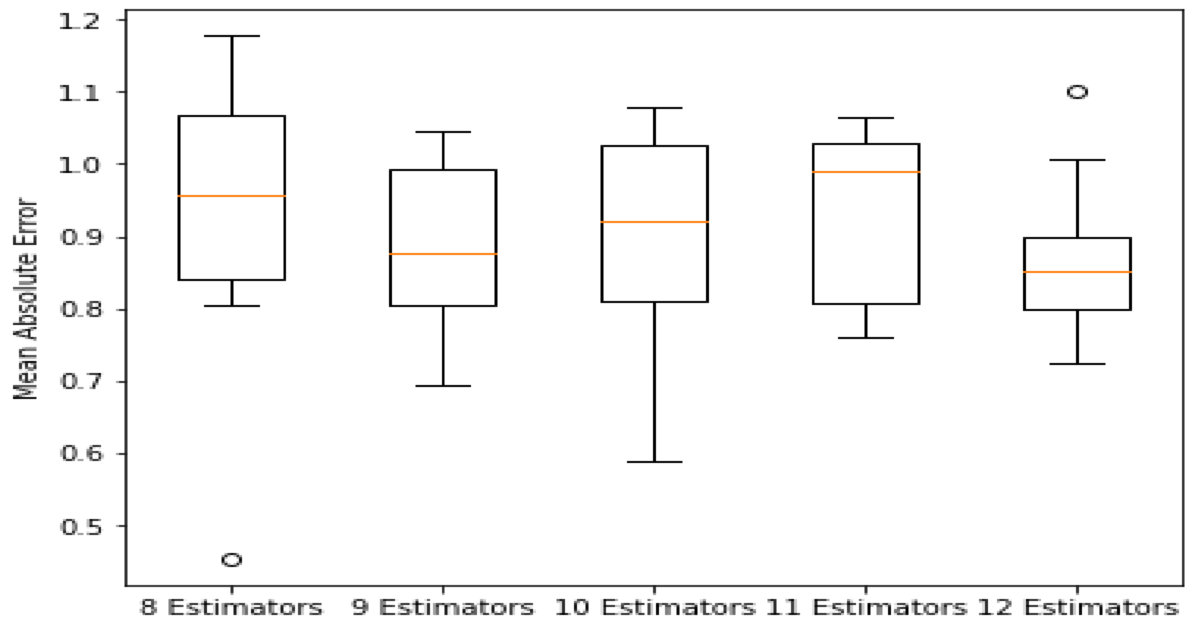


Fig. 18 Boxplot for MAE in DTR

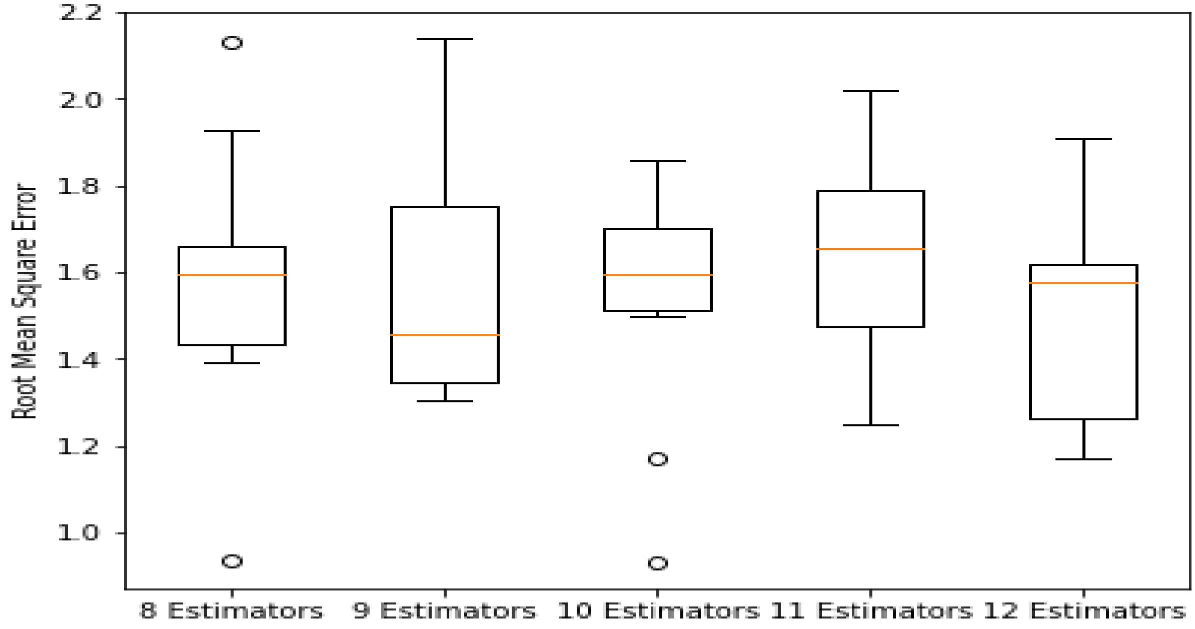


Fig. 19 Boxplot for RMSE in DTR

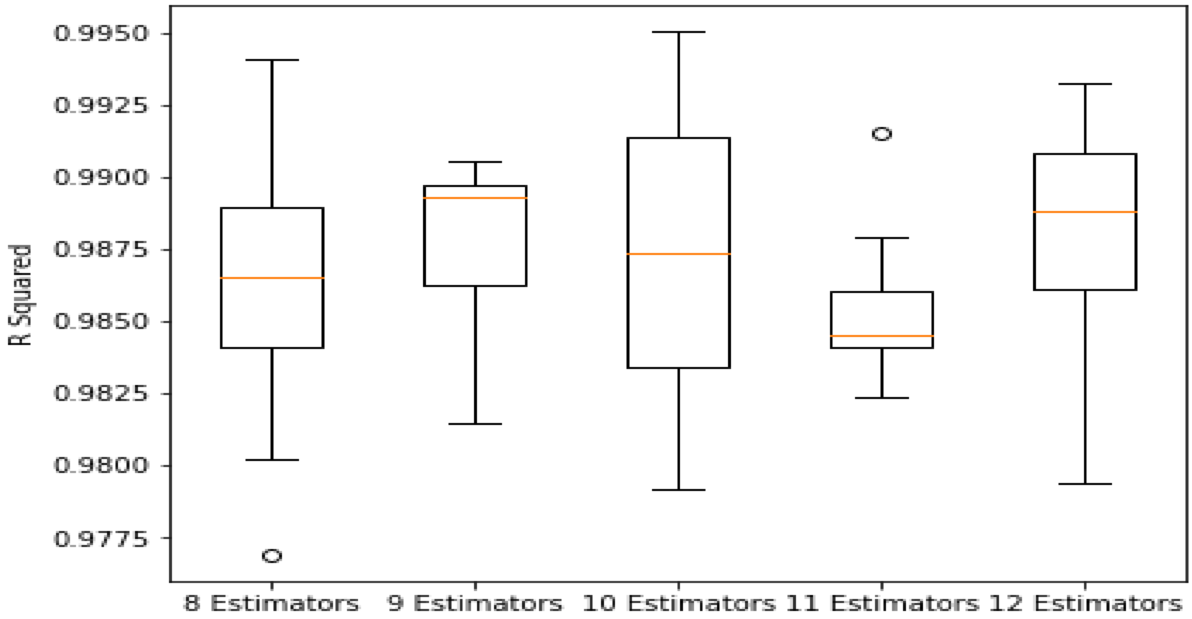


Fig. 20 Boxplot for R squared in DTR

- In fig. 18, a Box and Whisker plot shows a range in which 'Mean Absolute Error' varies for 5 different tree depths(estimators). A tree with '12 estimators' yields lower median value than others but it has one significant outlier. Average 'Mean Absolute Errors' are 0.926, 0.881, 0.897, 0.932 and 0.866 ranging from 8 to 12 estimators respectively. Value of variances are 0.0378, 0.0138, 0.0221,0.0139 and 0.0125 ranging from 8 to 12 estimators respectively. Thus, to avoid overfitting and high bias, a tree with 9 estimators is chosen.
- In fig. 19, a 'Box and Whisker' plot shows a range in which 'Root Mean Square Error' varies for 5 different tree depths(estimators). A tree with '9 estimators' yields lower median value than others with no significant outlier. 'Root Mean Square Error' are 1.580, 1.451, 1.542, 0.1.646 and 1.487 ranging from 8 to 12 estimators respectively. Value of variances are 0.0915, 0.0866, 0.0737,0.0563 and 0.0518 ranging from 8 to 12 estimators respectively. Thus, to avoid overfitting and high bias, a tree with 9 estimators is chosen.
- In fig. 20, a 'Box and Whisker' plot shows a range in which 'R squared' value varies for 5 different tree depths(estimators). A tree with '9 estimators' yields higher median value than others with no significant outlier. 'R squared' values are 0.9860, 0.9875, 0.9875, 0.9854 and 0.9887 ranging from 8 to 12 estimators respectively. Value of variances are 0.00002, 0.00001, 0.00002,0.00006 and 0.00001 ranging from 8 to 12 estimators respectively. Thus, to avoid overfitting and high bias, a tree with 9 estimators is chosen.

4.2.4 K-Nearest Neighbor:

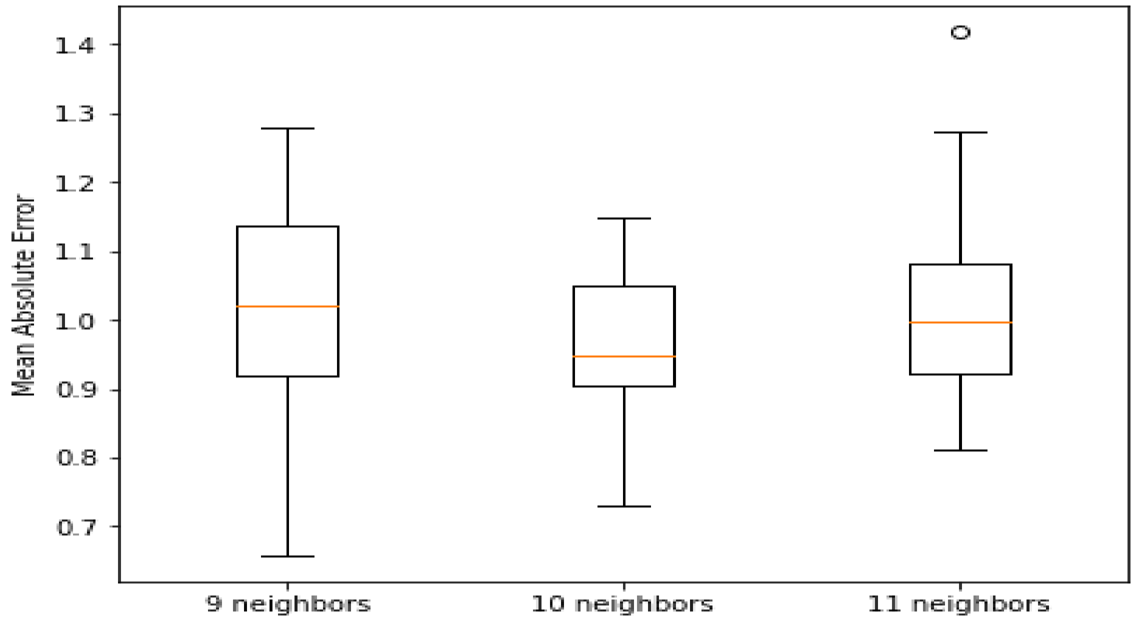


Fig. 21 Boxplot for MAE in KNN

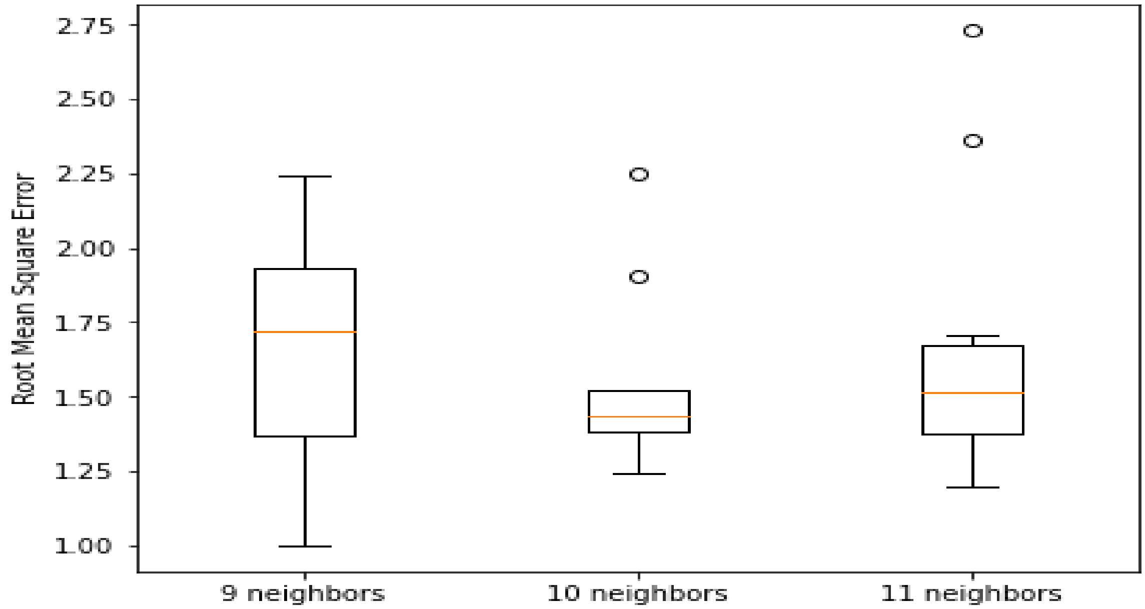


Fig. 22. Boxplot for RMSE in KNN

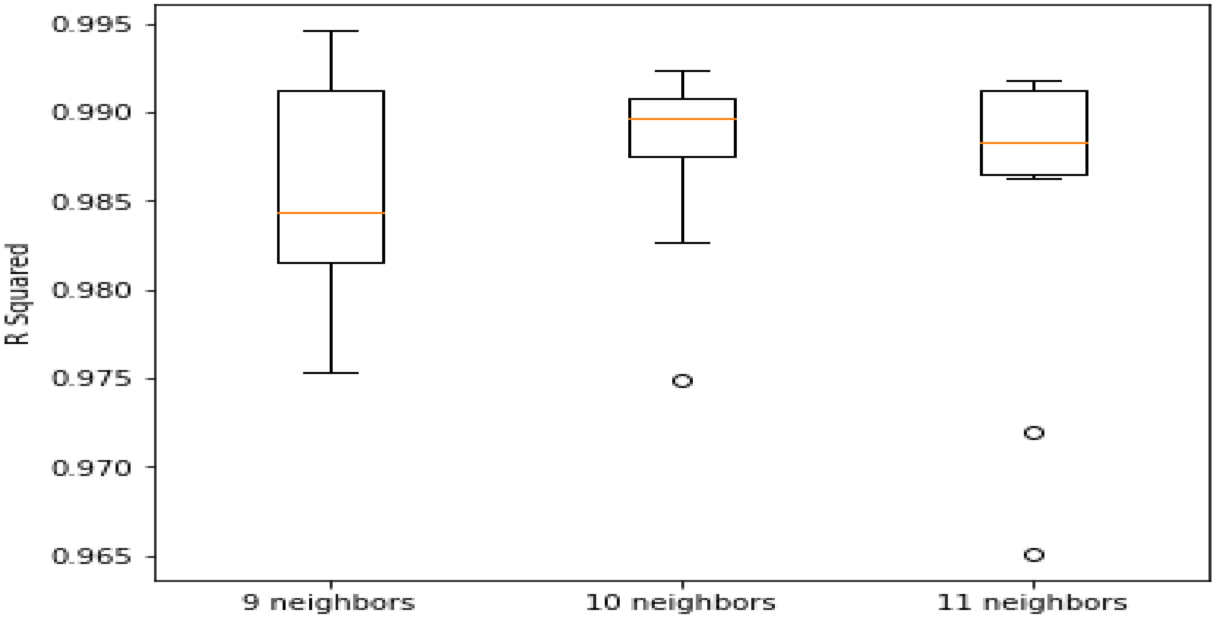


Fig. 23 Boxplot for R squared in KNN

- In fig. 21, a ‘Box and Whisker’ plot shows a range in which ‘Mean Absolute Error’ varies for all 3 different clusters. ‘Ten neighbours’ yields lower median value than others and it has now significant outliers. Average ‘Mean Absolute Errors’ are 1.006, 0.968 and 1.038 for respectively. Value of variances are 0.0309, 0.0144 and 0.0313 respectively. Thus, a cluster with 10 neighbours is well suited for datasets in order avoid high bias.
- In fig. 22, a ‘Box and Whisker’ plot shows a range in which ‘Root Mean Square Error’ varies for all 3 different clusters. ‘Ten neighbours’ yields lower median value than others and it has now significant outliers. Values of ‘Root Mean Square Error’ are 1.66, 1.54 and 1.674 for respectively. Value of variances are 0.1342, 0.0832 and 0.2160 for respectively. Thus, a cluster with 10 neighbours is well suited for datasets in order avoid high bias.
- In fig. 23, a ‘Box and Whisker’ plot shows a range in which ‘R Squared’ value varies for all 3 different clusters. ‘Ten neighbours’ yields higher median value than others and it has now significant outliers. ‘R Squared’ value are 0.989, 0.987 and 0.985 for respectively. Value of variances are 3.749, 2.482 and 7.603 for respectively. Thus, a cluster with 10 neighbours is well suited for datasets in order avoid high bias.

4.3 Testing:

Dataset has been split into 0.8 to 0.2 ration for training and testing respectively. Initial dataset has total 880 no. of rows. It has 5 inputs and 1 output. After feature extraction and Data transformation 2 columns were omitted and 1 more row was added. This tested data then compared with predicted data. Graphs for all Models and their estimators are given below.

4.3.1 Linear Regression:

- Comparison of curves

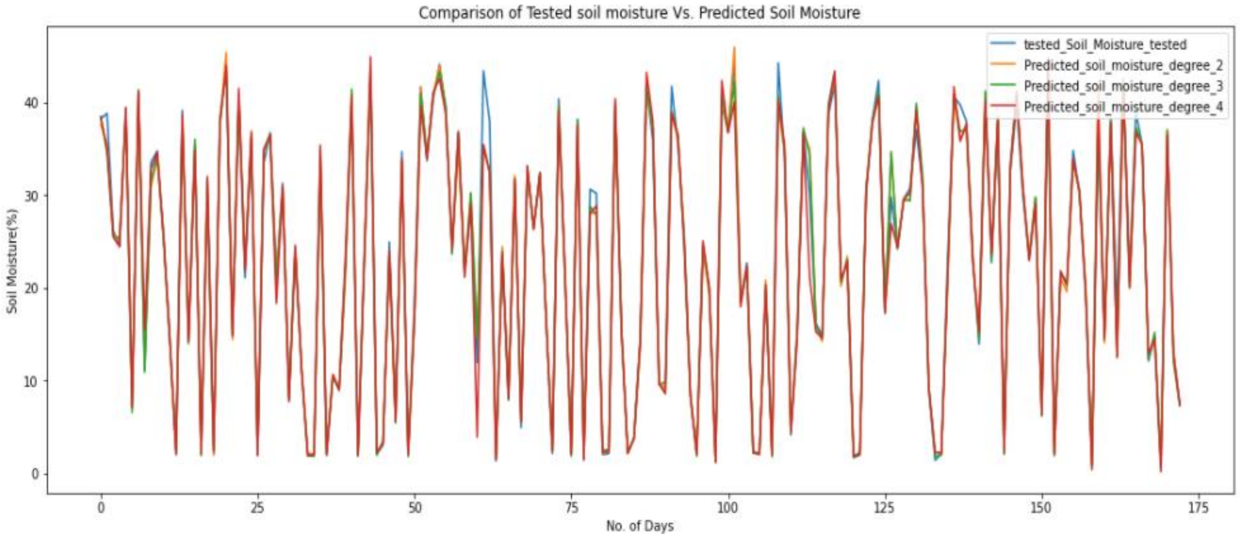


Fig. 24 Curve Comparison for LR.

- Error Calculations:

Degree of polynomial	Mean Absolute Error	Root Mean Squared Error	R Squared Value
2	0.969	3.022	0.901
3	0.921	1.605	0.922
4	**0.713**	**1.422**	**0.923**

Table No. 2 Errors in testing for LR

4.2.1 Support Vector Machines:

- Comparison of Curves

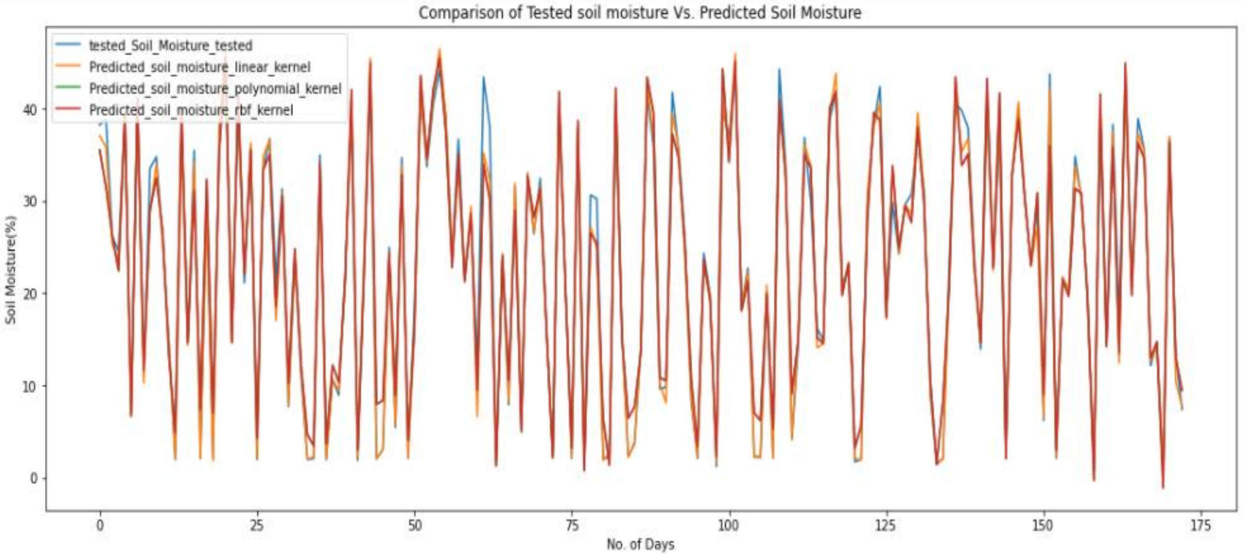


Fig. 25 Curve Comparison for SVM

- Calculation of Errors:

Kernel	Mean Absolute Error	Root Mean Squared Error	R Squared Value
Linear	**0.909**	**1.720**	**0.930**
Polynomial	1.921	2.671	0.926
Radial basis function	0.922	1.831	0.921

Table No. 3 Errors in testing for SVM

4.2.3 Decision Tree Regression:

- Comparison of curves

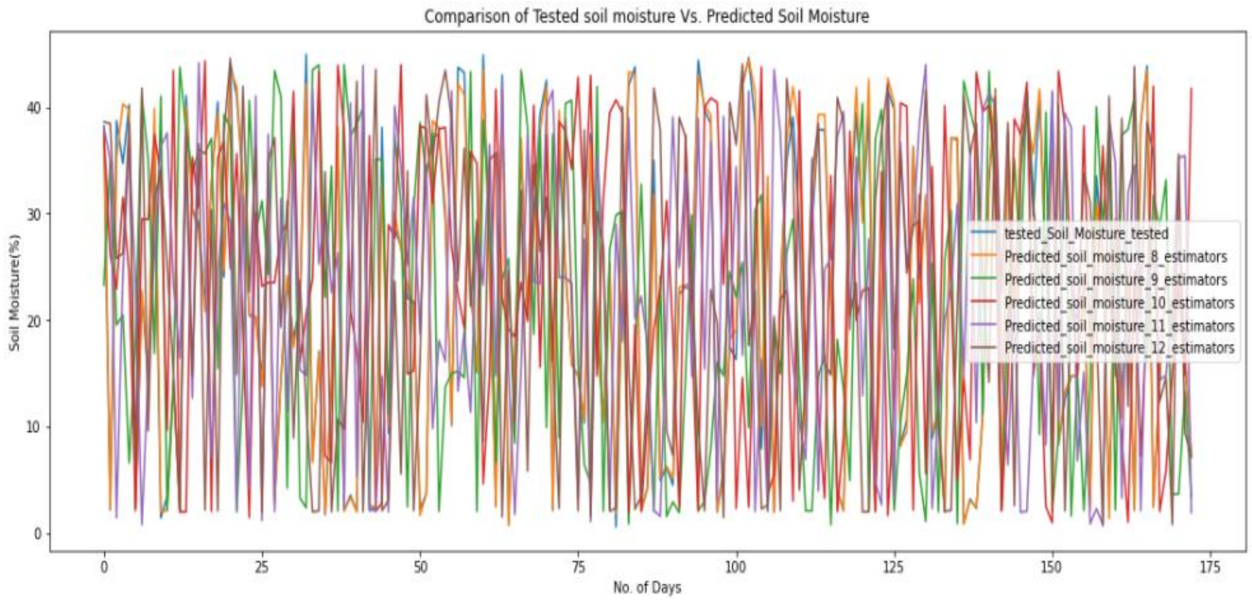


Fig. 26 Curve Comparison for DTR

- Calculation of Errors

No. of Trees	Mean Absolute Error	Root Mean Squared Error	R Squared Value
8	**0.652**	**1.347**	**0.985**
9	0.889	1.853	0.987
10	0.919	1.451	0.987
11	0.965	1.642	0.988
12	0.667	1.358	0.984

Table No. 4. Errors in testing for DTR

4.2.4 K Nearest Neighbors:

- Comparison of Curves

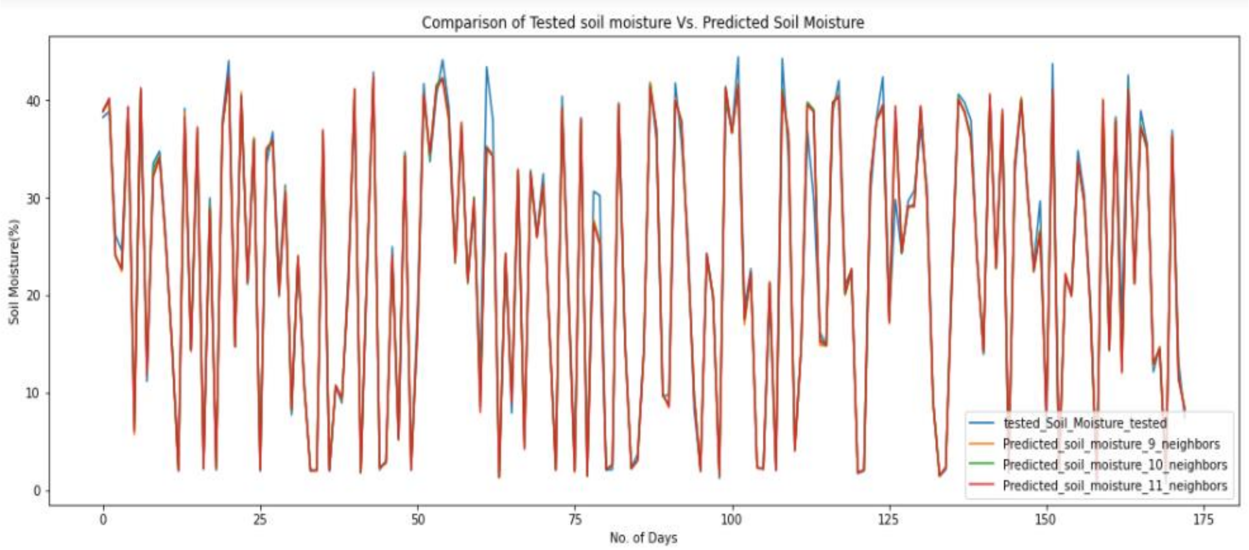


Fig. 27. Curve comparison for KNN

- Calculation of Errors:

No. of Neighbors	**Mean Absolute Error**	**Root Mean Squared Error**	**R Squared Value**
9	1.006	1.887	0.912
10	**0.885**	**1.213**	**0.931**
11	1.215	1.887	0.909

Table. No. 5 Errors in testing for KNN

CHAPTER-5: CONCLUSION

5.1. Conclusion and Future Scope:

- This Project Addresses prediction of soil moisture using four different supervised machine leaning techniques which are Linear Regression, Support Vector Machine, Discission Tree Regression and K Nearest Neighbors. There are three major input parameters are applied to model which are Rainfall, Temperature and Evapotranspiration. In addition, Historical Soil Moisture Data has been provided to model to improve the accuracy. The need of smart water management in irrigation, Sensor based irrigation, application of machine learning techniques in irrigation and water resources are range of topics revied in literature.
- Evaluation of machine learning model is a complex problem and there are variety of different ways in which model can be evaluated. Evaluation depends on goal of a project. Aim of the project is prediction of soil moisture with low bias and low variance. So, in the end model with less MAE and RMSE but high R squared value supposed to perform better. In four different methods, linear regression and decision trees performs much better than K-NN and SVM. A linear regression of degree 4 and decision tree with depth 8 performs more accurate than other estimators. Since Decision Tree is generally more preferred algorithm for problems with higher dimensions than linear regression. So, the most suitable algorithm for prediction of soil moisture is decision tree with depth 8. It satisfies all fitness criteria for MAE, RMSE and R squared value satisfactorily in both training and testing stage.
- Data Transformation and Feature Extraction have extremely high impact on model accuracy. One of the main reasons behind this, is seasonality of climatological data. Understanding of ‘data structure’ is key step successful feature extraction. Understanding, the background of variables helps to understand the relationship between the parameters which is clearly not visible. Soil Moisture is highly dependent on evapotranspiration but in practical scenario evapotranspiration has little daily variation. Same, Applies for the temperature. In case of rainfall though, despite of less correlation value sudden change in rainfall happens frequently and rainfall patterns are uncertain. This uncertainty causes drastic fluctuations in soil moisture which isn’t visible through correlation analysis. Moving Averages have been implemented to compensate this seasonality. All models performed better after feature extraction which is done by using

mathematical tool called MASH. Average accuracy for decision trees improved by 4 times after feature extraction.

- More work is required to increase the fitness criterion for the prediction of soil moisture. In this project, datasets used has daily variation and dataset is obtained from various remote sensing sources. For practical use in the irrigation, data should be either in real time or multiple instances in single day. A lack of strong ground truth data can restrict the capabilities of model in the same way that a lack of strong features might lead algorithm to perform badly. Thanks to rise of Internet of Things (IOT) in the field of water resources and irrigation which has made real time analysis possible. When sensors are linked to additional sensors detecting environmental factors such as temperature, pressure, and humidity, machine learning may be used to assist calibrate and rectify them. Application of sensor-based irrigation in farms should be enhanced with both sources of data that is remote sensing and ground-based sensors.
- Machine learning models are stochastic and they are not deterministic. This means these models do not understand nature of variables and physics behind them. Researchers needs to feed more features to make machine learn a nature of problem. It varies with application of model and parameters. Mathematically, machine learning models including ANN can be enhanced with the physics of processes. Further research needs to be conducted on algorithms to improve their understanding of physics and natural processes.

References:

1. Allen, Pereira, Richard G, Luis S, Howell, Terry A, and Jensen, Marvin E. 2011. Evapotranspiration information reporting: I. Factors governing measurement accuracy. *USDA-ARS / UNL* Faculty. 829.
2. Anghileri D., Pianosi, F., Soncini-Sessa, R. 2014. Trend detection in seasonal data: from hydrology to water resources, *Journal of Hydrology*.
3. Cortes C., Vapnik V. 1995. Support-vector networks, Machine. Learning. Mach Learn 20, 273–297.
4. Dabach Sharon, Naftali Lazarovitch, Jirka Simunek, Uri Shani. 2011. Numerical investigation of irrigation scheduling based on soil-water status. Springer-Verlag.
5. Davis S.L., M.D. Dukes, G.L. Miller, 2009, Landscape irrigation by evapotranspiration-based irrigation controllers under dry conditions in Southwest Florida. Agricultural Water Management.
6. Dimri Tripti, Ahmad Shamshad and Sharif Mohammad. 2020. Time series analysis of climate variables using seasonal ARIMA approach. J. Earth Syst. Sci. 129:149
7. Dhawal H and Mishra N. 2016. A Survey on rainfall prediction techniques; *Int. J. Comput. Appl*. 6(2) 1797–2250.
8. Dukes D Michael, Eric H. Simonne, Wayne E. Davis, David W. Studstill, Robert Hochmuth. 2003. Effect of sensor based high frequency irrigation on bell paper yield and water use. *Proceeding's 2nd International Conference on Irrigation and Drainage*, Phoenix, AZ May 12-15, pp. 665-674.
9. Fix E. and Hodges J.L. 1951. An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges. School of mathematical sciences, University of Bath, BA2 7AY U.K.
10. Grabow G. L., Ghali I. E., Huffman R. L., Miller G. L., Bowman D., and A. Vasanth. 2013. Water Application Efficiency and Adequacy of ET-Based and Soil Moisture–Based Irrigation Controllers for Turfgrass Irrigation. *Journal of Irrigation and Drainage Engineering Vol. 139, No. 2*.
11. Geerts Sam, Dirk Raes. 2009. Deficit irrigation as an on-farm strategy to maximize crop water productivity in dry areas. Division of Soil and Water Management, Celestijnenlaan 200 E, B-3001Leuven, Belgium

12. Gill. M. Kashif, Tirusew Asefa, Mariush W. Kemblowski, and Mac McKee. 2006. JAWRA Journal of the American Water Resources Association.
13. Granger, R.J., 1989. Evaporation from natural monosaturated surfaces. J. Hydrol. 111, 21–29.
14. Jiang H and Cotton W.R. Cotton. 2004 Soil moisture estimation using an artificial neural network: a feasibility study Atmospheric Science Department, Colorado State University, Fort Collins, CO 80523, USA.
15. K. P. Sudheer, A. K. Gosain, and K. S. Ramasastri. 2003. Estimating Actual Evapotranspiration from Limited Climatic Data Using Neural Computing Technique. *Journal of irrigation and drainage Engineering*. ISSN 0733-9437/ 2003/3-214–218.
16. Karthikeya H. K., Sudarshan K., Shetty S. Disha. 2020. Prediction of Agricultural Crops using KNN Algorithm. *International Journal of Innovative Science and Research Technology*. ISSN No: -2456-2165
17. Kite G.W, Droogers P. 1999. Comparing evapotranspiration estimates from satellites, hydrological models and field data. *Journal of Hydrology* 229 (2000) 3–18. 8
18. Khan Mohammad Sajjad and Coulibaly Paulin. 2006. Application of Support Vector Machine in Lake Water Level Prediction. *Journal of Hydrologic Engineering*, Vol. 11, No. 3, May 1, 2006. ©ASCE, ISSN 1084-0699/2006/3-199–205/$25.00.
19. Kleinbaum G. David. 2013. Applied Regression Analysis and other multivariable methods.
20. Kwame A L, Abdul-Aziz A R, Anokye M, , Munyakazi and Nsowah-Nuamah. 2013. Modelling and Forecasting rainfall pattern in Ghana as a seasonal arima process: The case of Ashanti region. *International Journal of Irrigation Engineering and Drainage*. 3(3) 224–233.
21. Narayanamoorthy A. 2003. Drip irrigation in India: can it solve water scarcity? Gokhale Institute of Politics and Economics. IWA Publishing.
22. P. Vinciya. 2016. Agriculture Analysis for Next Generation High Tech Farming in Data Mining", Anna University, Trichy, Tamilnadu, India.
23. Parag P., Bhagwat, M. Rajib, Multistep ahead river flow prediction using LS-SVR at daily scale. 2012. *Journal of Water Resource and Protection* 4 528–539.

24. Pai D.S., Latha Sridhar, Rajeevan M., Sreejith O.P., Satbhai N.S. and Mukhopadhyay B., 2014: Development of a new high spatial resolution (0.25° X 0.25°)Long period (1901-2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region; MAUSAM, 65, 1(January 2014), pp1-18.
25. Qi, Y, Huo, Z, Feng, S, Adeloye, AJ & Dai, X 2018, 'Prediction of Consumptive Use Under Different Soil Moisture Content and Soil Salinity Conditions Using Artificial Neural Network Models', *Irrigation and Drainage, vol. 67, no. 4*, pp. 615-624.
26. Raghavendra Sujay, N. Paresh Chandra Deka. 2014. Support vector machine applications in the field of hydrology: A review. Applied Soft Computing 19 (2014) 372–386.
27. R. Dipak. Samal, Shirish Gedam, Rama Nagarajan, 2015, GIS based drainage morphometry and its influence on hydrology in parts of Western Ghats region, Maharashtra, India. *Geocarto International, Taylor and Francis*. CSRE and IIT Bombay Mumbai, India.
28. Russell J. Qualls, Joshua M. Scott, and William B. DeOreo. 2002. Soil Moisture sensors for urban landscape irrigation: Effectiveness and reliability. *JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION*.
29. Srivastava A. K., Rajeevan M., Kshirsagar S. R. Development of High Resolution Daily Gridded Temperature Data Set (1969-2005) for the Indian Region. ATMOSPHERIC SCIENCE LETTERS Atmos. Sci. Let. (2009) DOI: 10.1002/asl.232
30. Taylor and Sun. 2011. A review of optimization methodologies in support vector machines. Volume 74, Issue 17,Neurocomputing.
31. Walison Alves B., Glauco De S. Rolim, Lucas E. De, O. Aparecida. 2017. Reference Evapotranspiration forecasting by artificial neural networks. *Journal of the Brazilian Association of Agricultural Engineering.*
32. Wei-Yin Loh. 2011. Classification and regression. Volume 1. International Statistical Review. Wiley Online Library.
33. Zanetti S. S. Sousa E. F., V. P. S. Oliveira, F. T. Almeida, and S. Bernardo, Estimating Evapotranspiration Using Artificial Neural Network and Minimum Climatological Data. *Journal of Irrigation and Drainage Engineering,* Vol. 133, No. 2, April 1, 2007. ASCE, ISSN 0733-9437/2007/2-83–89.