**Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize**

**Abstract**

In this project we are using machine learning and deep learning algorithms to predict future crop yield based on weather data such as temperature and rainfall. If farmers know the crop yield before sowing based on historical weather data, then he may take better decision. So, by employing machine/deep learning algorithms we can inform farmers about future crop yield. In proposed method we are using Irish Maize and Potato yield dataset to train all machine learning models and then these models can be used to predict future crop yield. In proposed method we are using random forest, SVR, **DNN, CNN, ANN and LSTM**. So, we have implemented all 6 algorithms on both datasets. To evaluate performance of each algorithm we are calculating MSE and R2 Score where MSE refers to mean square error (difference between TEST crop yield and predicted yield). R2 refers to correct prediction rate. So, for any algorithm MSE must be lower and R2 must be higher for better crop yield prediction.

**Keywords:** crop yield prediction, Mean square error, R2 value , RF, ANN, DNN LSTM.

**INDEX**

**CHAPTER 1: INTRODUCTION**

1.1Aim of study

1.2 Objective of study

1.3 Introduction

1.4 Machine Learning

1.5 Crop Management

**CHAPTER 2: LITERATURE SURVEY**

**CHAPTER 3: PROPOSED METHOD**

**CHAPTER 4: HARDWARE AND SOFTWARE REQUIREMENT**

**CHAPTER 5: RESULT**

**CHAPTER 6: CONCLUSION**

**CHAPTER 7: ADVANTGAES AND APPLICATIONS**

**APPENDIX**

PYTHON

**LIST OF FIGURES**

Figure1.1 Machine learning algorithms

Figure 1.2 Simple Artificial Neural Network

Figure 3.1 Flowchart

Fig.5.1 Dataset folder

Fig.5.2 dataset

Fig.5.3 Importing all require python packages

Fig.5.4 defining MINMAX and defining MSE

Fig.5.5 Read dataset function

Fig.5.6 training with SVR algorithm

Fig.5.7 Graphical representation

Fig.5.8 defining and training with LSTM

Fig.5.9 LSTM MSE and R2 values

Fig 5.10 training with ANN model

Fig.5.11 ANN MSE and R2 values

Fig.5.12 training with CNN

Fig.5.13 CNN MSE and R2 values

Fig.5.14 training with DNN

Fig.5.15 DNN graph

Fig.5.16 graph representation

Fig.5.17 plotting MIN, MAX and median

Table 1.1 Difference between supervised learning and unsupervised learning.

**CHAPTER 1**

**INTRODUCTION**

* 1. **Aim of study**

The aim of the study on "Crop Yield Prediction Using Machine Learning Models: Case of Irish Potato and Maize" is to develop and implement machine learning models to accurately predict crop yields for Irish Potato and Maize crops.

* 1. **Objective of study**

1. Create machine learning models that can accurately predict crop yields for Irish Potato and Maize.
2. Analyze the data to identify key factors influencing crop yields.
3. Identify which factors have the most significant impact.

**1.3 Introduction**

Agriculture, since its invention and inception, be the prime and pre-eminent activity of every culture and civilization throughout the history of mankind. It is not only an enormous aspect of the growing economy, but it’s essential for us to survive. It’s also a crucial sector for Indian economy and also human future.

It also contributes an outsized portion of employment. Because the time passes the requirement for production has been increased exponentially. So as to produce in mass quantity people are using technology in an exceedingly wrong way. New sorts of hybrid varieties are produced day by day.

However, these varieties don’t provide the essential contents as naturally produced crop. These unnatural techniques spoil the soil. It all ends up in further environmental harm. Most of these unnatural techniques are wont to avoid losses. But when the producers of the crops know the accurate information on the crop yield it minimizes the loss. Machine learning, a fast-growing approach that’s spreading out and helping every sector in making viable decisions to create the foremost of its applications.

Most devices nowadays are facilitated by models being analysed before deployment. The main concept is to increase the throughput of the agriculture sector with the Machine Learning models. Another factor that also affects the prediction is the amount of knowledge that’s being given within the training period, as the number of parameters was higher comparatively. The core emphasis would be on precision agriculture, where quality is ensured over undesirable environmental factors.

So as to perform accurate prediction and stand on the inconsistent trends in temperature and rainfall various machine learning classifiers like Logistic Regression, Naïve Bayes, Random Forest etc. are applied to urge a pattern. By applying the above machine learning classifiers, we came into a conclusion that Random Forest algorithm provides the foremost accurate value.

System predicts crop prediction from the gathering of past data. Using past information on weather, temperature and a number of other factors the information is given. The Application which we developed, runs the algorithm and shows the list of crops suitable for entered data with predicted yield value.

Aruvansh Nigam, Saksham Garg, Archit Agrawal [1] conducted experiments on Indian government dataset and it’s been established that Random Forest machine learning algorithm gives the best yield prediction accuracy. Sequential model that’s Simple Recurrent Neural Network performs better on rainfall prediction while LSTM is good for temperature prediction.

The paper puts factors like rainfall, temperature, season, area etc. together for yield prediction. Results reveals that Random Forest is the best classifier when all parameters are combined. Leo Brieman [2], is specializing in the accuracy and strength & correlation of random forest algorithm.

Random forest algorithm creates decision trees on different data samples and then predict the data from each subset and then by voting gives better the answer for the system. Random Forest used the bagging method to train the data. To boost the accuracy, the randomness injected has to minimize the correlation ρ while maintaining strength.

Balamurugan [3], have implemented crop yield prediction by using only the random forest classifier. Various features like rainfall, temperature and season were taken into account to predict the crop yield.

Other machine learning algorithms were not applied to the datasets. With the absence of other algorithms, comparison and quantification were missing thus unable to provide the apt algorithm. Mishra [4], has theoretically described various machine learning techniques that can be applied in various forecasting areas.

However, their work fails to implement any algorithms and thus cannot provide a clear insight into the practicality of the proposed work. Dr. Y. Jeevan Nagendra Kumar [5], have concluded Machine Learning algorithms can predict a target/outcome by using Supervised Learning. This paper focuses on supervised learning techniques for crop yield prediction.

To get the specified outputs it needs to generate an appropriate function by set of some variables which can map the input variable to the aim output. The paper conveys that the predictions can be done by Random Forest ML algorithm which attain the crop prediction with best accurate value by considering least number of models.

Agriculture is the fundamental source of food industry. It is one of the most oldest and important economic activities which is being practiced in the world wide since thousands of years. Its development has taken over the period of many years with the emergence of new technology, equipment, techniques of farming and domestication.

Huge advancement and growth can be seen in this sector with the time period. This sector, not only witnessed the enormous growth but also gave rise to many other sectors with significant progress.

Majority of the people are being involved in this occupation as it is the basic need of human beings survival. More than 50% of the land in the world has been devoted to agriculture. Agriculture sector accounts for 14% of Gross Domestic Product (GDP) of the Indian economy. About 70% of the population of India lives in rural areas and majority of them depend upon agriculture as their primary source of income.

Agriculture not only helps people to survive but keeps economy on-going. It plays vital role in the economic development of India. Government of India has shown concern about the improvement of cultivator’s knowledge of the soil, improvement of the fertility of the soil, irrigation facilities, fertilizer utilization, cattle-manure utilization, precise pesticides usage and grazing in forest area. Thus productivity has to be increased with the increase in population.

In agriculture planning to obtain maximum crop yield with restricted area of land is the largest task in an agro-based country like India. Yield rate of the crop can be increased with the help of indicators by investigating crop related problems. Crop selection will be more accurate and beneficial with minimum loss, whether unfavourable condition occurs [10]. Maximum crop yield can be obtained in favourable growing condition.

Improving production rate of crop can be an important topic for research for the agro-meteorologists, for the development of economic growth of the country. The two main factors responsible for the yield rate of the crop is, first one is quality of seeds which can be improved by genetic development using hybridization technology and second one is the selection of crop based on the favourable and unfavourable conditions. The two techniques: statistical and machine learning both these techniques modelled.

Many researchers had been tried to get an efficient and accurate model for crop yielding prediction, soil classification, crop classification, weather predictions [13], crop disease prediction classification of crops [5, 6]. Thus this new method called cropRS for Crop Disease Detection and Yield Prediction 143 selection method (CSM) developed to increase in net yield rate of crops over seasons.

Crop production rate depends on the topography and geographic condition of the region (e.g. mountainous region, hilly area, river ground, depth regions), weather condition (e.g. humidity, temperature, rainfall, cloud) [14, 15], soil type (e.g. sandy, clay, peaty, saline, silty, loam soil), soil composition (e.g. PH value, nitrogen, phosphorous, magnesium, calcium, sulphur, potassium, organic carbon, copper, iron) and harvesting methods.

Different prediction models are used for different parameters of different crops. Some of these prediction models are studied thoroughly through researches for the crop production. The prediction models are of two types: statistical model and machine learning. This chapter describes overall study on Machine learning concepts, algorithms, and methods.

This also gives an overview of recommender system and how recommender system is used in agriculture for disease detection and prediction. The chapter also describes crop management activities such as crop yield prediction, disease detection, weed detection and crop quality. Lastly application of agriculture and recommender system is discussed.

**1.4 Machine Learning**

**1.4.1 Overview**

Machine learning is emerging technology day by day in different fields. But now-a-days agriculture is the sector where machine learning applications are in greater demand.

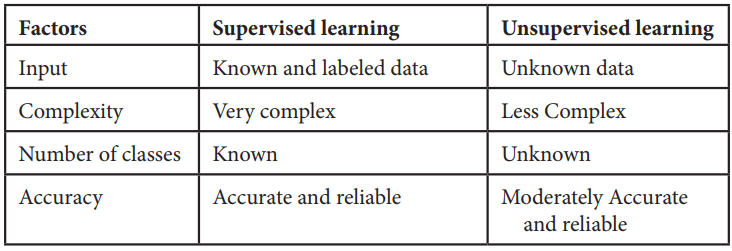
**Now the question arises what is machine learning?**

Actually it is nothing but machine that learns from experiences in order to perform specific task. It provides ability to learn. Huge real time data set is provided to the system.

This data set contains set of attributes called as features. Learning uses these features for further analysis. The performance is measured with performance metric. As more data is added over a time, this performance metric is improved with experience. Various mathematical models are used to calculate performance of machine learning algorithm.

Machine learning tasks are broadly classified into supervised learning, unsupervised learning, reinforcement learning, Semi-supervised learning and learning models (classification, regression, clustering, and dimensionality reduction). Table 1.1 describes difference between supervised learning and unsupervised learning.

**Table 1.1 Difference between supervised learning and unsupervised learning.**



**1. Supervised learning:**

In supervised learning labelled dataset is used. Here the model is trained on labelled dataset. This dataset contains both input and output parameters. Supervised learning is learning where there is input data, output data and algorithm that maps to input and output.

Learning means input and output is provided to machine and machine will develop its own logic for the given task. Some supervised learning algorithms are: Linear Regression, Nearest Neighbour, Gaussian Naive Bayes, Decision Trees, Support Vector Machine (SVM), Random Forest, etc.

The learning is called supervised learning because it is similar to a teacher who is acting as supervisor on entire learning process. The predictions on trained data are generated by learning algorithms. These predictions are corrected by the teacher and learning process is stopped when correct output is achieved [5, 6].

**2. Unsupervised learning:**

Unsupervised learning is learning that contains only input data and no output data is present. This is applied where there is need to model data distribution in order to get more and more data and there is no any supervisor (like teacher) to supervised the things. Algorithms themselves learn, discover and present structure in data. Here algorithm itself create data pattern. Some recommendation systems for marketing automation use this type of learning.

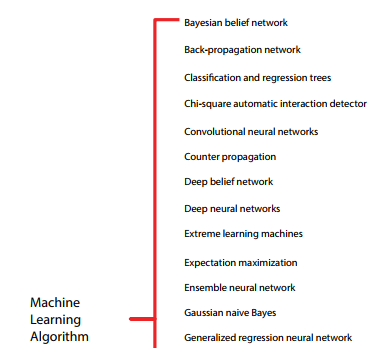
**3. Semi-supervised learning:**

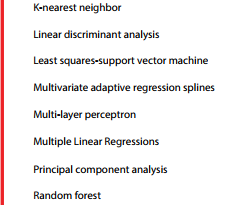
The supervised learning has disadvantage that it required labelled dataset. This process is very costly while dealing with large volume of data. Unsupervised learning also has disadvantage that the range of its application is limited.

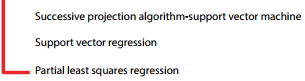
To solve these problems, semi supervised learning algorithms concept was developed. This concept used both labelled data and unlabelled data so that it can work on any type of data. Mostly it contains small amount of labelled data and huge amount of unlabelled data.

**1.4.2 Machine Learning Algorithms**

There are various machine learning algorithms as described in Figure 1.1 below. These algorithms can be applied in any area to solve various problems.







**Figure1.1 Machine learning algorithms**

**1.4.3 Machine Learning Methods**

There are various Machine learning methods such as boosting techniques (RGF, GBDT, and Ada boost), Regression Tree (ID3, C4.5), Random Forest, SVM, K Nearest Neighbor, and ANN, etc.

**1.4.3.2 Artificial Neural Network**

Artificial Neural Network is a simple mathematical model of the brain. This is used to process nonlinear relationships between inputs and outputs in parallel form, for example human brain. So Artificial Neural Networks can be used in different variety of tasks one of the best use is classification. We can learn Artificial Neural Network speedily. The Information flows through a neural network in two different ways.

Firstly when the model is learning or operating normally, the information from the dataset is given to the network through the input neurons, which then trigger the layers of hidden neurons, and then it is converted to the output neurons. So this is called as feed forward network. Each neuron receives inputs to its left, and then they are multiplied by the weights. So every neuron adds up all the inputs. If the sum is more than a certain threshold value then the neuron “fires”.

Whenever we use large datasets, the neural networks are more powerful at that time. So in the basic structure of an Artificial Neural Network we can create 3 layers of “neurons”- The input layer, the hidden layer and the output layer as shown in Figure 8.2. The information flows from the input layer, from the hidden layer to the output layer.

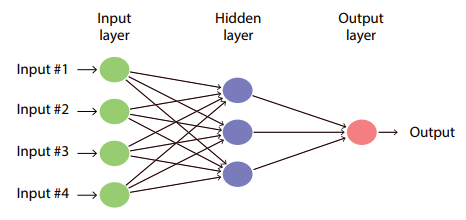
As each of the connections has a number associated with it, and it is called the connection weight. Also, each of the neurons has a number and a special formula associated with it called as threshold value.

The neural network can be trained and then it can be provided with a set of inputs and outputs. Each neuron transforms the input and forwards it to the next layer and so on. The result is received on the output layer. Then the layer is compared to the outputs through special algorithms is used to produced outputs as close to each other as possible.

As this process is repeated many times and here completes the training part. Whenever new inputs are provided to the network, we can get the actual outputs. So, the Artificial Neural Network can be used in predicting house prices and classifying objects and images.

**1.4.3.2 Support Vector Machines**

SVM stands for Support Vector Machine. It is a simple algorithm in machine learning. It belongs to the supervised learning category in



**Figure 1.2 Simple Artificial Neural Network**

machine learning which is used for both regression and classification analysis but it is a discriminant classifier which is widely used as classification algorithm. It produces significant accuracy with less computation power.

This algorithm is used to create a line or a hyper plane in order to separate the data into various classes. It takes the data as input and outputs a line that separates two classes. It is used in variety of applications such as face detection, classification of emails, news articles, web pages, handwriting recognition.

**Crop identification:**

With the help of advanced technology, lot of improvement has been done in agriculture sector. When a farmer purchases a land for farming the most important question arises is which crop to be grown to have better production? If the selected crop is unsuitable for that land, then there will be less productivity.

Generally, farmers are unaware about agriculture land requirements that is minerals needed, soil moisture and other soil requirements. The concept of precision agriculture solves this problem of farmer. In this characteristic of soil are identified and used for identifying the suitable crop for cultivation.

Soil characteristics are soil type, texture, pH value, moisture, temperature and other environment conditions. Due to this risk of cultivating improper crop is avoided. Always farmer will cultivate proper crop that results in good crop yield and get enormous profit for particular land area.

Here recommendation system helps in recommending the most suitable crop that yields in better productivity. Also, it prevents from crop loss making farmer to be financial stable.

**1.5 Crop Management**

**1.5.1 Yield Prediction**

The most important topic in agriculture is yield prediction because yield prediction contributes to profit [5]. It includes activities such as yield mapping and estimation, pairing crop supply with demand and crop management.

All these activities lead to increase production. One of the low costs and efficient machine learning application was developed that contributes in automatically counting coffee fruits on branch of a tree.

The calculation of coffee fruits is classified into three types: fruits that are harvestable, fruits that are non-harvestable and fruits that are ignored from stage of maturation. This system also calculated the weight of coffee fruit and maturation percentage.

This system aims to help the factory owner who wants to grow coffee by providing the best economic benefits and also plan. The machine learning application named a machine vision system was developed that automatically shake and catch cherries while harvesting process.

This system does segmentation and detects the obstructed branches even though these branches are full of leaves and not visible clearly. This system aims to reduce labor work and manual handling of operation during harvesting. One of the systems introduced was yield mapping system that identifies immature green citrus.

This system aims to provide information of good citrus in citrus groove that will result in better profit and yield. This is very useful system for citrus growers. Based upon the Artificial Neural Network, a new model for calculating grassland biomass was developed. The study was also done on yield prediction.

For study, wheat crop was considered and images were taken with the help of satellite. Using image processing crop growth characteristics were identified and data were fused to predict the result. Another application was developed to detect the tomatoes using images that were sensed remotely.

These images were captured using Unmanned Aerial vehicle. Based on SVM, new rice development stage prediction model was developed. All the basic information that is required for processing was obtained from weather forecasting station in China.

The study was done that focuses on helping farmers to overcome imbalances that happened in the market supply and demand. This may be due to quality of crop. A generalized method based on ENN application was developed for good agricultural yield prediction.

The following table describes study about various crops along with their functionality, algorithms and results. Table 8.2 describes study about various crops along with their functionality, algorithms and results.

**CHAPTER 2**

**LITERATURE REVIEW**

**[1] S. S. Sannakki and V. S. Rajpurohit, proposed**

A “Classification of Pomegranate Diseases Based on Back Propagation Neural Network” which mainly works on the method of Segment the defected area and Color and texture are used as the features. Here they used neural network classifier for the classification. The main advantage is it Converts to L\*a\*b to extract chromaticity layers of the image and Categorisation is found to be 97.30% accurate. The main disadvantage is that it is used only for the limited crops.

**[2] P. R. Rothe and R. V. Kshirsagar introduced**

A” Cotton Leaf Disease Identification was using Pattern Recognition Techniques” which Uses snake segmentation; here Hu’s moments are used as distinctive attribute. Active contour model used to limit the vitality inside the infection spot, BPNN classifier tackles the numerous class problems. The average classification is found to be 85.52%.

**[3] Aakanksha Rastogi, Ritika Arora and Shanu Sharma,”**

Leaf Disease Detection and Grading using Computer Vision Technology &Fuzzy Logic”. K-means clustering used to segment the defected area; GLCM is used for the extraction of texture features, Fuzzy logic is used for disease grading. They used artificial neural network (ANN) as a classifier which mainly helps to check the severity of the diseased leaf.

**[4] Godliver Owomugisha, John A.**

Quinn, Ernest Mwebaze and James Lwasa, proposed” Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease “Color histograms are extracted and transformed from RGB to HSV, RGB to L\*a\*b. Peak components are used to create max tree, five shape attributes are used and area under the curve analysis is used for classification.

They used nearest neighbours, Decision tree, random forest, extremely randomized tree, Naïve Bayes and SV classifier. In seven classifiers extremely, randomized trees yield a very high score, provide real time information provide flexibility to the application.

**[5] Uan Tian, Chunjiang Zhao, Shenglian Lu and Xinyu Guo,**”

SVM-based Multiple Classifier System for Recognition of Wheat Leaf Diseases,” Color features are represented in RGB to HIS, by using GLCM, seven invariant moment are taken as shape parameter. They used SVM classifier which has MCS, used for detecting disease in wheat plant offline.

**[1] Aruvansh Nigam, Saksham Garg, Archit Agrawal “Crop Yield Prediction using ML Algorithms “, 2019**

Agriculture is the pillar of the Indian economy and more than 50% of India's population are dependent on agriculture for their survival. Variations in weather, climate, and other such environmental conditions have become a major risk for the healthy existence of agriculture. Machine learning (ML) plays a significant role as it has decision support tool for Crop Yield Prediction (CYP) including supporting decisions on what crops to grow and what to do during the growing season of the crops.

The present research deals with a systematic review that extracts and synthesizes the features used for CYP and furthermore, there are a variety of methods that were developed to analyse crop yield prediction using artificial intelligence techniques.

The major limitations of the Neural Network are reduction in the relative error and decreased prediction efficiency of Crop Yield. Similarly, supervised learning techniques were incapable to capture the nonlinear bond between input and output variables faced a problem during the selection of fruits grading or sorting.

Many studies were recommended for agriculture development and the goal was to create an accurate and efficient model for crop classification such as crop yield estimation based on the weather, crop disease, classification of crops based on the growing phase etc., This paper explores various ML techniques utilized in the field of crop yield estimation and provided a detailed analysis in terms of accuracy using the techniques.

The present research work discussed about the variety of features that are mainly dependent on the data availability and each of the research will investigated CYP using ML algorithms that differed from the features.

The features were chosen based upon the geological position, scale, and crop features and these choices were mainly dependent upon the data-set availability, but the more features usage was not always giving better results.

Therefore, finding the fewer best performing features were tested that also have been utilized for the studies. Most of the existing models utilized neural networks, random forests, KNN regression techniques for CYP and a variety of ML techniques were also used for best prediction. From the studies most of the common algorithms used were CNN, LSTM, DNN algorithms but still improvement was still required further in CYP.

The present research shows several existing models that consider elements such as temperature, weather condition, performing models for the effective crop yield prediction. Ultimately, the experimental study showed the combination of ML with the agricultural domain field for improving the advancement in crop prediction.

**[2] Leo Brieman, “Random Forests”, 2001**

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of over fitting to their training set. 587–588 Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

The first algorithm for random decision forests was created in 1995 by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg.

An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).

The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance. Random forests are frequently used as "black box" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

The general method of random decision forests was first proposed by Ho in 1995. Ho established that forests of trees splitting with oblique hyper planes can gain accuracy as they grow without suffering from overtraining, as long as the forests are randomly restricted to be sensitive to only selected feature dimensions.

A subsequent work along the same lines concluded that other splitting methods behave similarly, as long as they are randomly forced to be insensitive to some feature dimensions. Note that this observation of a more complex classifier (a larger forest) getting more accurate nearly monotonically is in sharp contrast to the common belief that the complexity of a classifier can only grow to a certain level of accuracy before being hurt by over fitting. The explanation of the forest method's resistance to overtraining can be found in Kleinberg's theory of stochastic discrimination.

The early development of Breiman's notion of random forests was influenced by the work of Amit and Geman who introduced the idea of searching over a random subset of the available decisions when splitting a node, in the context of growing a single tree.

The idea of random subspace selection from Ho was also influential in the design of random forests. In this method a forest of trees is grown, and variation among the trees is introduced by projecting the training data into a randomly chosen subspace before fitting each tree or each node.

Finally, the idea of randomized node optimization, where the decision at each node is selected by a randomized procedure, rather than a deterministic optimization was first introduced by Thomas G. Dietterich. The proper introduction of random forests was made in a paper by Leo Breiman.

This paper describes a method of building a forest of uncorrelated trees using a CART like procedure, combined with randomized node optimization and bagging. In addition, this paper combines several ingredients, some previously known and some novel, which form the basis of the modern practice of random forests, in particular: [3] Priya, P., Muthaiah, U., Balamurugan, and M.” Predicting Yield of the Crop Using Machine Learning Algorithm”, 2015

The agriculture plays a dominant role in the growth of the country’s economy. Climate and other environmental changes has become a major threat in the agriculture field. Machine learning (ML) is an essential approach for achieving practical and effective solutions for this problem.

Crop Yield Prediction involves predicting yield of the crop from available historical available data like weather parameter, soil parameter and historic crop yield. This paper focus on predicting the yield of the crop based on the existing data by using Random Forest algorithm. Real data of Tami lnadu were used for building the models and the models were tested with samples.

The prediction will helps to the farmer to predict the yield of the crop before cultivating onto the agriculture field. To predict the crop yield in future accurately Random Forest, a most powerful and popular supervised machine learning algorithm is used.

Agriculture is the backbone of every economy. In a country like India, which has ever increasing demand of food due to rising population, advances in agriculture sector are required to meet the needs. From ancient period, agriculture is considered as the main and the foremost culture practiced in India.

Ancient people cultivate the crops in their own land and so they have been accommodated to their needs. Therefore, the natural crops are cultivated and have been used by many creatures such as human beings, animals and birds.

The greenish goods produced in the land which have been taken by the creature leads to a healthy and welfare life. Since the invention of new innovative technologies and techniques the agriculture field is slowly degrading. Due to these, abundant invention people are been concentrated on cultivating artificial products that is hybrid products where there leads to an unhealthy life.

Nowadays, modern people don’t have awareness about the cultivation of the crops in a right time and at a right place. Because of these cultivating techniques the seasonal climatic conditions are also being changed against the fundamental assets like soil, water and air which lead to insecurity of food.

By analysing all these issues and problems like weather, temperature and several factors, there is no proper solution and technologies to overcome the situation faced by us. In India there are several ways to increase the economic growth in the field of agriculture. There are multiple ways to increase and improve the crop yield and the quality of the crops.

Data mining also useful for predicting the crop yield production. Generally, data mining is the process of analysing data from different perspectives and summarizing it into useful information.

The Results shows that we can attain an accurate crop yield prediction using the Random Forest algorithm. Random Forest algorithm achieves a largest number of crop yield models with a lowest models. It is suitable for massive crop yield prediction in agricultural planning. This makes the farmers to take the right decision for right crop such that the agricultural sector will be developed by innovative ideas.

**[5] Dr. Jeevan Kumar, “Supervised Learning Approach for Crop Production”, 2020**

Machine learning (ML) is a crucial perspective for acquiring real-world and operative solution for crop yield issue. From a given set of predictors, ML can predict a target/outcome by using Supervised Learning. To get the desired outputs need to generate a suitable function by set of some variables which will map the input variable to the aim output.

Crop yield prediction incorporates forecasting the yield of the crop from past historical data which includes factors such as temperature, humidity, ph, rainfall, crop name. It gives us an idea for the finest predicted crop which will be cultivate in the field weather conditions. These predictions can be done by a machine learning algorithm called Random Forest.

It will attain the crop prediction with best accurate value. The algorithm random forest is used to give the best crop yield model by considering least number of models. It is very useful to predict the yield of the crop in agriculture sector.

Implement a system to predict crop production from the collection of past data. Using data mining techniques crop yield is predicted. Here, using Random Forest algorithm for predicting the best crop yield as output. In agriculture field, the crop yield prediction is mostly appropriate.

The more increase in accuracy results in more profit to the crop yield. The proposed technique helps farmers to acquire apprehension in the requirement and price of different crops. It helps farmers in decision making of which crop to cultivate in the field. The more increase in accuracy results in more profit to the crop yield.

This work is employed to search out the gain knowledge about the crop that can be deployed to make an efficient and useful harvesting. Under this system, maximum types of crops will be covered. The accurate prediction of different specified crops across different districts will help farmers of India.

**[6] Ramesh Medar,Vijay S, Shweta, “Crop Yield Prediction using Machine Learning Techniques”, 2019**

Agriculture is the field which plays an important role in improving our countries economy. Agriculture is the one which gave birth to civilization. India is an agrarian country and its economy largely based upon crop productivity. Hence we can say that agriculture can be backbone of all business in our country.

Selecting of every crop is very important in the agriculture planning. The selection of crops will depend upon the different parameters such as market price, production rate and the different government policies. Many changes are required in the agriculture field to improve changes in our Indian economy.

We can improve agriculture by using machine learning techniques which are applied easily on farming sector. Along with all advances in the machines and technologies used in farming, useful and accurate information about different matters also plays a significant role in it.

The concept of this paper is to implement the crop selection method so that this method helps in solving many agriculture and farmers problems. This improves our Indian economy by maximizing the yield rate of crop production.

**[7] Ranjini B Guruprasad, Kumar Saurav, Sukanya Randhawa,”Machine Learning Methodologies for Paddy Yield Estimation in India: A CASE STUDY”, 2019**

In-season crop yield estimation has various applications such as the farmer taking corrective measures to increase the yield, optimizing the supply-demand chain of fertilizers, pesticides and agricultural commodities, price prediction and determining the risk levels in agriculture insurance.

Crop yield estimation models based on remote sensed satellite data and weather data have been widely adopted as do not require specific information about the crop and farming practices and hence are scalable across crops, growing locations and conditions.

In this paper, we present a case study of weather and soil data based yield estimation modelling for paddy crop at different spatial resolution (SR) levels, namely, at the district (coarser SR) and taluk (finer SR) levels in India. We provide a detailed analysis of accuracy of the yield estimation models across varied sets of features and different machine learning (ML) techniques.

Further, we perform dis-aggregation of district yield data by applying the machine learning models trained using district level data to predict yields at taluk level. Taluk level yield prediction by dis-aggregation of district level data has average error of 6% and maximum error of 25%.

In this paper, we have explored several ML methodologies to estimate the yield of paddy crop using weather and soil data as input features. Using the taluk data with weekly averaging, we achieved an average error of 3.14% and maximum error of 9.66%.

Further, we presented dis-aggregation of district yield data using district/taluk data to predict taluk yield. The taluk yield predicted by dis-aggregation of district yield data has an average error of 6% with maximum error of 25%.

**[8] Sangeeta, Shruthi G, “Design and Implementation of Crop Yield Prediction Model In Agriculture”, 2020**

Agriculture is something that individuals have started to finish up moderate on, disregarding that it's miles what is holding us alive. However, there is regardless some driving forward, enthusiastic ranchers whose life continues running on essentially developing. Regardless, there's in addition the pollution that is extending packages these days.

The Main intention of the Department of Agricultural Marketing and Agricultural Business is to have a reasonable cost to the cultivating network who are pushed behind the current focused showcasing situation and the mission of accomplishing the reasonable cost is by making the current demonstration and principles solid and progressively compelling by executing new innovations and systems went for lessening pre and post gather misfortunes through legitimate and sorted out techniques and urge enhancing the market.

The vital motivation behind making a managed market is to put off the undesirable exchange work out, to diminish the charges inside the commercial centre and to offer reasonable expenses to the Farmers.

A few activities have been taken to advance rural showcasing a decent method to cultivate and keep up the place of country monetary improvement. To advantage the cultivating from the new worldwide market get admission to potential outcomes, the inward rural promoting device inside the United States of America moreover wishes to be joined and strengthened.

This project is undertaken using machine learning and evaluates the performance by using Random forest, Polynomial Regression and Decision Tree algorithms. In our proposed model among all the three algorithm Random forest gives the better yield prediction as compared to other algorithms. Along with random forest, Polynomial Regression, Decision Tree model classify the output that shows improvements in dataset.

So we analysed that proposed model has got more efficiency than the existing model for finding crop yield. The implementation of above system would help in better cultivation of the agricultural practices of our country. Further it can be used to reduce the loss faced by the farmers and improve the crop yield to get better capital in agriculture.

The model can be improved by integrating this with other departments like horticulture, sericulture, and others towards the agricultural development of our country.

**[17] Ebrahimi, M.A., Khoshtaghaza, M.H., Minaei, S., Jamshidi, B., Vision-based pest detection based on SVM classification method. Comput. Electron. Agric., 137, 52–58, 2017.**

Automatic pest detection is a useful method for greenhouse monitoring against pest attacks. One of the more harmful pests that threaten strawberry greenhouses is thrips (Thysanoptera).

Therefore, the main objective of this study is to detect of thrips on the crop canopy images using SVM classification method. A new image processing technique was utilized to detect parasites that may be found on strawberry plants.

SVM method with difference kernel function was used for classification of parasites and detection of thrips. The ratio of major diameter to minor diameter as region index as well as Hue, Saturation and Intensify as color indexes was utilized to design the SVM structure.

Also, mean square error (MSE), root of mean square error (RMSE), mean absolute error (MAE) and mean percent error (MPE) were used for evaluation of the classification. Results show that using SVM method with region index and intensify as color index make the best classification with mean percent error of less than 2.25%.

It was shown that the developed image processing procedure is capable of identifying parasites in the greenhouse environment. Also, the parasites can be classified using the SVM classification method and the target parasite can be detected. In this study, incorporation of the image processing technique with SVM method and choice of suitable region and color index was successful in detecting the target (thrips) with an error less than 2.5%.

**[18]. Chung, C.L., Huang, K.J., Chen, S.Y., Lai, M.H., Chen, Y.C., Kuo, Y.F., Detecting Bakanae disease in rice seedlings by machine vision. Comput. Electron. Agric., 121, 404–411, 2016.**

Bakanae disease, or ‘‘foolish seedling”, is a seed-borne disease of rice (Oryza sativa L.). Infected plants can yield empty panicles or perish, resulting in a loss of grain yield. The disease occurs most frequently when contaminated seeds are used. Once the seeds are contaminated, the pathogen Fusarium fujikuroi spreads in the field.

Therefore, infected plants must be screened at early developmental stages. This work proposes an approach to non-destructively distinguish infected and healthy seedlings at the age of 3 weeks using machine vision.

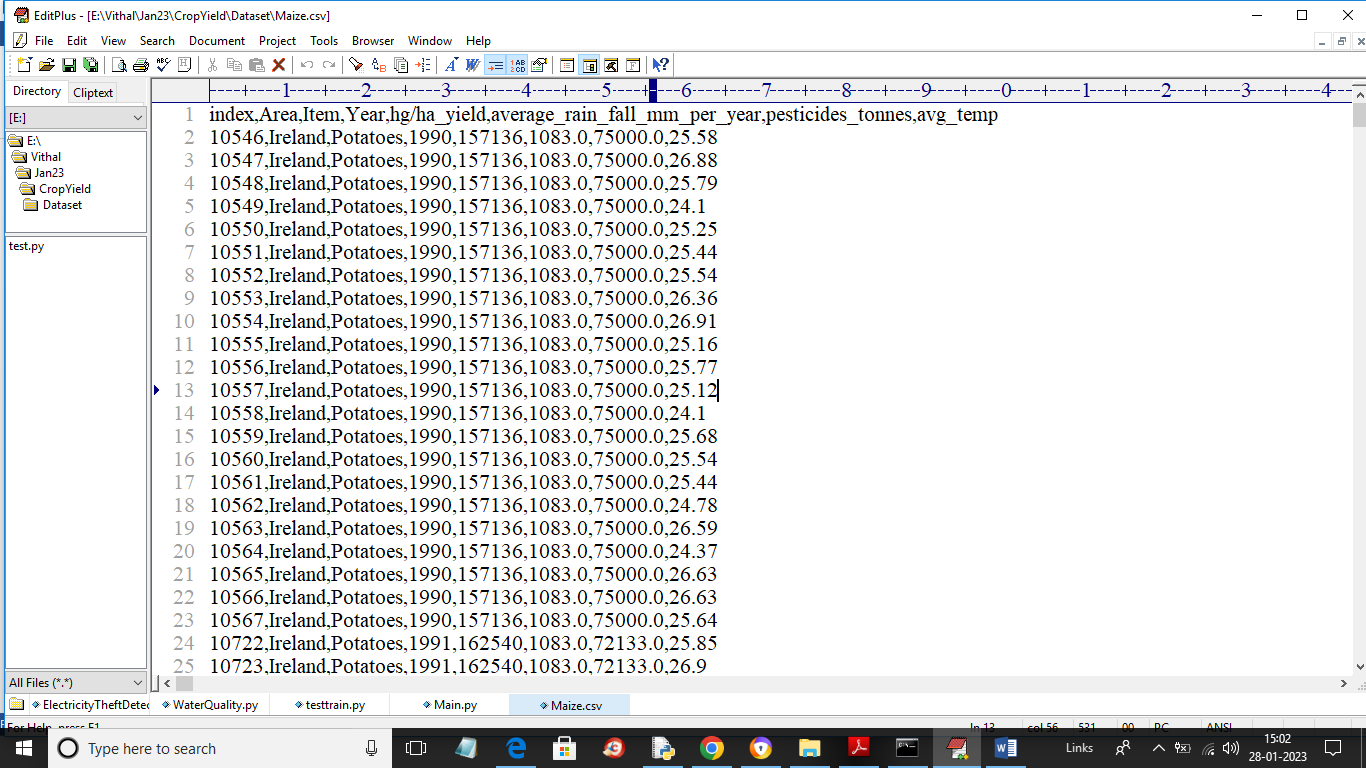
Seeds of the rice cultivars Tainan 11 and Toyonishiki were inoculated with a conidial suspension of F. fujikuroi. The seedling were cultivated in an incubator for 3 weeks. The images of infected and control seedlings were acquired using flatbed scanners to quantify their morphological and color traits.

Support vector machine (SVM) classifiers were developed for distinguishing the infected and healthy seedlings. A genetic algorithm was used for selecting essential traits and optimal model parameters for the SVM classifiers. The proposed approach distinguished infected and healthy seedlings with an accuracy of 87.9% and a positive predictive value of 91.8%.

**CHAPTER 3**

**PROPOSED METHOD**

In propose paper author using Random Forest and SVR algorithm but you asked us to implement RF, SVR, DNN, CNN, ANN and LSTM so we have implemented all 5 algorithms on both datasets. To evaluate performance of each algorithm we are calculating MSE and R2 Score where MSE refers to mean square error (difference between TEST crop yield and predicted yield). R2 refers to correct prediction rate. So for any algorithm MSE must be lower and R2 must be higher,

****

**3.1 Dataset**

In above dataset screen first row contains dataset column names and remaining rows contains dataset values. We have coded this project using JUPYTER notebook and below are the output screens with code and below colour comments

Upload Crop Dataset

(Maize, Potato)

Cluster Based Weighting

Run Algorithm (RF, SVR, **DNN, CNN, ANN and LSTM**)

MSE Comparison Graph

**Fig.3.2 Flowchart**

**CHAPTER 4**

**SOFTWARE AND HARDWARE REUIREMENT**

**Hardware Requirement:**

• Processor Type: Pentium -IV

• ROM: 512 MB

• RAM: 4 GB

• Hard disk: 20 GB

**Software Requirement:**

• Operating System: Windows 2007/8/10/11

• Script: **python Jupyter notebook**

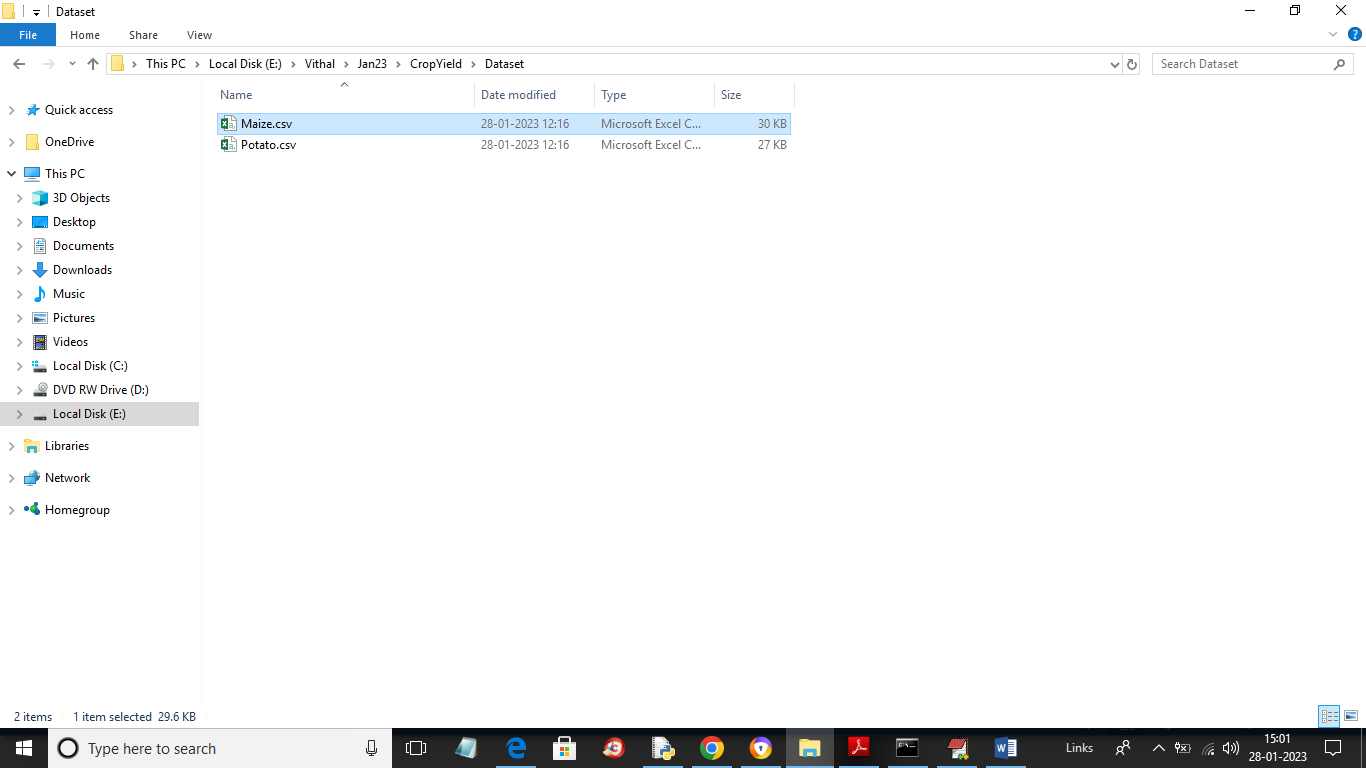
**CHAPTER 5**

**RESULT**

In this paper author employing Random Forest and SVR algorithm to predict future crop yield based on weather data such as temperature and rainfall. If farmers know the crop yield before sowing based on historical weather data then he may take better decision. So by employing machine learning algorithms we can inform farmers about future crop yield. In propose paper author using Irish Maize and Potato yield dataset to train all machine learning models and then this models can be used to predict future crop yield.

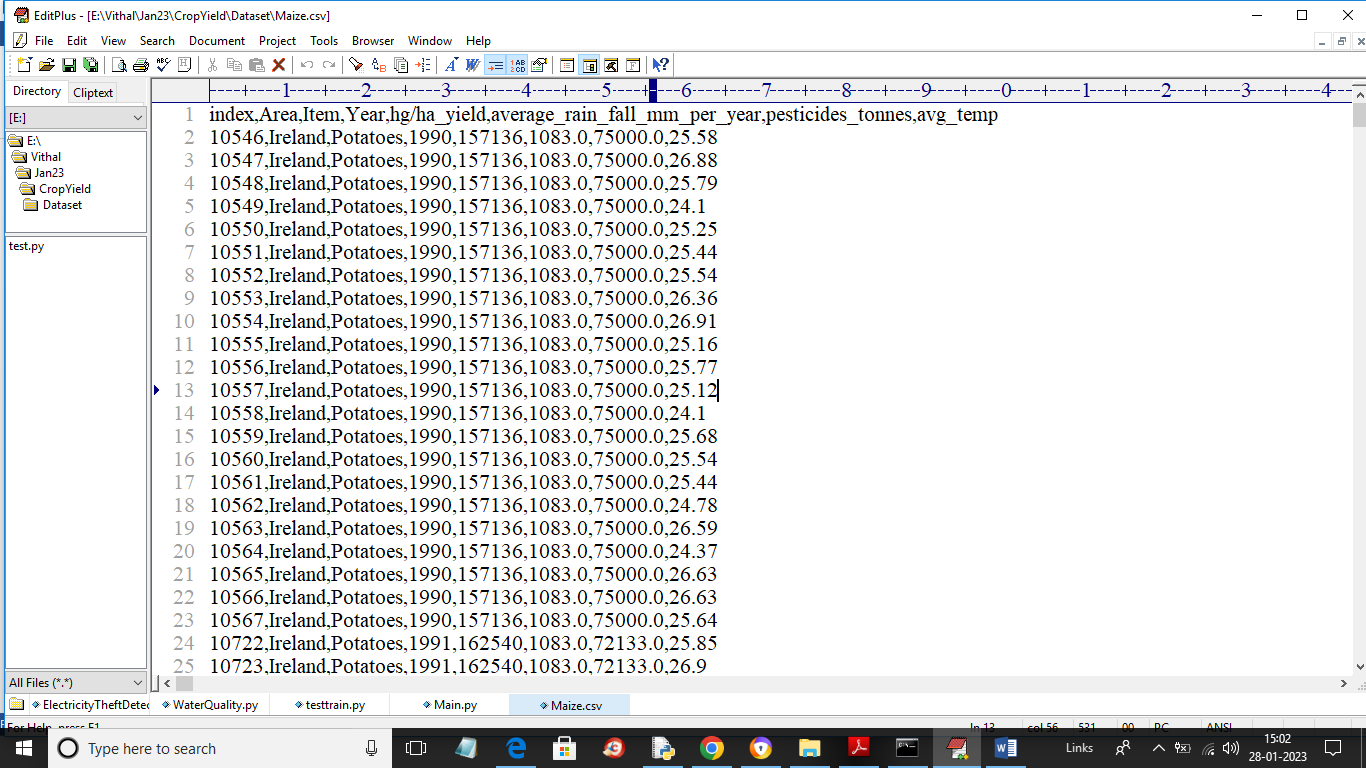
In propose paper author using Random Forest and SVR algorithm but you asked us to implement SVR, DNN, CNN, ANN and LSTM so we have implemented all 5 algorithms on both datasets. To evaluate performance of each algorithm we are calculating MSE and R2 Score where MSE refers to mean square error (difference between TEST crop yield and predicted yield). R2 refers to correct prediction rate. So for any algorithm MSE must be lower and R2 must be higher.

In below screen we are showing dataset details used in this project

****

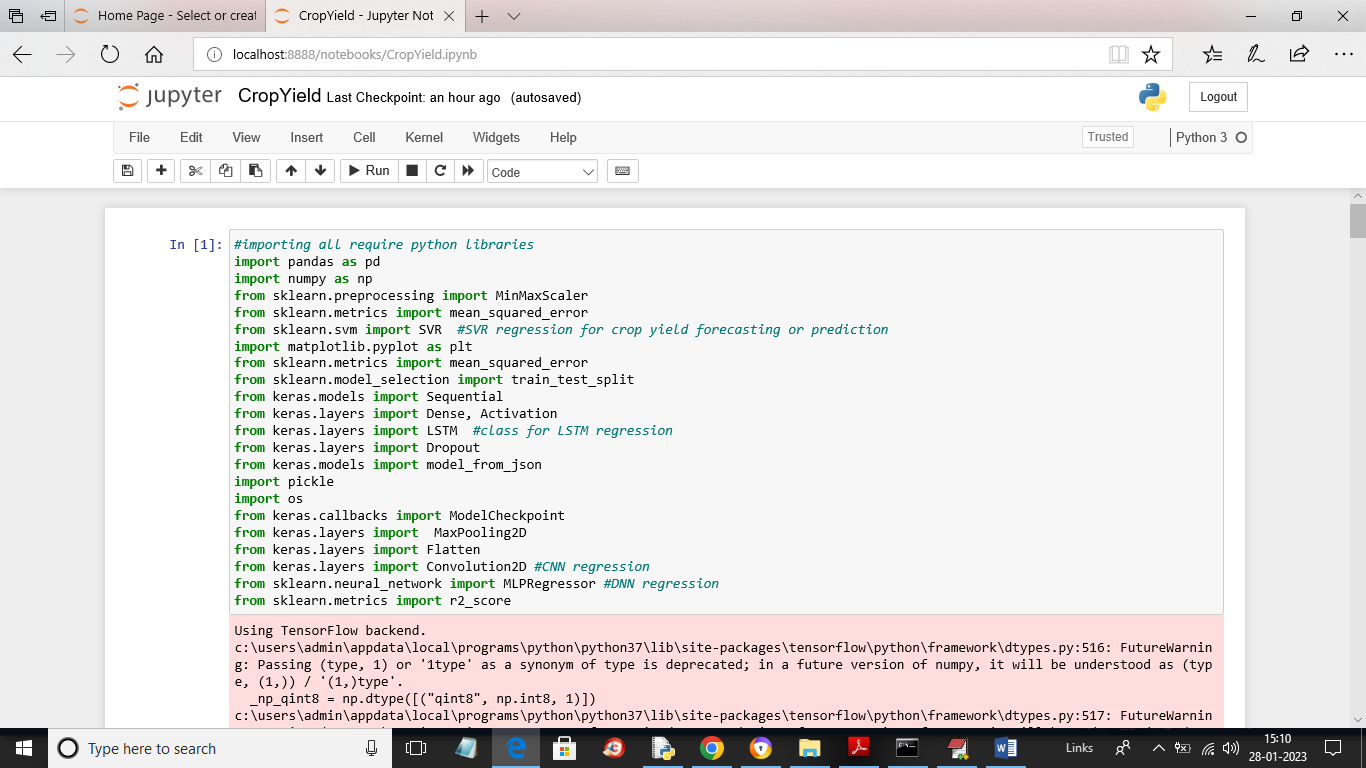
**Fig.5.1 Dataset folder**

In above dataset folder we have 2 folders called as Maize and Potato and just open those files to view below data

****

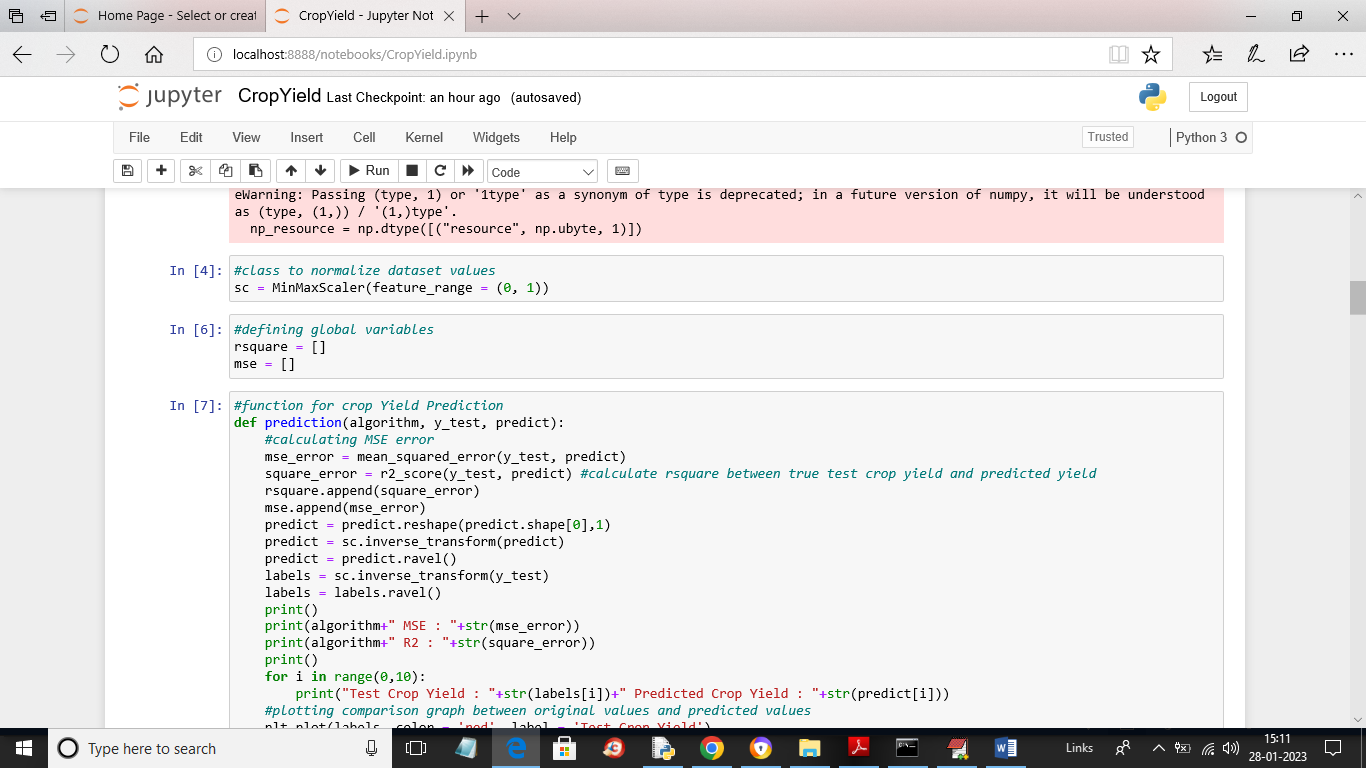
**Fig.5.2 dataset**

In above dataset screen first row contains dataset column names and remaining rows contains dataset values. We have coded this project using JUPYTER notebook and below are the output screens with code and below colour comments

****

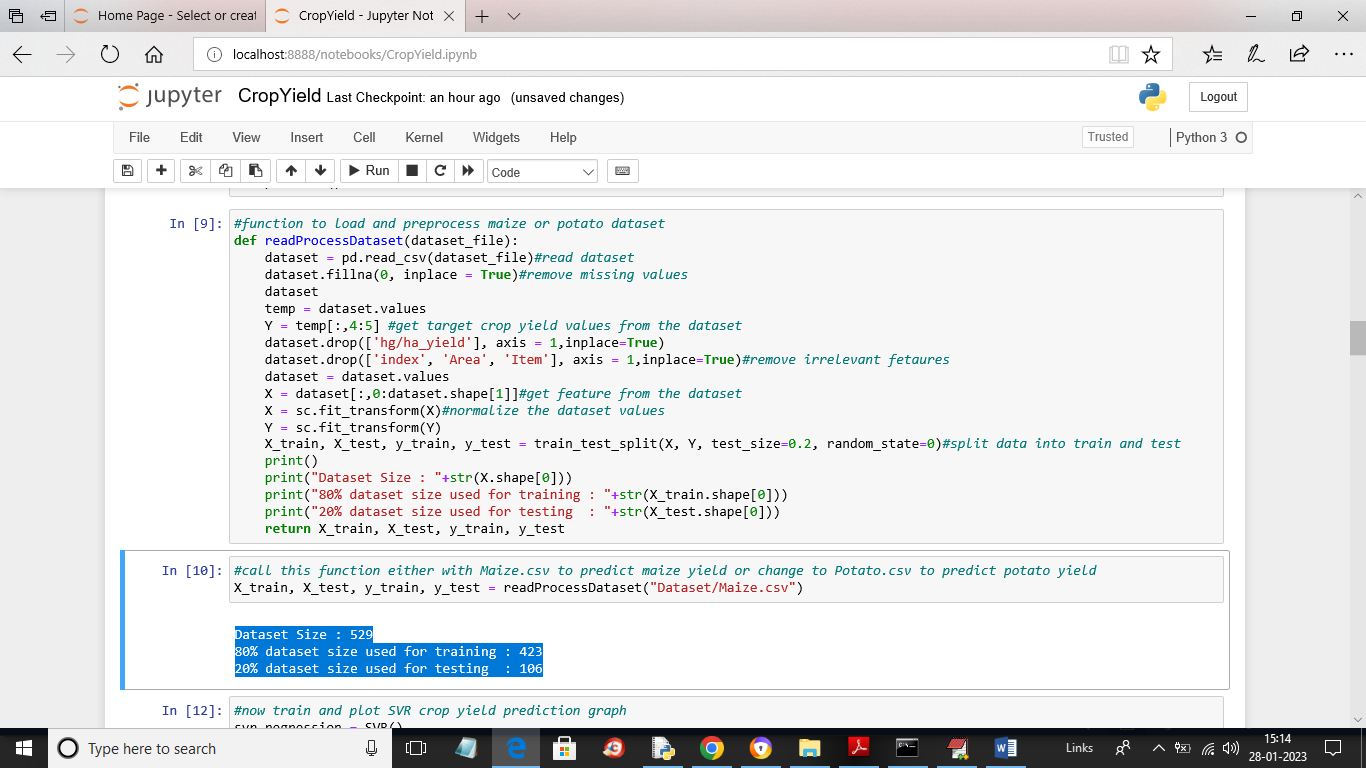
**Fig.5.3 Importing all require python packages**

In above screen we are importing all require python packages

****

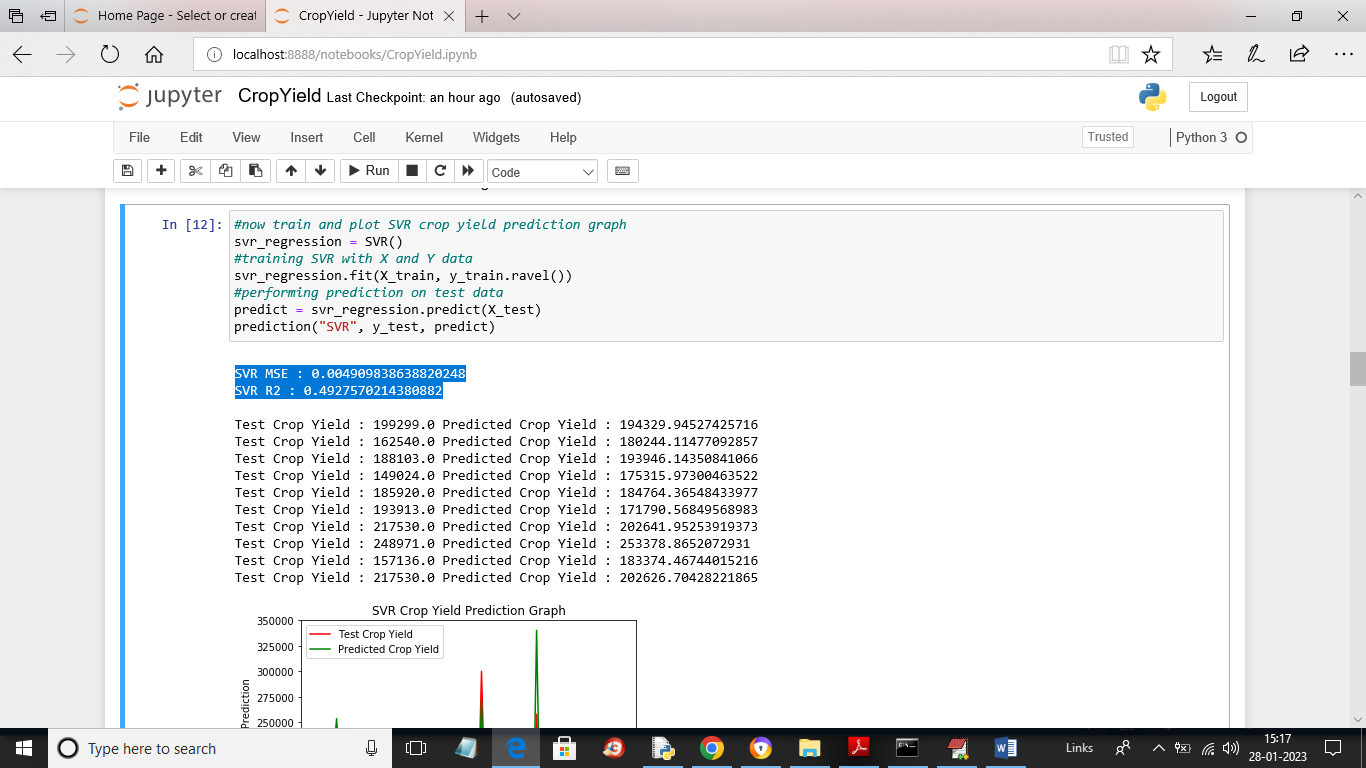
**Fig.5.4 defining MINMAX and defining MSE**

In above screen in first block defining MINMAX class to normalize dataset values and then defining MSE and square variables and then defining function predict and plot Crop Yield graph

****

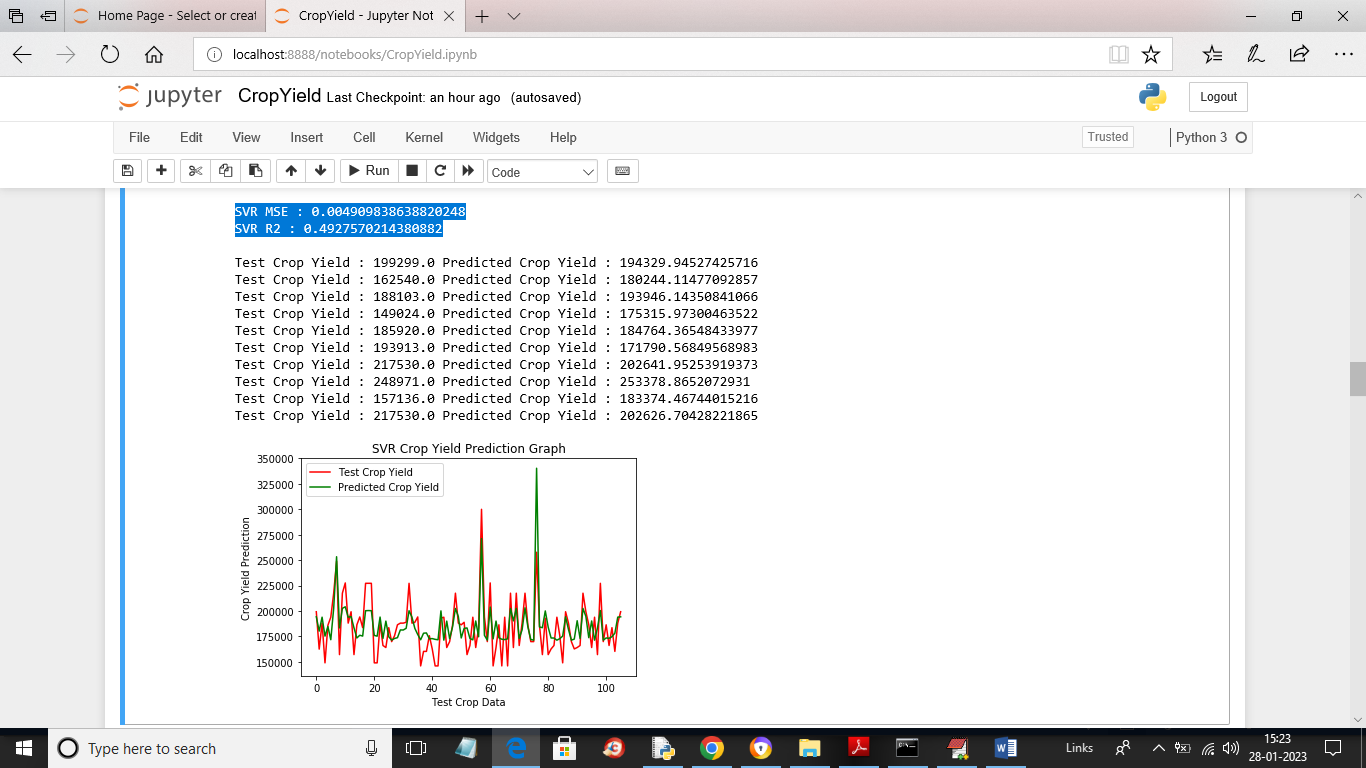
**Fig.5.5 Read dataset function**

In above screen we are calling Read dataset function to read and process dataset and then get train and test data and now using train and test data we will train and test performance of all algorithms and in blue colour we can see total dataset size and training and testing size

****

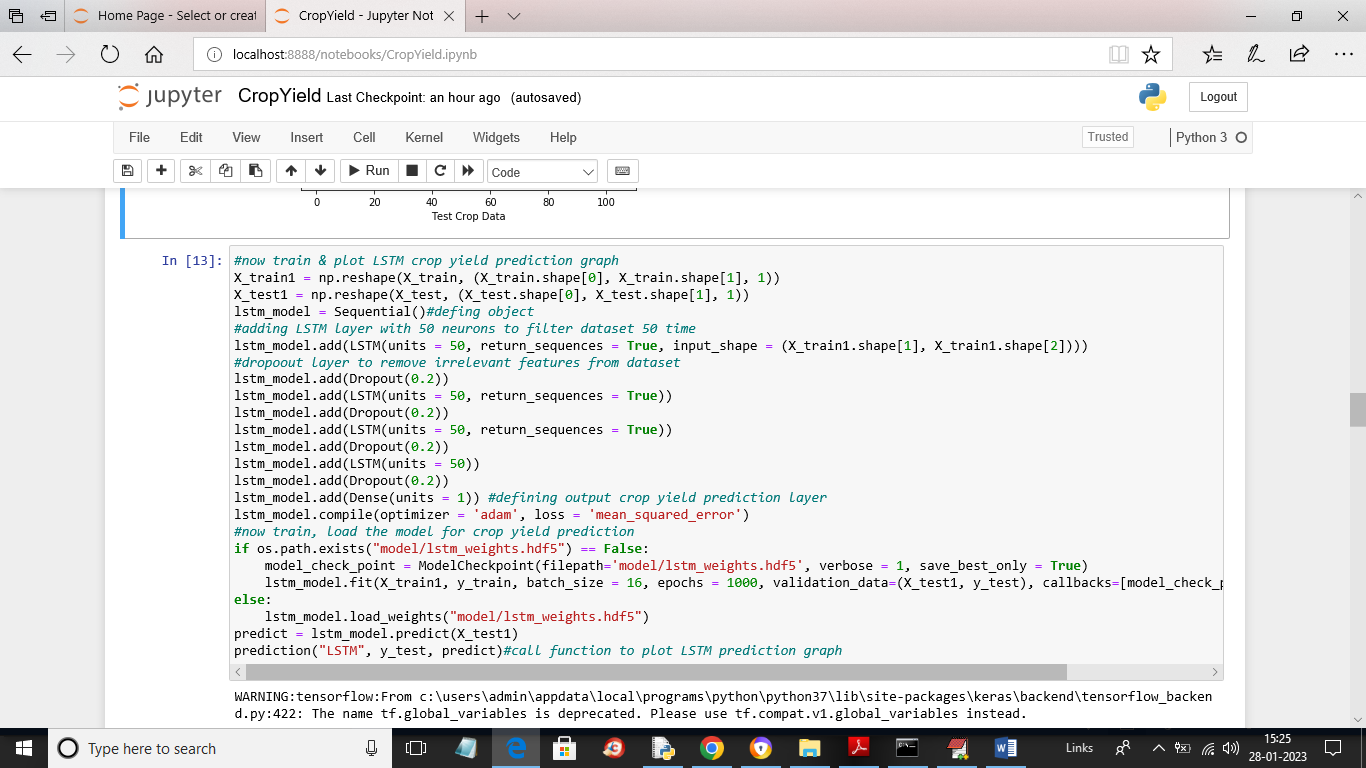
**Fig.5.6 training with SVR algorithm**

In above screen we are training with SVR algorithm and then in blue colour text we can see SVR MSE and R2 values and then in next lines we can see TEST crop yield values and SVR predicted crop yield values and in below screen we can see Test yield and predicted yield graph

****

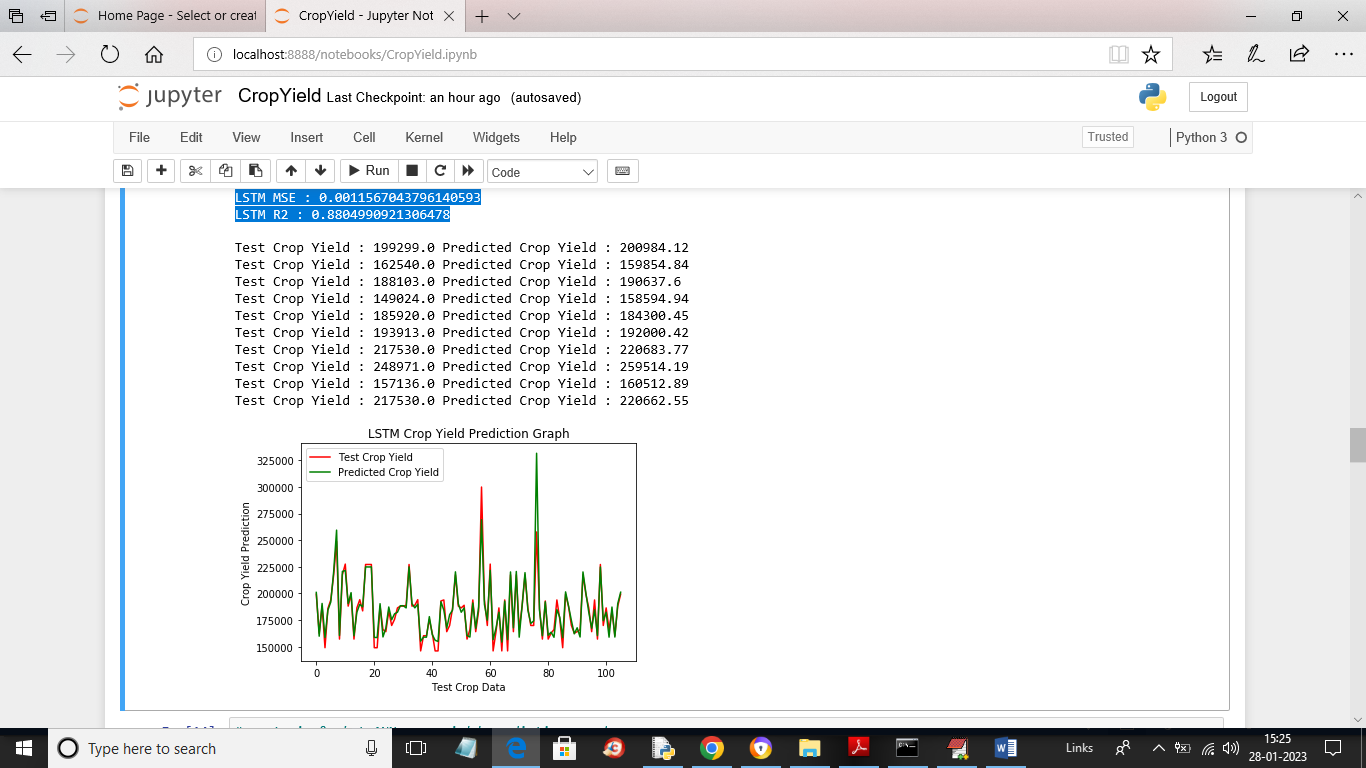
**Fig.5.7 Graphical representation**

In above screen x-axis represents number of test records and y-axis represents Yield values where red line represents TEST yield and green line represents predicted yield and in above graph we can see there lots of gap between red and green line so SVR prediction is not accurate

****

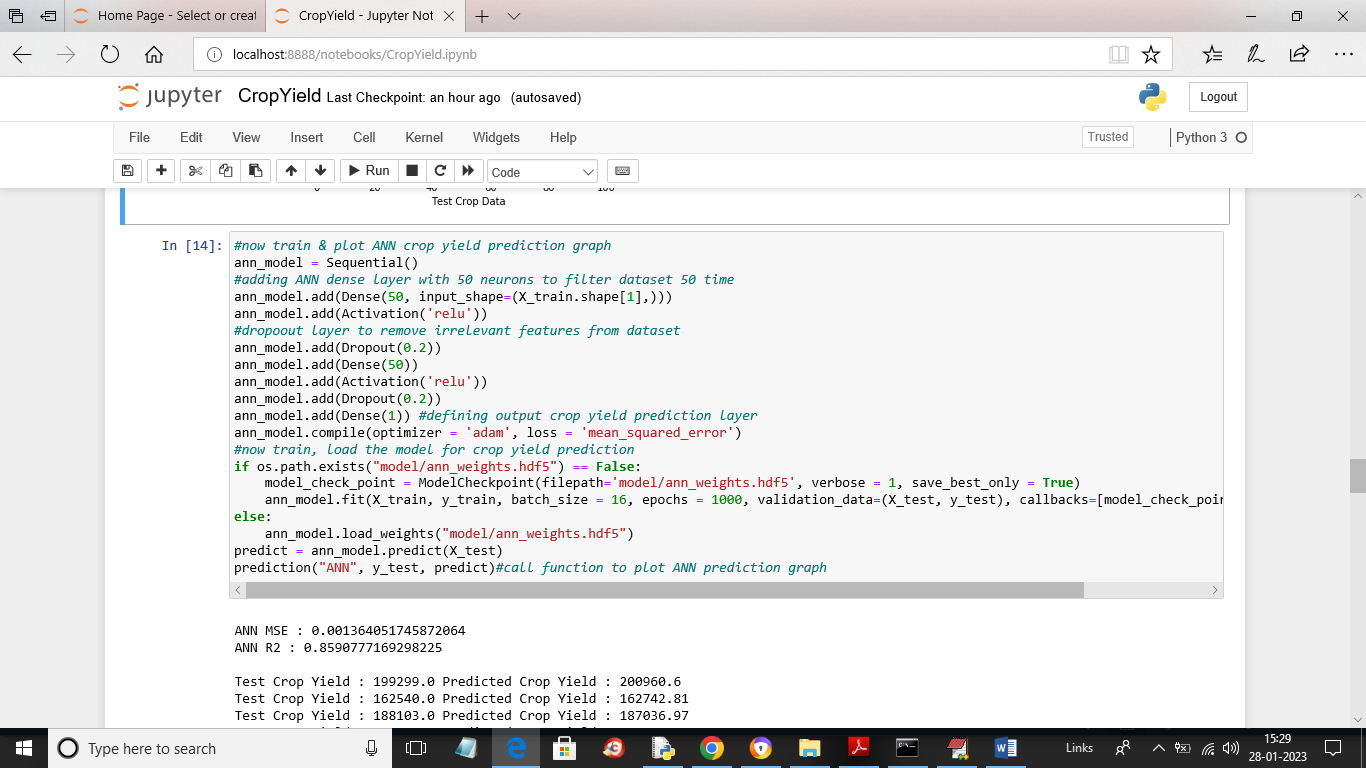
**Fig.5.8 defining and training with LSTM**

In above screen we are defining and training with LSTM and after executing above block will get below output

****

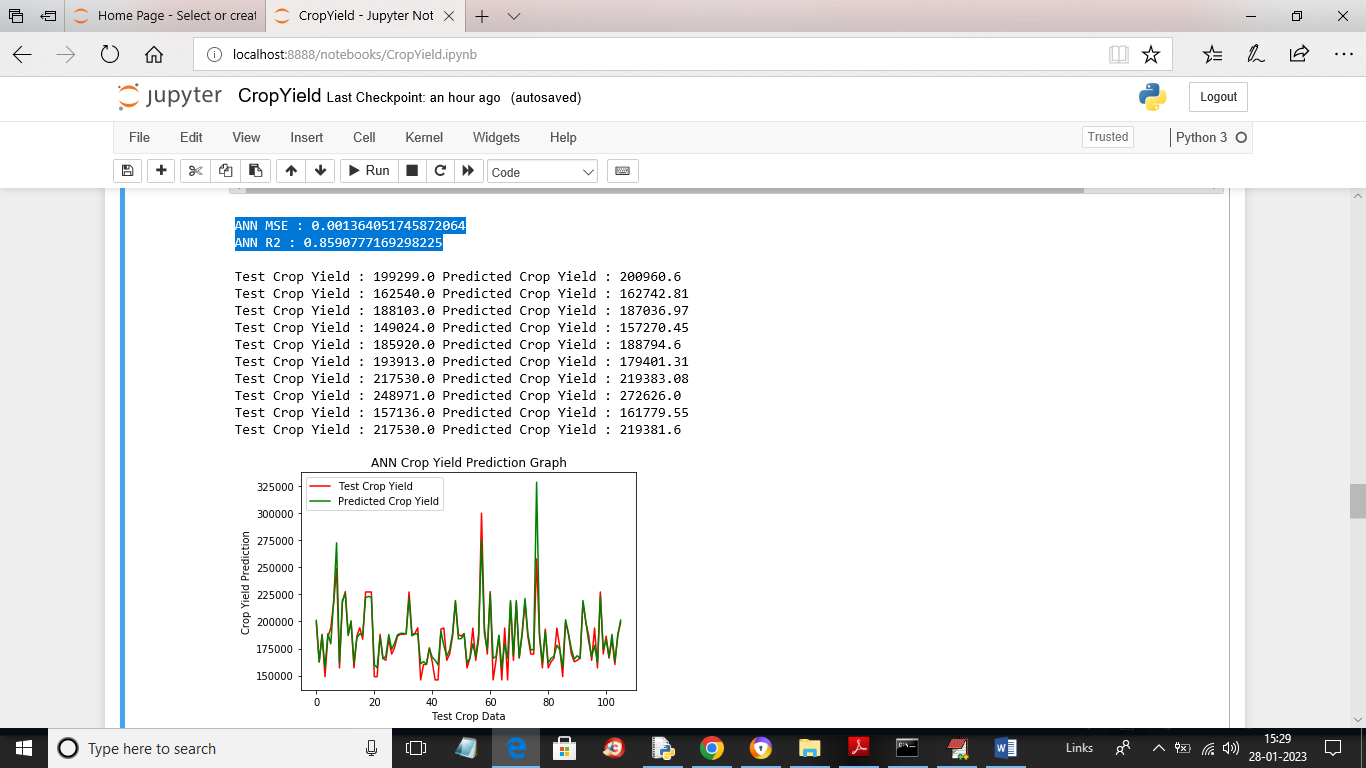
**Fig.5.9 LSTM MSE and R2 values**

In above screen in blue colour we can see LSTM MSE and R2 values and then in next lines we can see TEST crop yield and LSTM predicted crop yield and in LSTM graph we can see both green line and red is fully overlapping so LSTM prediction is accurate

****

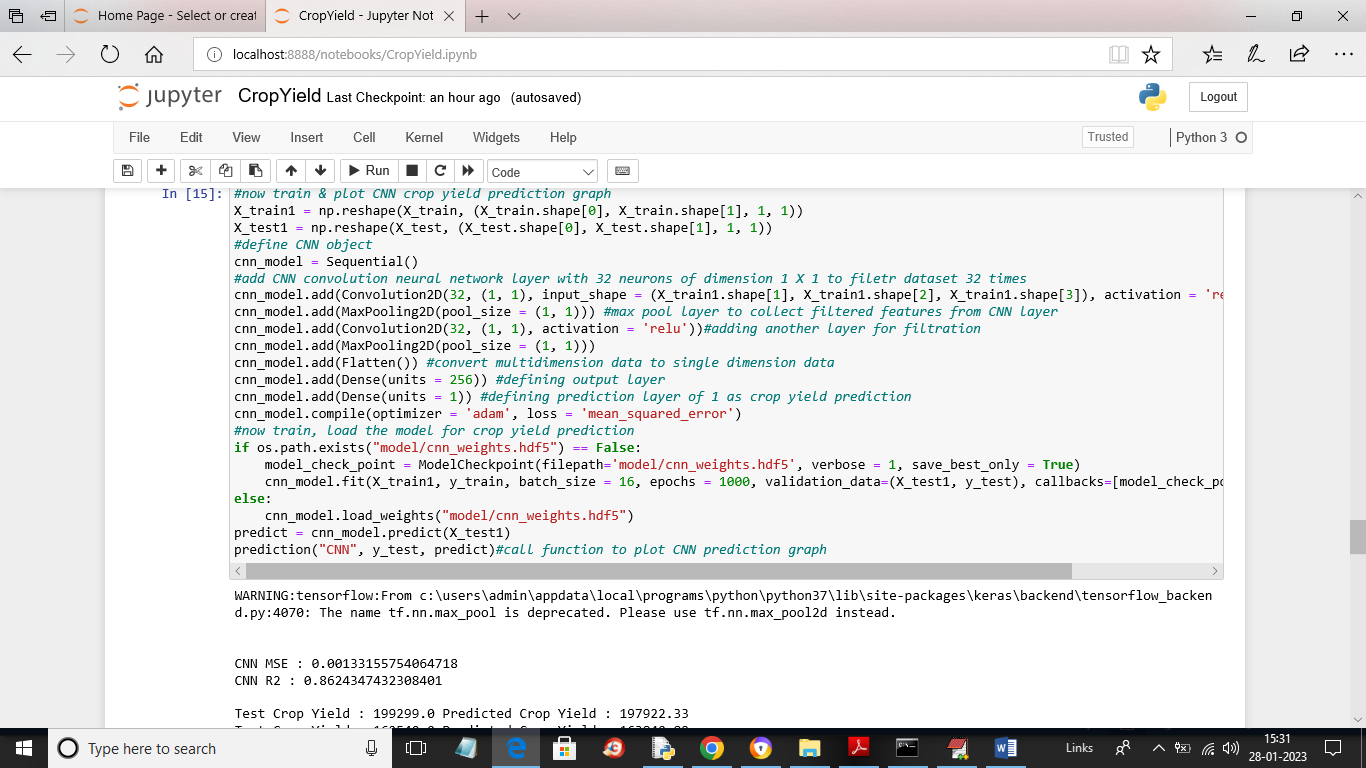
**Fig 5.10 training with ANN model**

In above screen we are training with ANN model and after executing above block will get below output

****

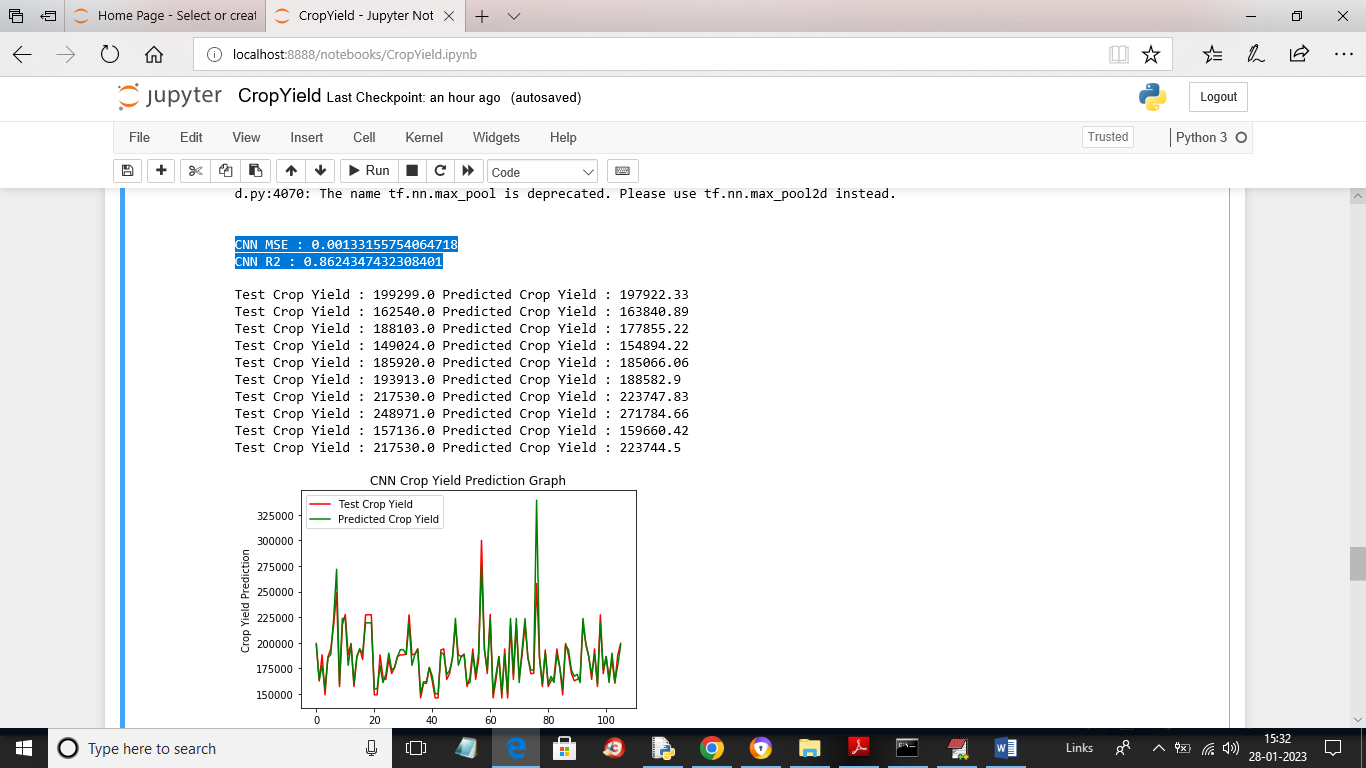
**Fig.5.11 ANN MSE and R2 values**

In above screen in blue colour text we can see ANN MSE and R2 values and then in next lines we can see TEST and predicted crop yield for ANN and then in ANN graph we can see both lines are fully overlapping so ANN prediction also accurate

****

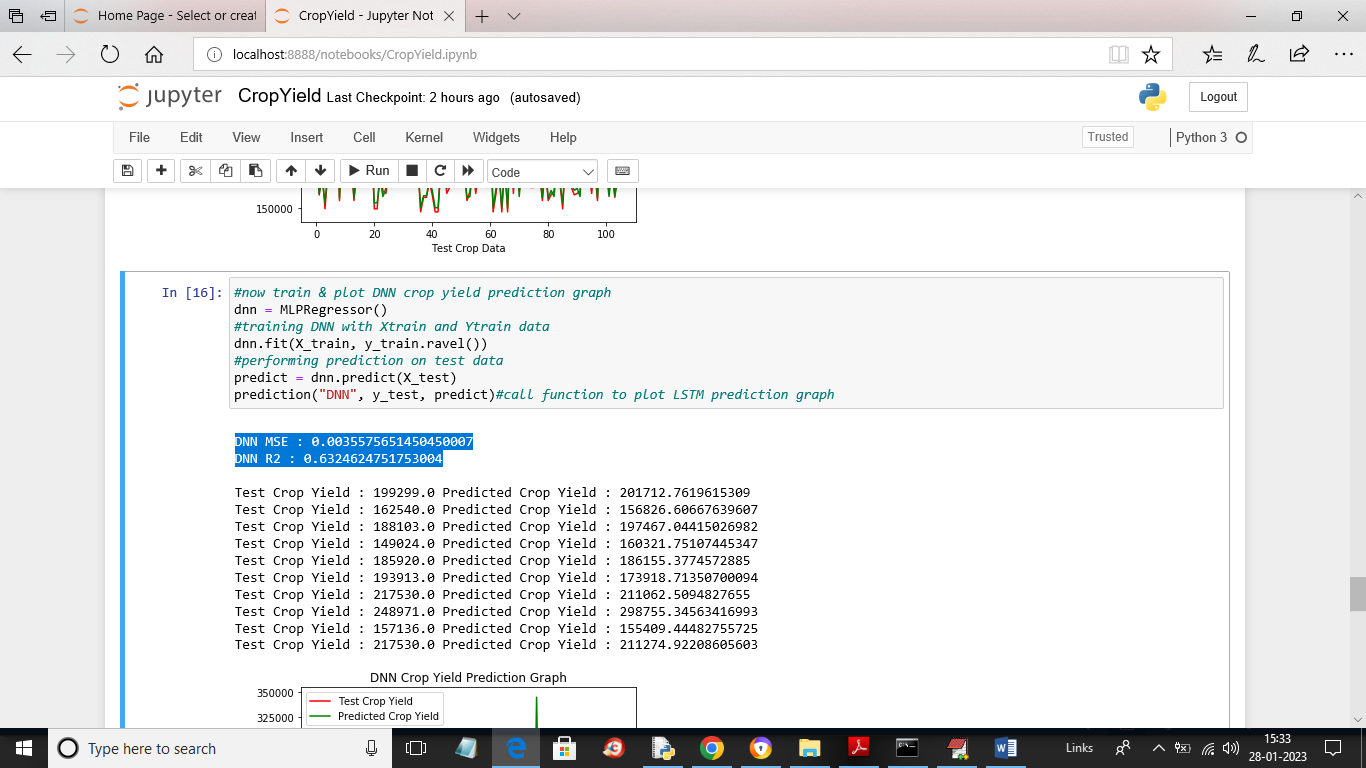
**Fig.5.12 training with CNN**

In above screen we are training with CNN and after executing above block will get below output

****

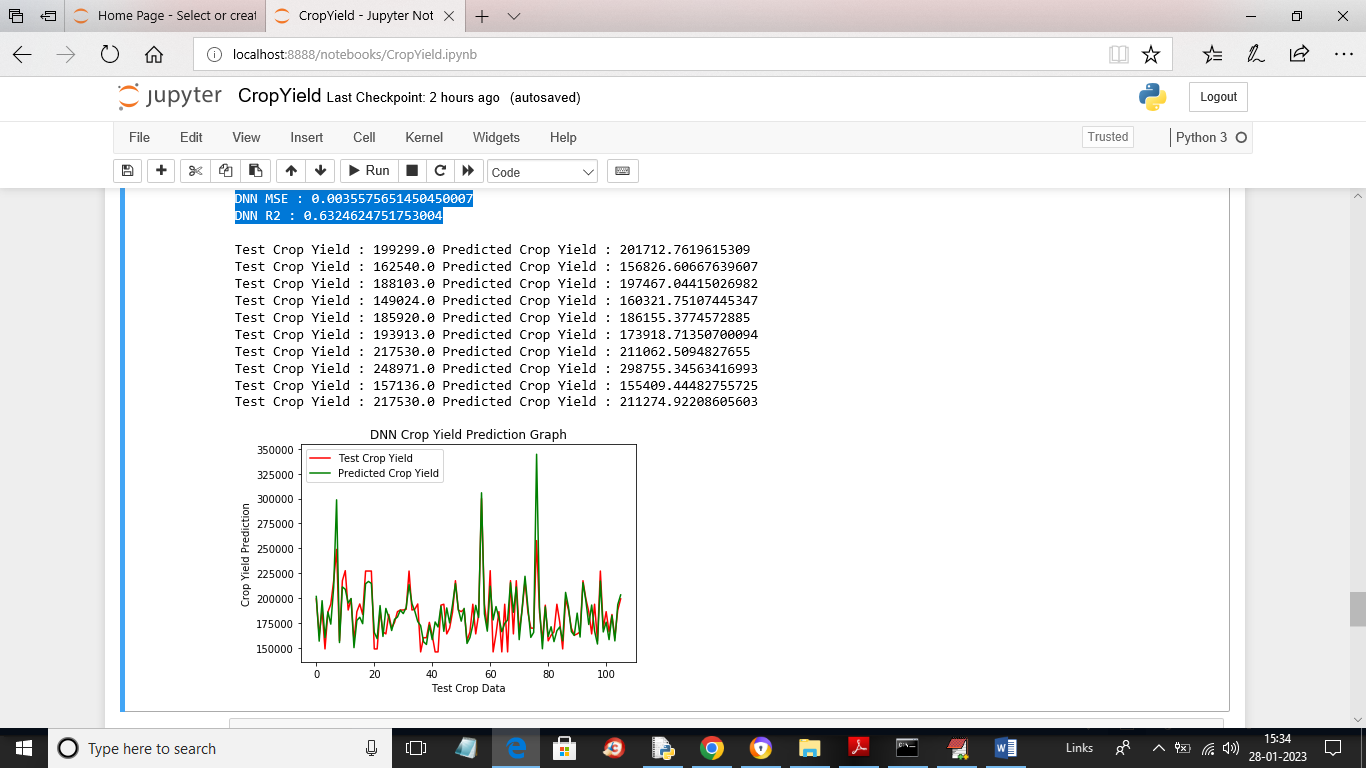
**Fig.5.13 CNN MSE and R2 values**

In above screen we can see CNN MSE and R2 values and then we can see TEST and predicted yield for CNN and then in graph we can see both lines are fully overlapping so CNN prediction is also accurate

****

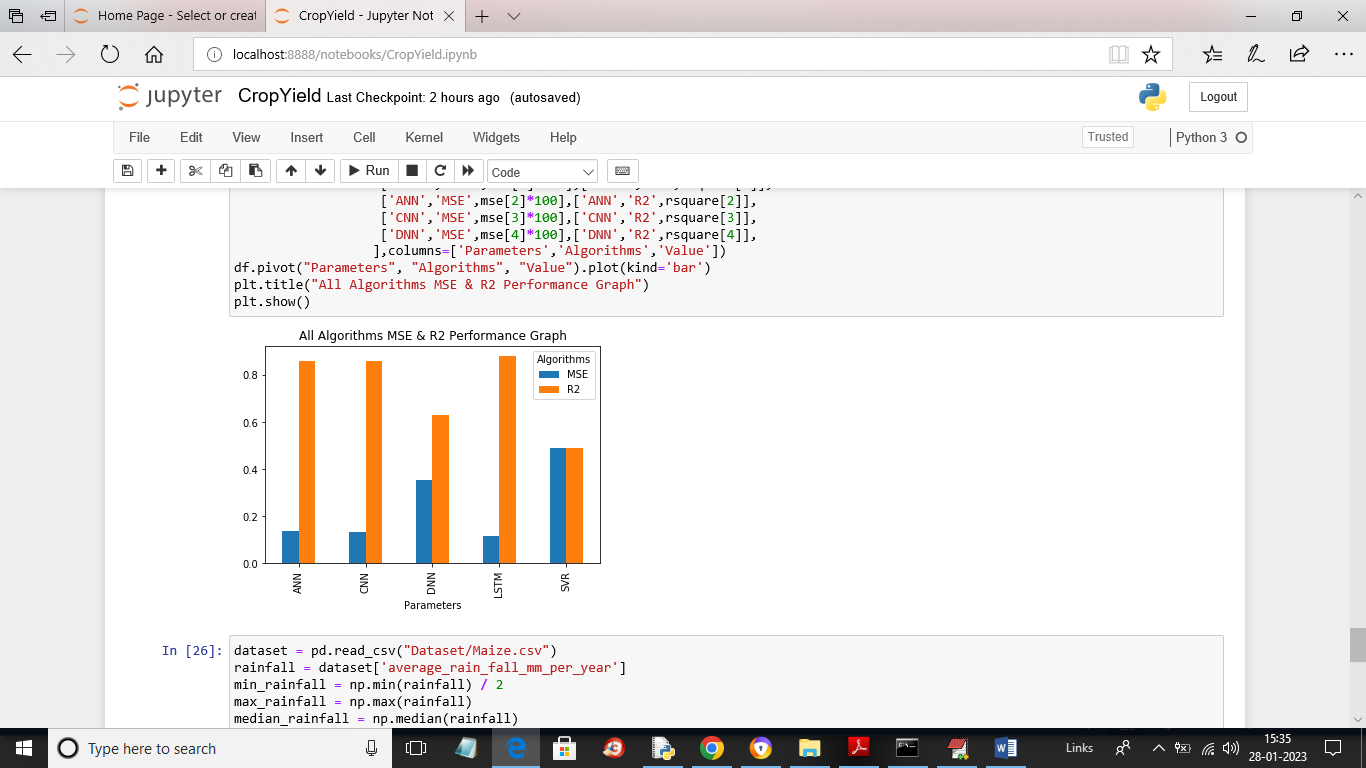
**Fig.5.14 training with DNN**

In above screen we are training with DNN and then we can see DNN MSE and R2 values and then we can see test and predicted crop yield for DNN

****

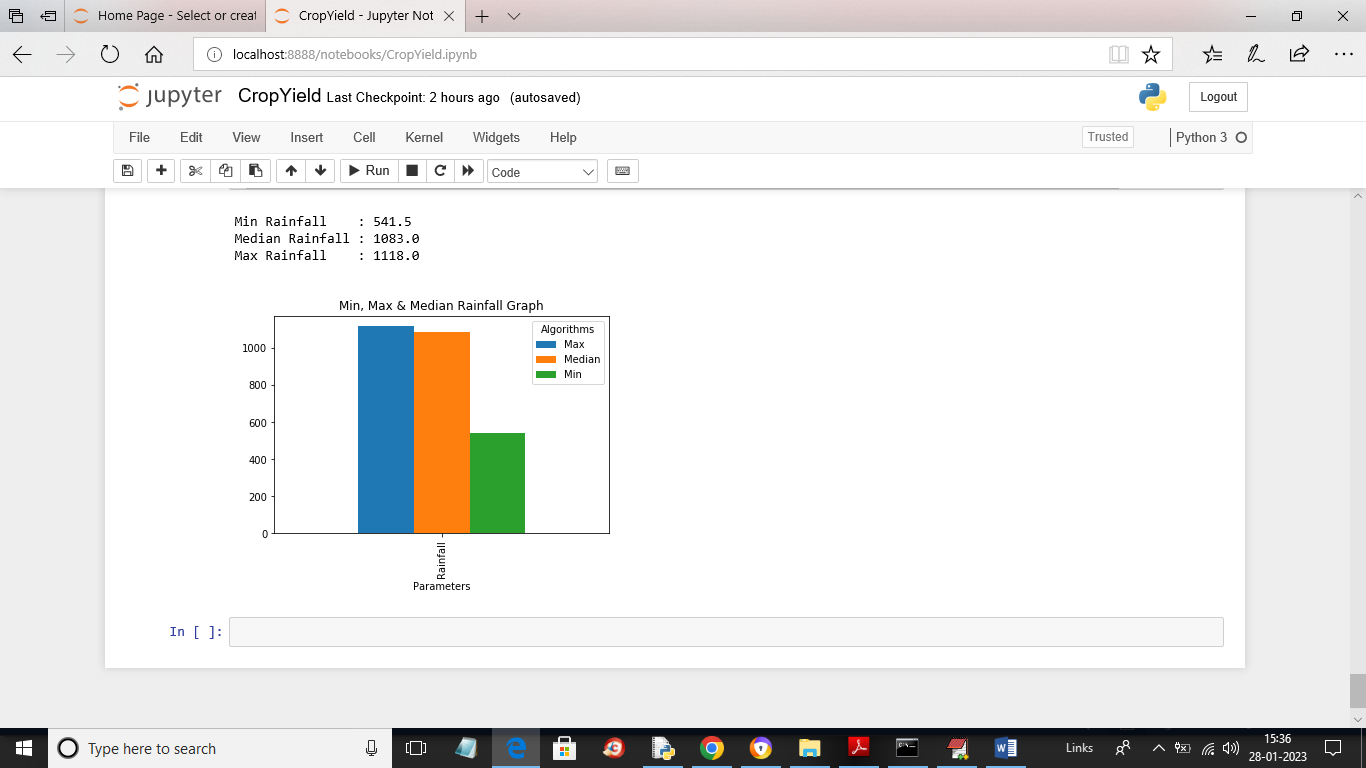
**Fig.5.15 DNN graph**

In above DNN graph there is little difference in red and green line as its contains some gap so DNN prediction is good but not accurate

****

**Fig.5.16 graph representation**

In above graph blue bar represents MSE and orange bar represents R2 and x-axis represents algorithm names and y-axis represents values and in all algorithms we can see LSTM got high R2 and less MSE compare to all algorithms so we can say LSTM is good at crop yield prediction.

****

**Fig.5.17 plotting MIN, MAX and median**

In above graph we are plotting MIN, MAX and median rainfall graph found in the dataset

**CHAPTER 6**

**CONCLUSION**

The research on crop yield prediction using machine learning models holds promise for transforming agriculture. In this project we are using machine learning and deep learning algorithms to predict future crop yield based on weather data such as temperature and rainfall. In proposed method we are using Irish Maize and Potato yield dataset to train all machine learning models and then these models can be used to predict future crop yield. In proposed method we are using random forest, SVR, **DNN, CNN, ANN and LSTM**. So, we have implemented all 6 algorithms on both datasets. To evaluate performance of each algorithm we are calculating MSE and R2 Score where MSE refers to mean square error.

**CHAPTER 7**

**ADVANTAGES AND APPLICATIONS**

**Advantages**

1. Higher accuracy
2. Time and cost will be saved
3. Values of yield will be same every time

**Applications**

1. Agricultural departments
2. Farmers

**REFERENCES**

[1] Aruvansh Nigam, Saksham Garg, Archit Agrawal “Crop Yield Prediction using ML Algorithms “, 2019

[2] Leo Brieman, “Random Forests”, 2001

[3] Priya, P., Muthaiah, U., Balamurugan, M.” Predicting Yield of the Crop Using Machine Learning Algorithm”, 2015

[4] Mishra, S., Mishra, D., Santra, G. H., “Applications of machine learning techniques in agricultural crop production”, 2016

[5] Dr. Jeevan Kumar, “Supervised Learning Approach for Crop Production”, 2020

[6] Ramesh Medar,Vijay S, Shweta, “Crop Yield Prediction using Machine Learning Techniques”, 2019

[7] Ranjini B Guruprasad, Kumar Saurav, Sukanya Randhawa,”Machine Learning Methodologies for Paddy Yield Estimation in India: A CASE STUDY”, 2019

[8] Sangeeta, Shruthi G, “Design and Implementation of Crop Yield Prediction Model In Agriculture”, 2020

[9] <https://power.larc.nasa.gov/data-access-viewer/>

[10] <https://en.wikipedia.org/wiki/Agriculture>

[11] <https://www.ibm.com/weather>

[12] https;//flutter.dev

[13] <https://openweathermap.org>

[14] https;//builtin.com/data-science/random-forest-algorithm

[15] https://tutorialspoint/machine-learning/logistic-regression

[16] <http://scikit-learn.org/modules/naive-bayes>

[17] Ebrahimi, M.A., Khoshtaghaza, M.H., Minaei, S., Jamshidi, B., Vision-based pest detection based on SVM classification method. Comput. Electron. Agric., 137, 52–58, 2017.

[18]. Chung, C.L., Huang, K.J., Chen, S.Y., Lai, M.H., Chen, Y.C., Kuo, Y.F., Detecting Bakanae disease in rice seedlings by machine vision. Comput. Electron. Agric., 121, 404–411, 2016.

[19]. Pantazi, X.E., Moshou, D., Oberti, R., West, J., Mouazen, A.M., Bochtis, D., Detection of biotic and abiotic stresses in crops by using hierarchical self-organizing classifiers. Précis. Agric., 18, 383–393, 2017.

[20]. Moshou, D., Bravo, C., West, J., Wahlen, S., McCartney, A., Automatic detection of “yellow rust” in wheat using reflectance measurements and neural networks. Comput. Electron. Agric., 44, 173–188, 2004.

[21]. Richardson, A., Signor, B.M., Lid bury, B.A., Bad rick, T., Clinical chemistry in higher dimensions: Machine-learning and enhanced prediction from routine clinical chemistry data. Clin. Biochem. 49, 1213–1220, 2016.

[22]. Wildenhain, J., Spitzer, M., Dolma, S., Jarvik, N., White, R., Roy, M., Griffiths, E., Bellows, D.S., Wright, G.D., Tyers, M., Prediction of Synergism from Chemical-Genetic Interactions by Machine Learning. Cell Syst., 1, 383–395, 2015.

[23]. Kang, J., Schwartz, R., Flickinger, J., and Beriwal, S., Machine learning approaches for predicting radiation therapy outcomes: A clinician’s perspective. Int. J. Radiat. Oncol. Biol. Phys., 93, 1127–1135, 2015.

[24]. Craven, B.D. and Islam, S.M.N., Ordinary least-squares regression SAGE. Dict. Quant. Manag. Res., 224–228, 2011.

[25]. Friedman, J.H., Multivariate Adaptive Regression Splines. Ann. Stat., 19, 1–67, 1991.

[26]. Alonso, J., Villa, A., Bahamonde, A., Improved estimation of bovine weight trajectories using Support Vector Machine Classification. Comput. Electron. Agric., 110, 36–41, 2015.

**APPENDIX**

**PYTHON**

**1.1 Introduction**

\* One of the most popular languages is Python. Guido van Rossum released this language in 1991. Python is available on the Mac, Windows, and Raspberry Pi operating systems. The syntax of Python is simple and identical to that of English. When compared to Python, it was seen that the other language requires a few extra lines.

\*It is an interpreter-based language because code may be run line by line after it has been written. This implies that rapid prototyping is possible across all platforms. Python is a big language with a free, binary-distributed interpreter standard library.

\* It is inferior to maintenance that is conducted and is straightforward to learn. It is an object-oriented, interpreted programming language. It supports several different programming paradigms in addition to object-oriented programming, including functional and procedural programming.

\* It supports several different programming paradigms in addition to object-oriented programming, including practical and procedural programming. Python is mighty while maintaining a relatively straightforward syntax. Classes, highly dynamic data types, modules, and exceptions are covered. Python can also be utilised by programmes that require programmable interfaces as an external language.

Here are some key features and characteristics of Python:

* Readability: Python emphasizes code readability with its clean and intuitive syntax. It uses indentation and whitespace to structure code blocks, making it easy to understand and maintain.
* Easy to Learn: Python's simplicity and readability make it an excellent choice for beginners. Its straightforward syntax and extensive documentation make it accessible for newcomers to programming.
* Interpreted Language: Python is an interpreted language, meaning that it doesn't need to be compiled before running. The Python interpreter reads and executes the code directly, making the development process faster and more interactive.
* Cross-platform Compatibility: Python is available for major operating systems like Windows, macOS, and Linux. This cross-platform compatibility allows developers to write code once and run it on different platforms without modifications.
* Large Standard Library: Python comes with a vast standard library that provides ready-to-use modules and functions for various tasks. It covers areas such as file I/O, networking, regular expressions, databases, and more, saving developers time and effort.
* Extensible and Modular: Python supports modular programming, enabling developers to organize code into reusable modules and packages. Additionally, Python allows integrating modules written in other languages, such as C or C++, providing flexibility and performance optimizations.
* Wide Range of Libraries and Frameworks: Python has a vibrant ecosystem with numerous third-party libraries and frameworks. These libraries, such as NumPy, pandas, TensorFlow, and Django, extend Python's capabilities for specific domains, making it a powerful tool for diverse applications.
* Object-Oriented: Python supports object-oriented programming (OOP) principles, allowing developers to create and work with classes and objects. OOP provides a structured approach to code organization, promoting code reuse and modularity.
* Dynamic Typing: Python is dynamically typed, meaning variable types are determined at runtime. Developers do not need to declare variable types explicitly, which enhances flexibility and simplifies code writing.

**1.2 Installation**

To install Python on your computer, follow these basic steps:

* Step 1: Visit the Python website Go to the official Python website at <https://www.python.org/>.
* Step 2: Select the operating system Choose the appropriate installer for your operating system. Python supports Windows, macOS, and various Linux distributions. Make sure to select the correct version that matches your operating system.
* Step 3: Check which version of Python is installed; if the 3.7.0 version is not there, uninstall it through the control panel and
* Step 4: Install Python 3.7.0 using Cmd.
* Step 5: Install the all libraries that required to run the project
* Step 6: Run

**1.3 Python Features:**

1. **Easy:** Because Python is a more accessible and straightforward language, Python programming is easier to learn.
2. **Interpreted language:** Python is an interpreted language, therefore it can be used to examine the code line by line and provide results.
3. **Open Source:** Python is a free online programming language since it is open-source.
4. **Portable:** Python is portable because the same code may be used on several computer standard
5. **libraries:** Python offers a sizable library that we may utilize to create applications quickly.
6. **GUI:** It stands for GUI (Graphical User Interface)
7. **Dynamical typed:** Python is a dynamically typed language, therefore the type of the value will be determined at runtime.

**1.4 Python GUI (Tkinter)**

* Python provides a wide range of options for GUI development (Graphical User Interfaces).
* Tkinter, the most widely used GUI technique, is used for all of them.
* The Tk GUI toolkit offered by Python is used with the conventional Python interface.
* Tkinter is the easiest and quickest way to write Python GUI programs.
* Using Tkinter, creating a GUI is simple.
* A part of Python's built-in library is Tkinter. The GUI programs were created.
* Python and Tkinter together give a straightforward and quick way. The Tk GUI toolkit's object-oriented user interface is called Tkinter.

Making a GUI application is easy using Tkinter. Following are the steps:

1) Install the Tkinter module in place.

2) The GUI applicatioMakeske the primary window

3) Include one or more of the widgets mentioned above in the GUI application.

4) Set up the main event loop such that it reacts to each user-initiated event.

Although Tkinter is the only GUI framework included in the Python standard library, Python includes a GUI framework. The default library for Python is called Tkinter. Tk is a scripting language often used in designing, testing, and developing GUIs. Tk is a free, open-source widget toolkit that may be used to build GUI applications in a wide range of computer languages.

**1.5 Python IDLE**

* Python IDLE offers a full-fledged file editor, which gives you the ability to write and execute Python programs from within this program. The built-in file editor also includes several features, like code completion and automatic indentation, that will speed up your coding workflow.
* Guido Van Rossum named Python after the British comedy group Monty Python while the name IDLE was chosen to pay tribute to Eric Idle, who was one of the Monty Python's founding members. IDLE comes bundled with the default implementation of the Python language since the 01.5. 2b1 release
* IDLE is used to execute statements similar to Python Shell. IDLE is used to create, modify, and execute Python code. IDLE provides a fully-featured text editor to write Python scripts and provides features like syntax highlighting, auto-completion, and smart indent.
* IDLE has two modes: interactive and script. We wrote our first program, “Hello, World!” in interactive mode. Interactive mode immediately returns the results of commands you enter into the shell. In script mode, you will write a script and then run it.
* The IDE Python IDLE is a good place to start as it helps you become familiar with the way Python works and understand its syntax. This IDE is good to start programming in Python due to its great debugger, but once you are fluent and start developing projects it is necessary to jump to another, more complete IDE.
* Python IDLE (Integrated Development and Learning Environment) is an interactive development environment included with the Python programming language. It provides a convenient way to write, execute, and debug Python code.

When you install Python, IDLE is typically installed along with it. To open IDLE, you can follow these steps:

* Open the command prompt (Windows) or terminal (macOS/Linux).
* Type "idle" and press Enter. Alternatively, you can specify the version with "idle3" or "idle2" for Python 3 or Python 2, respectively.
* Once IDLE is launched, you will see the Python shell, which is an interactive environment where you can type and execute Python code directly.

Here are some features and functionalities provided by Python IDLE:

* Editor: IDLE includes a text editor where you can write your Python code. It offers syntax highlighting, automatic indentation, and code completion to enhance your coding experience.
* Interactive Shell: The Python shell in IDLE allows you to execute Python code interactively. You can type commands, statements, or function calls directly in the shell, and Python will execute them immediately.
* Debugging: IDLE provides basic debugging capabilities to help you find and fix errors in your code. You can set breakpoints, step through code, inspect variables, and track the program's execution.
* Python Help: IDLE provides access to the Python documentation and built-in help. You can access the help menu to find information about Python modules, functions, classes, and more.
* Script Execution: In addition to the interactive shell, IDLE allows you to run Python scripts stored in files. You can write your code in the editor and execute it as a script to see the output or interact with the program.
* Customization: IDLE can be customized to suit your preferences. You can modify settings related to syntax highlighting, indentation, fonts, and more.
* Python IDLE serves as a beginner-friendly development environment and learning tool. It is suitable for writing small scripts, testing code snippets, experimenting with Python features, and learning the language's basics. However, for more advanced development projects, you may consider using other code editors or integrated development environments (IDEs) that provide additional features and better project management capabilities.

**1.6 Libraries**

In Python, libraries (also referred to as modules or packages) are collections of pre-written code that provide additional functionality and tools to extend the capabilities of the Python language. Libraries contain reusable code that developers can leverage to perform specific tasks without having to write everything from scratch.

Python libraries are designed to solve common problems, such as handling data, performing mathematical operations, interacting with databases, working with files, implementing networking protocols, creating graphical user interfaces (GUIs), and much more. They provide ready-to-use functions, classes, and methods that simplify complex operations and save development time.

**Libraries in Python offer various advantages:**

* Code Reusability:
* Efficiency:
* Collaboration
* Domain-Specific Functionality
* To use a Python library, you need to install it first.

There are some libraries following:

* **Pandas:**

Pandas are a Python computer language library for data analysis and manipulation. It offers a specific operation and data format for handling time series and numerical tables. It differs significantly from the release3-clause of the BSD license. It is a well-liked open-source of opinion that is utilized in machine learning and data analysis.

Pandas are a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python. Pandas are a Python library used for working with data sets.

* It has functions for analysing, cleaning, exploring, and manipulating data.
* The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.
* Pandas allow us to analyse big data and make conclusions based on statistical theories.
* Pandas can clean messy data sets, and make them readable and relevant.

Relevant data is very important in data science. Pandas are a Python library for data analysis. Started by Wes McKinney in 2008 out of a need for a powerful and flexible quantitative analysis tool, pandas have grown into one of the most popular Python libraries. It has an extremely active community of contributors. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself.

* **NumPy:**

The NumPy Python library for multi-dimensional, big-scale matrices adds a huge number of high-level mathematical functions. It is possible to modify NumPy by utilizing a Python library. Along with line, algebra, and the Fourier transform operations, it also contains several matrices-related functions.

NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate on these arrays and matrices.

* NumPy is a Python library used for working with arrays.
* It also has functions for working in domain of linear algebra, Fourier transform, and matrices.
* NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.
* NumPy stands for Numerical Python.
* In Python we have lists that serve the purpose of arrays, but they are slow to process.
* NumPy aims to provide an array object that is up to 50x faster than traditional Python lists.
* The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy.
* Arrays are very frequently used in data science, where speed and resources are very important.
* **Matplotlib:**

It is a multi-platform, array-based data visualization framework built to interact with the whole SciPy stack. MATLAB is proposed as an open-source alternative. Matplotlib is a Python extension and a cross-platform toolkit for graphical plotting and visualization.

Matplotlib is a popular Python library for creating static, animated, and interactive visualizations. It provides a flexible and comprehensive set of tools for generating plots, charts, histograms, scatter plots, and more. Matplotlib is widely used in various fields, including data analysis, scientific research, and data visualization.

Here are some key features and functionalities of the Matplotlib library:

* Plotting Functions
* Customization Options
* Multiple Interfaces
* Integration with NumPy and pandas
* Subplots and Figures:
* Saving and Exporting
* **Scikit-learn:**

The most stable and practical machine learning library for Python is scikit-learn. Regression, dimensionality reduction, classification, and clustering are just a few of the helpful tools it provides through the Python interface for statistical modeling and machine learning. It is an essential part of the Python machine learning toolbox used by JP Morgan. It is frequently used in various machine learning applications, including classification and predictive analysis.

Scikit-learn (also referred to as sklearn) is a widely used open-source machine learning library for Python. It provides a comprehensive set of tools and algorithms for various machine learning tasks, including classification, regression, clustering, dimensionality reduction, model selection, and pre-processing.

Here are some key features and functionalities of the Scikit-learn library:

* Easy-to-Use Interface:
* Broad Range of Algorithms:
* Data Pre-processing and Feature Engineering:
* Model Evaluation and Validation:
* Integration with NumPy and pandas:
* Robust Documentation and Community Support:
* **Keras:**

\* Google's Keras is a cutting-edge deep learning API for creating neural networks. It is created in Python and is designed to simplify the development of neural networks. Additionally, it enables the use of various neural networks for computation. Deep learning models are developed and tested using the free and open-source Python software known as Keras.

Keras is a high-level deep learning library for Python. It is designed to provide a user-friendly and intuitive interface for building and training deep learning models. Keras acts as a front-end API, allowing developers to define and configure neural networks while leveraging the computational backend engines, such as Tensor Flow or Theano.

Here are some key features and functionalities of the Keras library:

* User-Friendly API
* Multi-backend Support
* Wide Range of Neural Network Architectures
* Pre-trained Models and Transfer Learning:
* Easy Model Training and Evaluation:
* GPU Support:
* **h5py:**

\* The h5py Python module offers an interface for the binary HDF5 data format. Thanks to p5py, the top can quickly halt the vast amount of numerical data and alter it using the NumPy library. It employs common syntax for Python, NumPy, and dictionary arrays.

h5py is a Python library that provides a simple and efficient interface for working with datasets and files in the Hierarchical Data Format 5 (HDF5) format. HDF5 is a versatile data format commonly used for storing and managing large volumes of numerical data.

Here are some key features and functionalities of the h5py library:

* + HDF5 File Access
  + Dataset Handling:
  + Group Organization:
  + Attributes:
  + Compatibility with NumPy
  + Performance
* **Tensor flow**

TensorFlow is a Python library for fast numerical computing created and released by Google. It is a foundation library that can be used to create Deep Learning models directly or by using wrapper libraries that simplify the process built on top of TensorFlow. TensorFlow is an end-to-end open source platform for machine learning. TensorFlow is a rich system for managing all aspects of a machine learning system; however, this class focuses on using a particular TensorFlow API to develop and train machine learning models.

TensorFlow is a popular open-source library for machine learning and deep learning. It provides a comprehensive set of tools, APIs, and computational resources for building and training various types of machine learning models, especially neural networks.

Here are some key features and functionalities of TensorFlow:

* Neural Network Framework:
* Computational Graphs
* Automatic Differentiation
* GPU and TPU Support
* Distributed Computing
* Deployment Capabilities
* **Tkinter**

Tkinter is an acronym for "Tk interface". Tk was developed as a GUI extension for the Tcl scripting language by John Ousterhout. The first release was in 1991. Tkinter is the de facto way in Python to create Graphical User interfaces (GUIs) and is included in all standard Python Distributions. In fact, it's the only framework built into the Python standard library.

Tkinter is a standard Python library used for creating graphical user interfaces (GUIs). It provides a set of modules and classes that allow you to develop interactive and visually appealing desktop applications.

Here are some key features and functionalities of Tkinter:

* Cross-Platform Compatibility
* Simple and Easy-to-Use
* Widgets and Layout Management
* Event-Driven Programming
* Customization and Styling
* Integration with Other Libraries
* **NLTK**

NLTK is a toolkit build for working with NLP in Python. It provides us various text processing libraries with a lot of test datasets. A variety of tasks can be performed using NLTK such as tokenizing, parse tree visualization, etc NLTK (Natural Language Toolkit) is the go-to API for NLP (Natural Language Processing) with Python. It is a really powerful tool to pre-process text data for further analysis like with ML models for instance. It helps convert text into numbers, which the model can then easily work with.

NLTK (Natural Language Toolkit) is a Python library widely used for working with human language data and implementing natural language processing (NLP) tasks. It provides a set of tools, corpora, and resources for tasks such as tokenization, stemming, tagging, parsing, sentiment analysis, and more.

Here are some key features and functionalities of NLTK:

* Text Processing
* Part-of-Speech Tagging
* Named Entity Recognition
* Chunking and Parsing
* Sentiment Analysis:
* WordNet Integration:
* **Scipy**

SciPy is a collection of mathematical algorithms and convenience functions built on the NumPy extension of Python. It adds significant power to the interactive Python session by providing the user with high-level commands and classes for manipulating and visualizing data.

SciPy is a powerful scientific computing library for Python that provides a wide range of mathematical algorithms and functions. It builds upon NumPy, another fundamental library for numerical computing, and extends its capabilities by adding additional tools for scientific and technical computing tasks.

Here are some key features and functionalities of SciPy:

* Numerical Integration:
* Optimization and Root Finding
* Linear Algebra
* Signal and Image Processing
* Statistics