Section A-Research Paper
ISSN 2063-5346

Identifying Key Crop Performance traits for improving yield using Data Mining

Yogesh Gandge
Department of Master of Computer Applications,
B.K.I.T. Bhalki,
Bidar, India
yogeshvin22@gmail.com

Dr. P. Sandhya
Department of Studies in Computer Applications,
Visvesvaraya Technological University,
Post graduate Centre, Mysuru, India.
sanjoshi17@yahoo.com

Abstract: Crop performance traits are various such as climatic condition, soil macro and micro nutrients. The interactions between these have an effect on yield. There is a fundamental relationship between the crop yield and the interactive factors like the climate and the soil nutrients.One needs to understand this complex correlation. To study such an association, it requires large datasets and optimized datamining algorithms to discover the relationships. This is where Data mining plays a crucial role in the area of identifying key crop performance traits for improving yield.

In this study an attempt is being made to identify the key trait responsible for high Soyabean crop yield.The dataset was obtained from KVK Bidar of past 04 years on Rainfall,Temperature, pH and suitable crop .Dataset consisting of 524 records were obtained. After preprocessing the data,the dataset was split into training data and test data.Outliers were removed, skewness was checked.Data was checked for a balanced dataset.Using Datamining algorithms the correlations among these input features were obtained. Random Forest was employed. Using GridSearchCV optimal parameters were obtained for the classifiers.Through this study it was identified that pH is having high influence on the Soyabean crop yield as compared to rainfall and temperature. Model was tested and the result obtained was 98%.

Keywords:Data mining,Machine Learning,Crop prerformance traits

I.Introduction

Crop yield, is effected by several parameters. These parameters are (i)agricultural practices, (ii)biological aspects like insects, pests, weeds and (iii)environmental parameters like climatic condition, soil nutrients, pH etc. It has been demonstrated that every 1% increase in agricultural yield translates into a 0.6–1.2% decrease in the number of poor households in the world ***[7]***.

Due to global warming, climate is changing rapidly and investment of money in agriculture has also increased . Due to factors like climatic condition, lack of soil nutrients and

over or underuse of fertilizers the farmers are not getting proper yield. High Crop yield is an integral part of agriculture and is majorly effected by soil,***[4][9]*** environmental features like rainfall, temperature, pH, etc. These features differ across large geographical area. Therefore the farmers are unable to cultivate the same yield in every area.

Accurately identifying the best traits for crop production, based on soil and environmental factors,[11] is crucial for agricultural yield. It is one of the research area. Most of the available techniques use machine learning (ML) for crop yield estimation or prediction but very little has been done to predict the region-specific traits which are crucial for high crop yield. Parameters such as soil type, nutrients content like N,P,K, micronutrients like zinc, boron, and manganese, temperature, rainfall, pH influence crop cultivation.These parameters differ for different regions, thus we need to collect region specific dataset on the features mentioned above. Therefore there is a need to identify the most important features that can maximize the crop yield.

In our study, a dataset of **04 years** on rainfall, temperature, pH and the respective crop has been collected. Dataset consisting of 524 records were obtained. From this huge dataset the major challenge is to identify the key attributes that influence more on crop performance across different farming conditions such as soil types, and climatic conditions.

The research presented in this paper makes use of data mining for identifying the key traits within the feature set of the data collected from Bidar district ,which can maximize the crop yield. ***[2].*** This makes it a perfect candidate for machine learning (ML) .Through ML it can be predicted for a particular crop which trait is most important along with other features.

Variations in rainfall, temperature, increase of CO2, effect negatively the crop yields [13].Temperature and rainfall changes are expected to negatively impact on agricultural activities. Crop plants are sensitive to excessive rainfall, temperature and soil factors since they were selected for high yield, and not for tolerating climatic stress[16].

The rise in temperature beyond a threshold level during the cultivation is sufficient to cause permanent damage to plant growth and development [18]. The Intergovernmental Panel on Climate Change (IPCC) projected rise of the temperature by 3–4° by 2050 [19, 20]. High temperature conditions due industries and transport vehicles affect the percentage of seed germination, photosynthetic efficiency, flowering times [16, 21].Excess temperature contributed about 40% to overall yield loss of wheat [22], 1.0–1.7% yield loss per day in maize for every rise in temperature above 30°C [23].
In several regions in the world, some agricultural fields have undergone human-induced soil degradation resulting in poor yield production per unit area. Around 40% of agricultural lands are affected by human induced land degradation.Overuse of fertilizers and chemicals without adhering to standard protocol leads to a decline of soil health, land degradation [24].

Soil pH and soil organic matter (SOM) affect the bioavailability of micronutrients (Stevenson, 1986; Tate, 1987). Many soil factors such as pH, SOM, temperature, and moisture affect the availability of micronutrients to crop plants. The effects of these factors vary

considerably from one micronutrient to another as well as in their degree of effectiveness to plant uptake.

The relationships associated with each of the soil factors are complicated. A good example of this is the negative correlation between Mo and Mn. The availability of both Mo and Mn is so strongly affected by soil pH that the other factors are of limited value.***[6].***While Mn in plants decreases extensively with increasing soil pH, Mo increases, and deficiencies of both Mn and Mo are not expected or do not usually occur in the same soil. Manganese deficiency is often combined with excess Mo and vice versa . Soil pH influences solubility, concentration in soil solution, ionic form, and mobility of micronutrients in soil, and consequently acquisition of these elements by plants (Fageria, Baligar and Edwards, 1990; Fageria, Baligar, and Jones, 1997). As a rule, the availability of B, Cu, Fe, Mn, and Zn generally decreases, and Mo increases as soil pH increases. Increasing soil pH favors adsorption of B. Availability and uptake of B decrease dramatically at pH >6.0

Zinc solubility is highly soil pH dependent and decreases 100-times for each unit increase in pH, and uptake by plants decreases as a consequence. The principal ionic Mn species in soil solution is Mn2+, and concentrations decrease100-fold for each unit increase in soil pH. In extremely acidic soil, Mn2+ solubility can be sufficiently high to induce toxicity problems in sensitive crops. Considerable soil Cu is specifically adsorbed as pH increases. For example, increasing the pH from 4 to 7 increased Cu adsorption (Cavallaro and McBride, 1984), and Cu was adsorbed on inorganic soil components and occluded by soil hydroxide and oxides (Martens andWestermann, 1991). Increasing soil pH decreases B availability by increasing B adsorption onto clay The highest availability of B was at pH 5.5–7.5, and the availability decreased below or above this pH range. Soil pH affects solubility, adsorption, desorption, oxidation of Mn, and reduction of Mn oxides in soil. As the pH decreases, Mn is mobilized from various fractions and increases Mn soil solution concentrations and availability. Soil pH is more important than any other single property for controlling Zn mobility in soils (Anderson and Christensen, 1988). Increasing soil pH generally decreased Zn availability to plants (Saeed and Fox, 1977), and such decreases were usually due to higher adsorption of Zn. As soil pH increases above pH 5.5, Zn is adsorbed on hydrous oxides of Al, Fe, and Mn (Moraghan and ascagni,1991). Low soil pH significantly affects crop growth and therefore decreases yield. In maize for instance, soil acidity causes yield loss of up to 69% ***[25]***.Therefore there is some form of correlation exist among the features of soil and crop yield. Through this research an attempt is being made to discover such features which can maximize the yield.

Related Work

*D. Diepeveen and L. Armstrong**[1]**,* They used data mining tools for identifying key traits within a location or agzone to better identify various seed variety performance and thereby enable the farmer to make better cropping decisions. This seed variety comparison information included variety trials, results for 8 nominated varieties since 1975 for 574 different trial locations. This information also included site specific information and trial management metadata. Traits included in this paper are: grain yield (GY); 100day develop score (DS); harvest height score (HH); straw strength (SS); harvested grain weight (XGWT) harvested grain protein (XPRO); harvest grain over 2mm sieve (XS20) and 2.5mm sieve (XS25). Through this analysis they have tried to identify relationships between traits and varieties specific for a location or region that the farmer can assess. In some cases, the farmer may be able to change his/her agriculture practice to enable greater yield performance.

Chongyuan Zhang et.al.[3] conducted a study to evaluate phenomics technologies for monitoring performance traits (e.g., seed yield, days to 50% flowering, and days to physiological maturity) and predict the yield of chickpea and pea in 03 growing seasons and 03 environments (or locations). Crucial correlations ($P < 0.05$) between the image features derived from multispectral UAV based imagery and the yields of chickpea ($r < 0.93$) and pea ($r < 0.85$) were observed at the early growth, flowering, and pod/seed development stages, with some exceptions. During seed yield prediction with the combined features dataset using LASSO regression, R^2 values up to 0.91 and 0.80 (model testing) were achieved for chickpea and pea, respectively. The image-based features were identified by the LASSO regression models as the yield predictors for chickpea (1-7 features) and pea (3–20 features). The results indicated that phenomics technologies can be used for collecting data and evaluating pulse crop performance in multiple agri plots, seasons and environments.This can save time & money for plant breeders.

G. Mariammal et.al**.*[4]*** presented **a** novel feature selection(FS) approach called modified recursive feature elimination (MRFE) to select appropriate features from a data set for crop prediction. The proposed MRFE technique selects and ranks salient features using a ranking method. The experimental results show that the MRFE method selects the most accurate features, while the bagging technique helps accurately predict a suitable crop. The performance of proposed MRFE technique is evaluated by various metrics such as accuracy (ACC), precision, recall, specificity, F1 score, area under the curve, mean absolute error, and log loss. From the performance analysis, it is justified that the MRFE technique performs well with 95% ACC than other FS methods.

Abde Sherefu and Israel Zewide***[5]*** concluded that all nutrient elements focused in their study (N, P, K, S, Ca, Mg, Fe, and Zn) influence crop quality. low N will lead to reduced amount of proteins where as low K will lead to reduced amount of proteins due to reduced activation of enzymes that metabolize carbohydrates for synthesis of amino acids and proteins. Too much NH4-N will suppress uptake of Ca and its functions. On the other hand, low levels of Mg and K will lead to reduced distribution of carbohydrates. It should be noted that nutrients do not work in isolation; therefore balanced nutrition is needed to optimize crop quality. Micronutrient deficiency is a severe problem in soil and plants globally (Imtiaz et al., 2010) while appropriate quantity of micronutrients is necessary for better growth, better flowering, higher fruit set, higher

yield, quality and post-harvest life of horticultural products while its deficiency leads in lowering the productivity.

FAGERIA *et al.**[6]* in their research article presented that deficiencies of micronutrient in crops has increased remarkably in recent years due to intensive cropping, loss of top soil by erosion, losses of micronutrients through leaching, liming of acid soils, decreased usage of natural manure compared with chemical fertilizers. Micronutrient deficiency problems are also aggravated by a high demand of modern crop cultivars. Increase in crop yields due to application of micronutrients have been reported in many countriues. Factors such as pH, redox potential, biological activity, SOM, cation-exchange capacity, and clay contents are important in determining the availability of micronutrients in soils.

Tandzi Ngoune Liliane et.al.*[7]* deliberated that Climate smart agriculture sustainably increased crop yields. The development of integrated soil-crop management system with diseases and pests' management for existing crop varieties should be the target. The goal should be to increase the usage of new improved and adapted high-yielding varieties. The application of genetically engineered seed seems to be a viable option for development of high-yielding crops.

Geraldin B. Dela Cruz, et.al.*[8]* , in this study, a data mining method based on PCA-GA is presented to characterize agricultural crops. Specifically it draws improvements to classification problems by using Principal Components Analysis (PCA) as a preprocessing method and a modified Genetic Algorithm (GA) as the function optimizer. The GA performs the optimization process, selecting the most suited set of features that determines the class of a crop it belongs to. The fitness function in GA is studied and modified accordingly using efficient distance measures. The experimental results show improved classification rates.

Upendra M. Sainju*[10]*in their research they presented the advantages and disadvantages of Nitrogen fertilization. It is one of the most commonly used practice to increase crop yields throughout the world because of abundant availability of N fertilizers and their effectiveness to increase yields compared with other organic fertilizers, such as manure and compost. However excessive application of N fertilizers in the last few years has resulted in soil acidification. N leaching in groundwater, and emissions of nitrous oxide (N2O), a potent greenhouse gas that contributes to global warming. Crop yields have declined in places where soil acidification is high due to unavailability of major nutrients and basic cations and toxic effect of acidic cations.. To reduce excessive N fertilization, composited soil sample to a depth of 60 cm should be conducted for NO3-N test prior to crop planting and N fertilization rate be adjusted by deducting soil NO3-N content from the desirable N rate.

VelidePhani Kumar and Lakshmi Velide[12],they applied data mining in selection of suitable rice variety to Warangal region by obtaining the data of 20 years. The eight rice varieties recommended by rice research station, focussing on six traits were considered for the present study. Results show that Ramappa, JGL 384, Swarna, Samba masuri are high yielding varieties in Warangal region. Kavya, MTU 1010 and WGL14 have high harvested grain weight whereas Sureka, Samba Masuri and Ramappa have least. Kavya and WGL14 showed high straw strength, panicle length and carbohydrate content.

Nitin N. Patil et.al.[14], described various approaches presented by different researchers for agriculture data analysis. They used Naive Bayesian classification technique to recommend the crops and fertilizers. The proposed technique uses five weather and soil related parameters to

obtain reliable crop recommendations. In this approach, the results are promising and useful for crop and fertilizer recommendation which will help the farmers according to crop fields.

ANITHA ARUMUGAM[15],this research aims to develop a predictive model that provides a cultivation plan for farmers to get high yield of paddy crops using data mining techniques. The data set used in this research for mining process is real data collected from farmers cultivating paddy along the Thamirabarani river basin.K-means clustering and various decision tree classifiers are applied to meteorological and agronomic data for the paddy crop. The performance of various classifiers is validated and compared. Based on experimentation and evaluation, it has been concluded that the random forest classifier outperforms the other classification methods. The outcome of this research is the identification of different combination of traits for achieving high yield in paddy crop.16 attributes were selected for the study. It has been concluded that the random forest classifier has high predictive power with an accuracy rate of 97.5%, ROC = 0.99, precision = 0.97, and recall = 0.98.
Veenadhari Suraparaju et.al. [17], here an attempt has been made to study the influence of climatic parameters on soybean productivity using decision tree induction technique. The findings of Decision tree were framed into different rules for better understanding by the end users. They suggested, there a correlation is present between climatic factors and soybean crop productivity. These variables influence the soybean crop. This was confirmed from the rule accuracy and Bayesian classification. They presented that the productivity of soybean crop was mostly influenced by Relative humidity followed by temperature and rainfall.

II.Methodology

Through this research work, an attempt is being made to identify the important feature which is required for high Soyabean crop yield along with other features.

A . Input

Most of the research papers that were studied have considered some climatic parameters like temperature, humidity, rainfall. Some agronomical parameters like soil nutrient contents like N, P, K, Zn, B,S and pesticides etc. For our study Dataset on Rainfall, Temperature, pH and the corresponding crop name has been collected from Krishi Vigyan Kendra over the past 04 year.

B. Preprocessing

The data which is collected contains some incomplete, redundant, inconsistent data. Therefore in this step such redundant, incomplete data were filtered. Outliers were removed. Dataset should be balanced and skewness should be checked.

C. Identifying key Features

Dividing the entire dataset into Training data and testing data in the ratio of 70:30.Identifying the important feature among features which is crucial for high yield of Soyabean. Then testing the model.

D. Output

Based on Data mining algorithms, identifying crucial Crop Performance attributes among all the feature set for improving yield.

An overview of identifying key crop performance traits:

1. Input Data:Data collected from KVK , over a period of 04 years, comprising of the attributes Rainfall, Temperature, pH and Name of crop .
2. Preprocessing Data:Removing outliers, incomplete data as shown in Fig 1 a & b

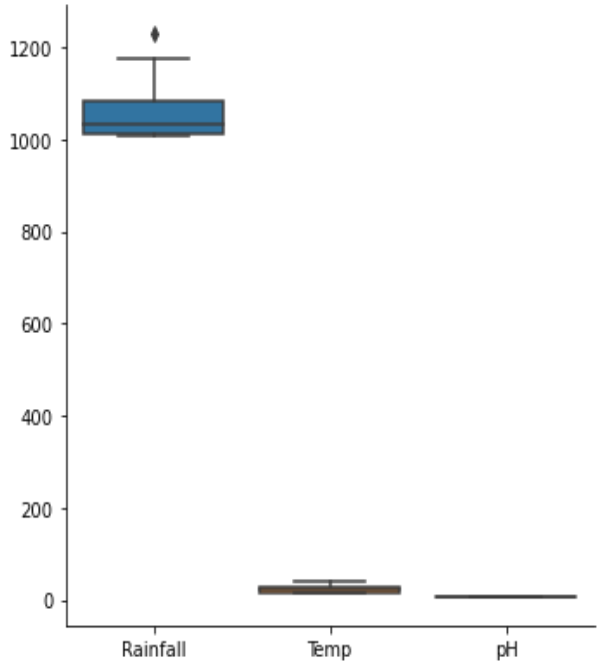

Fig 1(a) With outliers

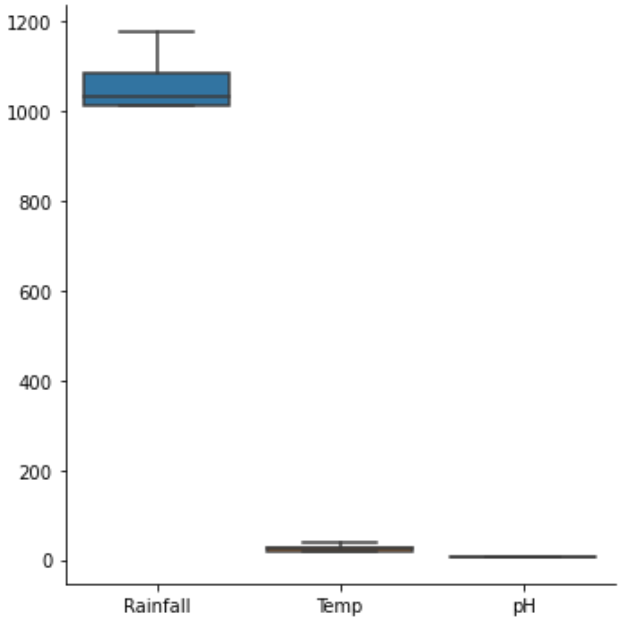

Fig 1(b)Without Outliers

Alt Text: During Data Preprocessing presence of outliers were checked and shown in fig 1.(a) and such outliers were removed and it can be seen in fig 1(b).In "Rainfall" attribute there were outliers which were removed during preprocessing.

3 Dataset balanced i.e.equal number of crops namely "Soyabean" and "other crop" as shown in Fig2.

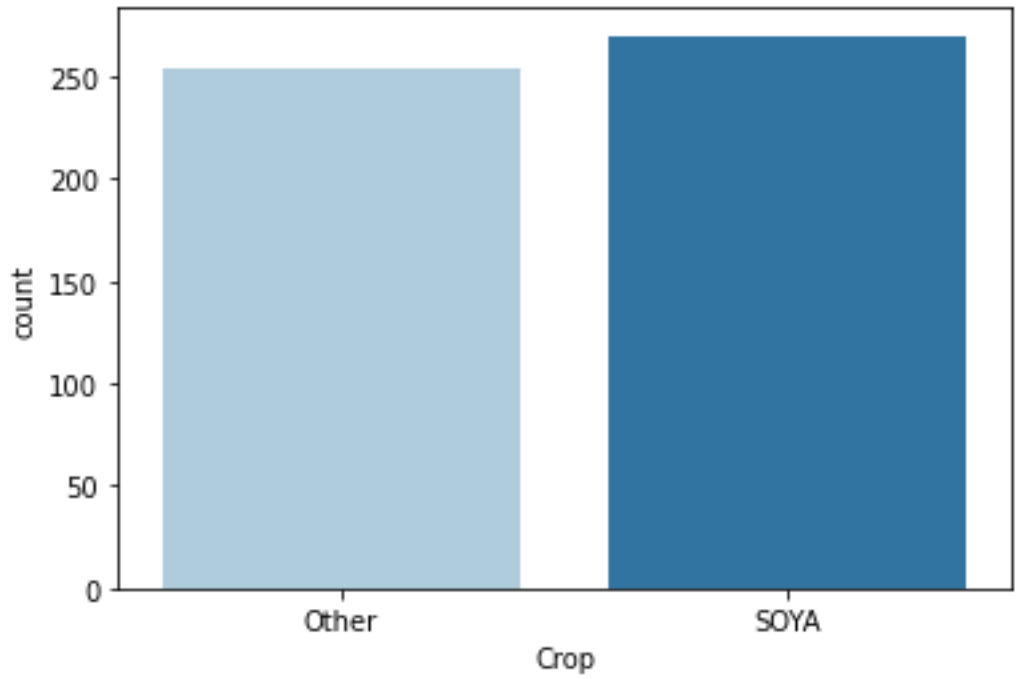

Fig 2.
Alt text: In the dataset , number of tuples with the attribute value 'Soya' and 'Other' were checked and is shown.

4. Dataset was checked for skewness.
5. Spliting the data into training and testing data using "train_test_split()"
6. Identification of the attribute which is very crucial for the soyabean crop as shown in Fig3..
7. Using GridSearchCV ,optimal values for the parameters of the Random Forest classifiers were calculated.
8. The model was tested using learning_curve() as shown in Fig 4..
9. The accuracy result obtained is 98.0%.

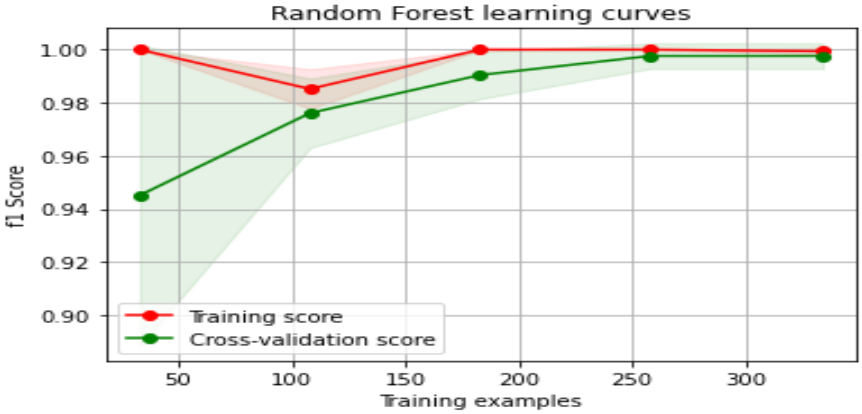

Fig 4.

Alt Text:The model was tested and it can be seen that after 250 sample both the curves are almost coinciding

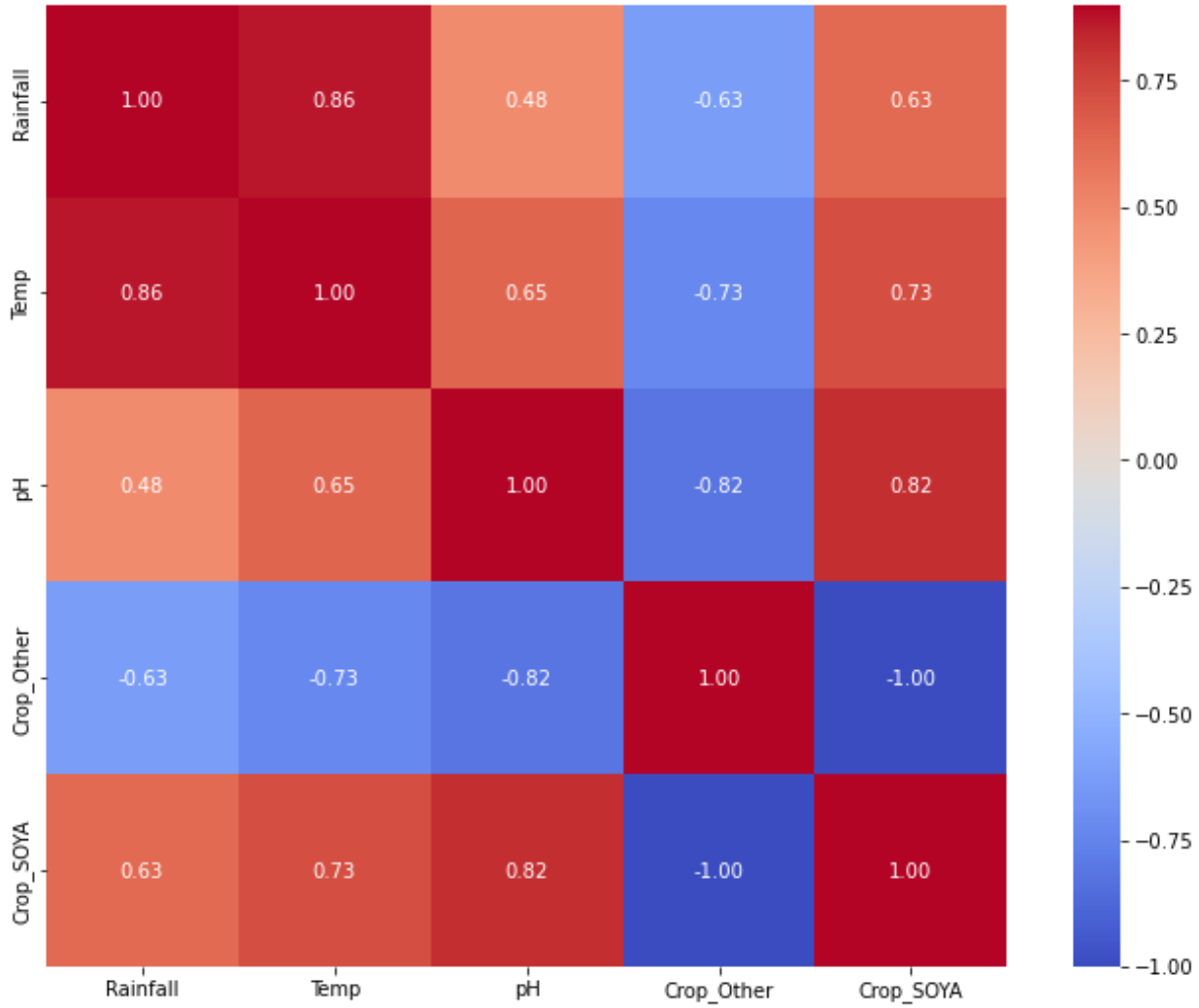

Fig 3.

Alt Text: From this diagram it can be viewed that pH is highly influencing Soya crop with 0.82 in comparison with Temp and Rainfall

III. Conclusion

In this paper, we presented an approach for identifying Key Crop Performance trait for improving soyabean yield of Bidar district.Clearly from the result it can be said that pH with a value of 0.82 is the crucial attribute as compared to rainfall (0.63) and temperature(0.73) for having high yield of Soyabean.It can be observed that pH is the most influencing feature as compared to rainfall and temperature.The model was tested and the result obtained was 98%.In

the proposed model we have used GridSearchCV for calculating the optimal values of the parameters for Random Forest classifier.

In general, the application of inappropriate practices such as untimely planting, incorrect plant spacing, poor sowing depth, delay in spraying of pesticide and weedicide, inappropriate use of fertilizers, and use of low yielding varieties, will significantly reduce crop yields.
Weather forecasting , crop yield prediction and specifying Key Crop Performance traits well in advance will provide a guideline to farmers in the process of crop cultivation.

IV. Future Work

There are still many challenges in this research area. To meet such challenges a more careful study by considering the soil attributes like OC,P,K Zn,B,S along with rainfall ,Temperature and pH should be carried out.And among these, identifying the most crucial attribute for the high yield of soyabean will be the objective of the research.But the bottleneck is the non availability of such dataset comprising of all the attributes mentioned.

V. Acknowledgement

We are very much thankful to the Krishi Vigyan Kendra for their unconditional support in carrying out this research work.

References

[1] D. Diepeveen and L. Armstrong "Identifying key crop performance traits using data mining" IAALD AFITA WCCA2008,World Conference on Agricultural Information and IT.

[2] Thirtle C, Irz X, Lin L, Mckenzie- Hill V, Wiggins S. Relationship between changes in agricultural productivity and the incidence of poverty in developing countries. In: DFID Report No. 7946. 2001

[3]Chongyuan Zhang, Rebecca J. McGee, George J. Vandemark and Sindhuja Sankaran ,"Crop Performance Evaluation of Chickpea and Dry Pea Breeding Lines Across Seasons and Locations Using Phenomics Data" ORIGINAL RESEARCH published: 25 February 2021,doi: 10.3389/fpls.2021.640259

[4] G. Mariammal , A. Suruliandi , S. P. Raja , and E. Poongothai ," Prediction of Land Suitability for Crop Cultivation Based on Soil and Environmental Characteristics
Using Modified Recursive Feature Elimination Technique With Various Classifiers", IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS, Digital Object Identifier 10.1109/TCSS.2021.3074534

[5] Abde Sherefu and Israel Zewide," Review Paper on Effect of Micronutrients for Crop Production", Journal of Nutrition and Food Processing Isreal Zewide,
Auctores Publishing LLC – Volume 4(7)-063 www.auctoresonline.org ISSN: 2637-8914

[6] N. K. Fageria,1 V. C. Baligar,2 and R. B. Clark," **MICRONUTRIENTS IN CROP PRODUCTION**", *Advances in Agronomy, Volume 77,*Copyright 2002, Elsevier Science (USA).

[7] *Tandzi Ngoune Liliane and Mutengwa Shelton Charles,"*Factors Affecting Yield of Crops*"*
DOI: http://dx.doi.org/10.5772/intechopen.90672

[8] Geraldin B. Dela Cruz, Bobby D. Gerardo and Bartolome T.Tanguilig III," An Improved Data Mining Mechanism Based on PCA-GA for Agricultural Crops Characterization" , *International Journal of Computer and Communication Engineering, Vol. 3, No. 3, May 2014,* DOI: 10.7763/IJCCE.2014.V3.324.

[9] A. Suruliandi, G. Mariammal & S.P. Raja, "Crop prediction based on soil and environmental characteristics using feature selection techniques**",** (Taylor & Francis) MATHEMATICAL AND COMPUTER MODELLING OF DYNAMICAL SYSTEMS 2021, VOL. 27, NO. 1, 117–140 https://doi.org/10.1080/13873954.2021.1882505

[10] *Upendra M. Sainju, Rajan Ghimire and Gautam P. Pradhan,"* Nitrogen Fertilization I: Impact on Crop, Soil, and Environment*", DOI: http://dx.doi.org/10.5772/intechopen.86028*

[11] Pierre Casadebaig, Bangyou Zheng, Scott Chapman, Neil Huth, Robert Faivre, Karine Chenu," Assessment of the Potential Impacts of Wheat Plant Traits across Environments by Combining Crop Modeling and Global Sensitivity Analysis", PLOS ONE | DOI:10.1371/journal.pone.0146385 January 22, 2016"

[12] VelidePhani Kumar and Lakshmi Velide ,"DATAMINING-A TOOL TO IDENTIFY CROP PERFORMANCE TRAITS FOR RICE VARIETIES OF WARANGAL REGION", *International Journal of Applied Biology and Pharmaceutical Technology Page: 177,Available online at www.ijabpt.com*

[13] Raza A, Razzaq A, Mehmood SS, Zou X, Zhang X, Lv Y, et al. "Impact of climate change on crop adaptation and strategies to tackle its outcome: A review". Plants. 2019;**8**(34):1-29. DOI: 10.3390/plants8020034

[14] **Nitin N. Patil, Mohmmad Ali M. Saiyyad," Machine Learning Technique for Crop recommendation in agriculture sector", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019,** *Retrieval Number: A1171109119/2019©BEIESP DOI: 10.35940/ijeat.A1171.109119 Journal Website: www.ijeat.org*

[15] ANITHA ARUMUGAM," A predictive modeling approach for improving paddy crop productivity using data mining techniques", Turkish Journal of Electrical Engineering and Computer Sciences: Vol. 25 No. 6, Article 28. https://doi.org/10.3906/elk-1612-361
Available at:https://journals.tubitak.gov.tr/elektrik/vol25/iss6/28

[16]Kumaraswamy S, Shetty PK." Critical abiotic factors affecting implementation of technological innovations in rice and wheat production: A review". Agricultural Reviews. 2016;**37**(4): 268-278. DOI: 10.18805/ag.v37i4.6457

[17] Veenadhari Suraparaju," Soybean Productivity Modelling using Decision Tree Algorithms", International Journal of Computer Applications (0975 – 8887) *Volume 27– No.7, August 2011*

[18] Kavi Kumar," Indian Agriculture and climate sensitivity", **Global Environmental Change**, *Vol. 11(2), pp: 147-154, 2001*

[19] IPCC. Climate change and biodiversity. In: IPCC Technical Paper V. 2002. pp. 1-86

[20] Meeh GA, Stocker TF, Collins WD." Climate change 2007: The physical science basis. In: Fourth Assessment Report of the Intergovernmental Panel on Climate Change". Cambridge, UK: University Press; 2007

[21]Antle JM, McGuckin T." Technological innovations, agricultural productivity and environmental quality". In: Carlson GA, Zilberman D, Miranowski J, editors. Agric. Environ. Resour. Econ. 1993. pp. 175-220

[22] Elbasyoni IS. "Performance and stability of commercial wheat cultivars under terminal heat stress". Agronomy. 2018;**8**:37

[23] Lobell DB, Banziger M, Magorokosho C, Vivek B. "Nonlinear heat effects on African maize as evidenced by historical yield trials" Nature Climate Change. 2011;**1**:42-45

[24] Shah F, Wu W. "Soil and crop management strategies to ensure higher crop productivity within sustainable environments. Sustainability". 2019;**11**(1485):1-19. DOI: 10.3390/ su11051485

[25] Tandzi NL, Mutengwa CS, Ngonkeu ELM, Gracen V. "Breeding maize for tolerance to acidic soils: A review. Agronomy". 2018;**8**(84):2-21. DOI: 10.3390/agronomy8060084