International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Supervised Machine Learning Techniques for Predicting Sugarcane Yield

Ramesh Medar, Vijay S. Rajpurohit

Abstract: ***Agriculture is the most important sector in the Indian economy and contributes 18% of Gross Domestic Product(GDP). India is the second largest producer of sugarcane crop and produces about 20% of the world's sugarcane. Sugarcane is cultivated in tropics and subtropic regions, on a wide range of soils from fertile well-drained mollisols to through heavy cracking vertisols, infertile acid oxisols, peaty histosols, to rocky andisols. Minimum moisture of 60cms, rich water supply and plenty of sunshine. In this paper, a novel approach to sugarcane yield forecasting in Karnataka, India region using Long Term Time Series(LTTS), weather-and-soil attributes, Normalized Vegetation Index(NDVI) and Supervised Machine Learning(SML) algorithms have been proposed. Sugarcane cultivation life cycle(SCLC) in the Karnataka region is about 12 months, with plantation beginning at three different seasons in weather condition. Our approach has been verified using historical dataset and results have shown that our approach has successfully modeled crop prediction. The application of the Custom-Kernel gives us a considerable boost in accuracy with SVM-Kernel Multiple giving 86.31% of accuracy, SVM-RBF kernel in second with an accuracy of 83.40%, GPR producing an accuracy score of 81.75%, Lasso giving an accuracy score of 26.81% and Kernel Ridge-RBF with the least accuracy score of 21.46% for final yield prediction.***

Keywords: ***Agriculture, Machine learning, Custom-kernel, Crop prediction.***

I. INTRODUCTION

Long Term Time Series (LTTS) forecasting has been a useful tool for governments, planning commissions and decision-makers in various applications such as solar energy, wind power energy, economic forecasting, and the agriculture sector. Historically LTTS has been applied at the regional and national level for planning, import and export decision making, and policy decisions [1]-[2].

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author
Ramesh Medar*, Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, India.. Email: rameshcs.git@gmail.com
Dr. Vijay S. Rajpurohit, Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, India. Email: vijaysr2k@yahoo.com

Traditionally, application of LTTS in the agriculture sector for yield prediction/crop forecasting is limited to empirical methods using ground-based observations and productions reports gathered by various organizations from different sources: meteorological data, agro-meteorological(yield), soil(water holding capacity), and remotely sensed agricultural statistics.

As Crop production rate depends on the geography of a region(e.g. hill area, river ground, depth region), weather condition(e.g. temperature, cloud, rainfall, humidity), soil type(e.g. sandy, salty, clay, peaty, saline soil), soil composition(e.g. PH value, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron) and harvesting methods, various combinations of subsets of these influencing parameters have been used by different prediction models for crop yield prediction using ground-based observations [3]-[7]. Datasets are whole and sole input to the supervised machine learning algorithms. Datasets consist of 'n' number of features and 'n' number of dimensions. The fine line between classes of data in the dataset becomes a major issue as ambiguity in the differentiation of classes increases as the data gets dense in a multi-class dataset. As long as the data is linear, it does not need any special manipulations to bifurcate between classes or types within itself. The second problem arises when the data is non-linear, the data is ambiguous to understand at first glance and needs some operations being made to transform the dimensions of the data which are known as kernel functions. Kernel functions are the ones which transform non-linear ambiguous data into understandable data. These kernels come in different forms and cover a broad spectrum of types of data. Kernel functions make the dataset as simple as possible, and thus kernel functions play a vital role in data analysis as well as in machine learning. Kernel functions are also known as the kernel trick, this describes the simple nature of the application of the kernel functions. Kernel functions come in a wide variety depending upon the type of data and the machine learning algorithm being used for prediction. Support Vector Machine (SVM) is the most famous machine learning algorithm which makes use of kernel functions extensively. We come across different kernel functions such as Polynomial kernel, Gaussian kernel, Gaussian Radial Basis kernel (RBF), Laplace RBF kernel, etc. This list of kernel functions is used under specific conditions of the dataset. The decision of making use of a specific kernel depends on how the initial plot of the dataset looks like.

Kernel functions can also be designed and tailor-made to obtain optimum score accuracy and provide better performance. These kernel functions are called custom-kernels. Thus, custom-kernels are the best way to tackle the optimization problem faced when none of the pre-defined kernels are effective.

This paper makes use of custom-kernels which are specifically designed to integrate with the Support Vector Machine (SVM) algorithm to give us desired results in the prediction. We will be making a comparative analysis of algorithms such as SVR-RBF, GPR, Kernel Ridge and Lasso Algorithms concerning accuracy score obtained after applying our custom-kernel.

II. LITERATURE SURVEY

Crop yield forecasting models could be categorized based on attribute measurement methods like ground-based observed data such as meteorological, agro-meteorological and soil data, remotely sensed data like various SVIs, weather data, soil data derived from SVIs, and a combination
of both ground-based
and remote sensed.

Another way to categorize yield forecasting models could be as classical empirical models and machine learning models [1]-[7]. Reference [8] shows different SVM kernels used for univariate and multivariate time series analysis. Each of these kernels model different assumptions on the process that generates the time series. Which kernel is optimal for a given learning task is still an unsolved problem. The experiments showed that the RBF kernel performs very well on different types of time series and learning tasks. Reference [9] shows that using a mixture of kernels can result in having both good interpolation and extrapolation abilities. The performance of this method is illustrated with an artificial as well as an industrial data set. It is shown that where the RBF kernels fail to extrapolate and a very high degree Polynomial kernel is needed to interpolate well, the mixture of the two kernels can do both. Reference [10] presented a comparative study on the performance of different SVM's kernels for classification of multi-temporal full-polarimetric L-band SAR data in the agricultural region. For classification, different SVM classifiers based on several well-known kernel functions(i.e. RBF, Linear, and polynomial) were applied to multi-temporal polarimetric features. The experimental evaluations demonstrated that the accuracies of RBF-based SVM classifier for various crop types were relatively better than the other two kernel functions. Reference [11] shows hybrid model of mixed kernels function (MKF) based support vector regression (SVR) model was developed to predict pure and impure CO2-oil minimum miscibility pressure (MMP) during a CO2-EOR process. In this MKF-SVR model, four factors (i.e. reservoir temperature, average critical temperature, the molecular weight of C5+ fraction of crude oil, and the ration of volatile components to intermediate components in crude oil) representing the most comprehensive and robust set were selected as the input variables while MMP was considered as the output variable. The mixed kernel function based support vector regression (SVR) model was successfully applied to predict the CO2-oil MMP value for both pure and impure CO2 gas. Different kernel functions affect the final performance of SVR significantly. Mixed kernel function, which combines the advantages of global radial basis function (RBF) and local (Polynomial) kernel functions, increases the applicability of SVR dramatically.

III. SUGARCANE YIELD PREDICTION MODEL

Dataset is a critical component of any ML algorithm and it needs to be understood and pre-processed before applying ML algorithms in any domain. Dataset used in this research comprises Weather and Soil Dataset(WSD), NDVI dataset and Sugarcane crop yield dataset. WSD is downloaded from "*https://www.meteoblue.com/en/weather/forecast/week/16.246N74.737E" [12]*, for the village Shirdhan, located at latitude and longitude of 16.2458^{o}N, 74.737^{o}E of Belagavi district, Karnataka(India). The dataset is multidimensional and application of multidimensional datasets require kernel functions to obtain a better prediction. The architecture diagram shown in Fig.1 gives us an overview of the process to be followed in achieving crop yield prediction. Proposed Sugarcane Crop Yield Forecasting Model(SCYFM) consists of three modules, which are by ML approaches. As shown in Fig.1, the first module known as Dataset Pre-processing Module(DPM) re-samples, scales and normalizes each attribute, select independent and important attributes, and divides into training and testing dataset. The second module is known as the Training and Testing
Module(TTM) trains SVR algorithm with built-in kernel functions first to make a rough estimate of the impact of it on the model and accuracy. This analysis will help us to formulate a Custom-Kernel function which will be specific to this dataset and will be modeled in phase 3. The third module is known as Prediction Module(PRM) forecasts sugarcane crop yield. All the three modules are implemented using Sci-Kit Learn package version 0.1.91 and Python 3.7.

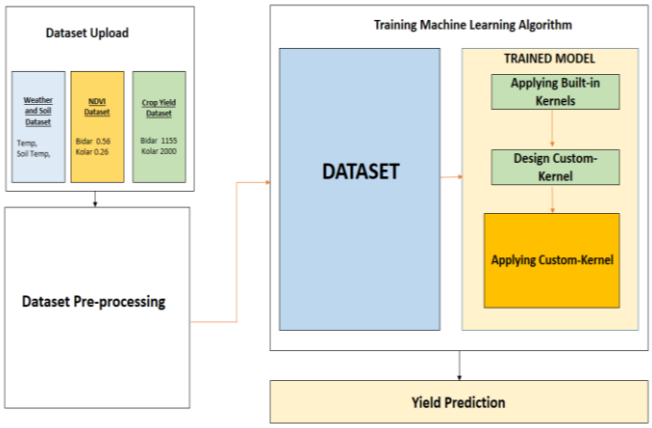

Fig.1. Architecture diagram of custom kernel crop prediction model

A. Dataset Pre-processing Module

Considering the three stages in the architecture of this research,
the dataset was collected from various districts around India in a 24 x 365 steps time series dataset. The dataset was re-sampled to 15 days aggregate. The dataset was further converted into 24 steps supervised ML dataset. We have about 4 years yield data per district, as supervised ML algorithms require as much data as possible, so we have considered all the growing districts in the state of Karnataka, Bihar, and Haryana. Phase wise NDVI aggregation is done according to planting timing in each district. The growing phases vary region wise. For Belagavi and Bijapur districts in Karnataka, sugarcane planting is done in January.

Dataset Distribution and Outliers: Using various graphs, we can understand the distribution of each attributes in dataset independently. Box and Whisker plot is shown in fig.2, indicates weather attributes are either skewed or have outliers. According to fig.2, attributes state 1, state 2, state 3 and yield have outliers. Outliers in the dataset could not be neglected, as they provide very important information about the nature of the overall condition. Histogram graph shown in fig.3, groups each attributes in the number of bins and provides several observations in each bin. The correlation matrix shown in fig.4 depicts independence in the correlation between the features. It is seen that state 1 has a positive relation with State 2 and State 3, and is a completely negative relation with yield. The density graph shown in fig.5, gives us an insight of a canvas of values where *n* number of data points are considered and these points give us a generic curve, in case of this particular dataset the density curve shows us that state 1, state 2, state 3 and yield have a single peak curve trend.

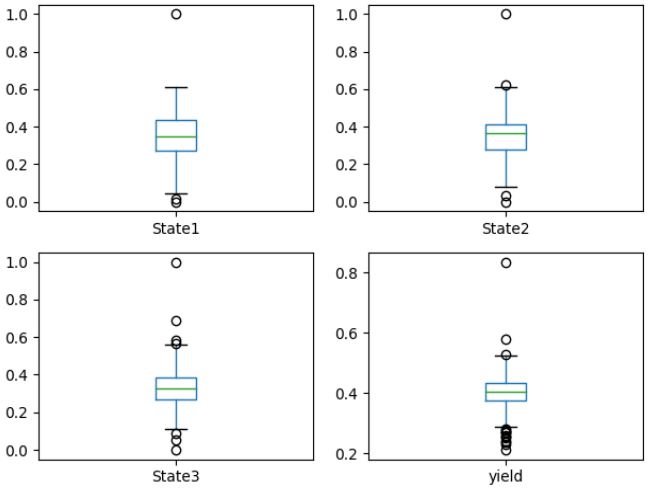

Fig.2. Box and Whisker Plot

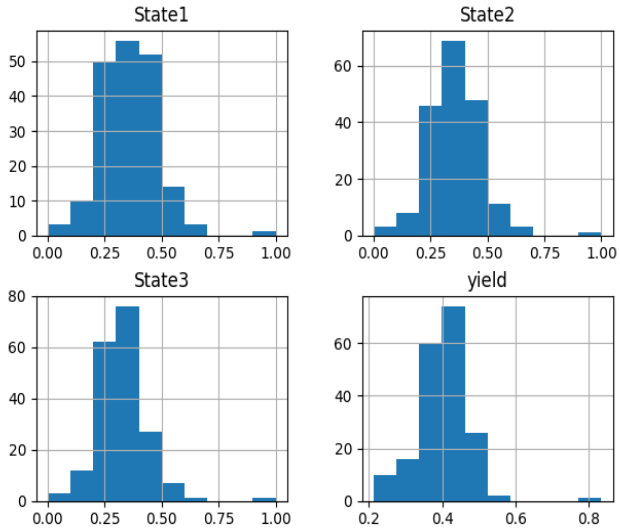

Fig.3. Histogram graph

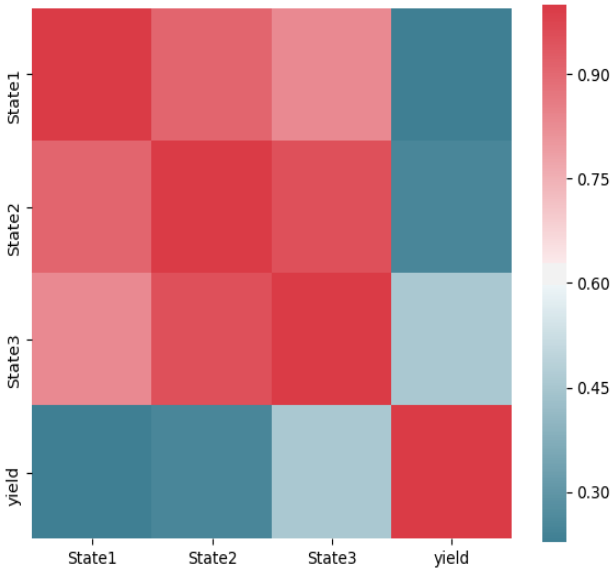

Fig.4. Correlation matrix

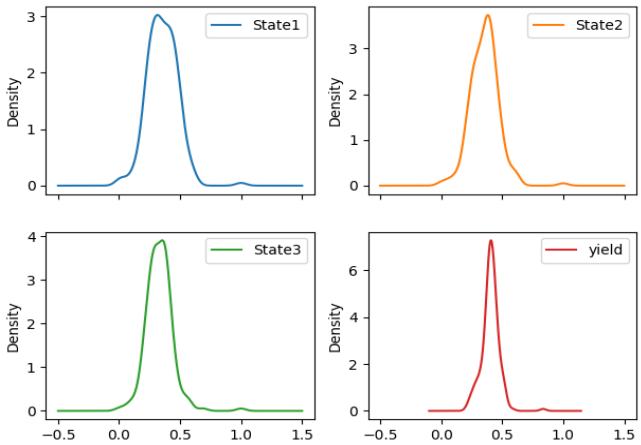

Fig.5. Density graph

B. Training and Testing Module

The training module according to the architecture diagram consists of the application of built-in kernels such as RBF-kernel, White-kernel, ExpSineSquared-kernel, and RationalQuadratic-kernel. The next phase in the training module consists of a formulation of data specific custom-kernel which is tailor-made to this particular dataset being used. Custom-kernel used in this paper is an additive custom-kernel where the summation of the built-in kernels is the final custom-kernel. Once the custom-kernel is ready, the selection of hyper-parameters is complete, which is explained in detail in the implementation part of the paper. Thus hyper-parameters have a considerable impact on the prediction model and cannot be ignored. We train the model with the custom-kernel once the selection of hyper-parameters is completed. The trained model is now ready for testing and prediction of yield.

IV. IMPLEMENTATION

The Custom-kernel is composed of several terms that are responsible for explaining different properties of the signal:

- a long term, a smooth rising trend is to be explained by an RBF kernel. The RBF kernel with a large length-scale enforces this component to be smooth. The specific length-scale and the amplitude are free hyper-parameters.
- a seasonal component, which is to be explained by the periodic ExpSineSquared kernel with a fixed periodicity of 15 days. The length-scale of this periodic component controlling its smoothness is a free parameter. To allow decaying away from exact periodicity, the product with an RBF kernel will be taken. The length-scale of this RBF component controls the decay time and is a further free parameter.
- smaller, medium-term irregularities are to be explained by a RationalQuadratic kernel component, whose length-scale and alpha parameter, which determines the diffuseness of the length-scales are to be determined.
- a noise term, consisting of an RBF kernel contribution, which will explain the correlated noise components such as local weather phenomena and a WhiteKernel contribution for the white noise. The relative amplitudes and the RBF's length scale are further free parameters.

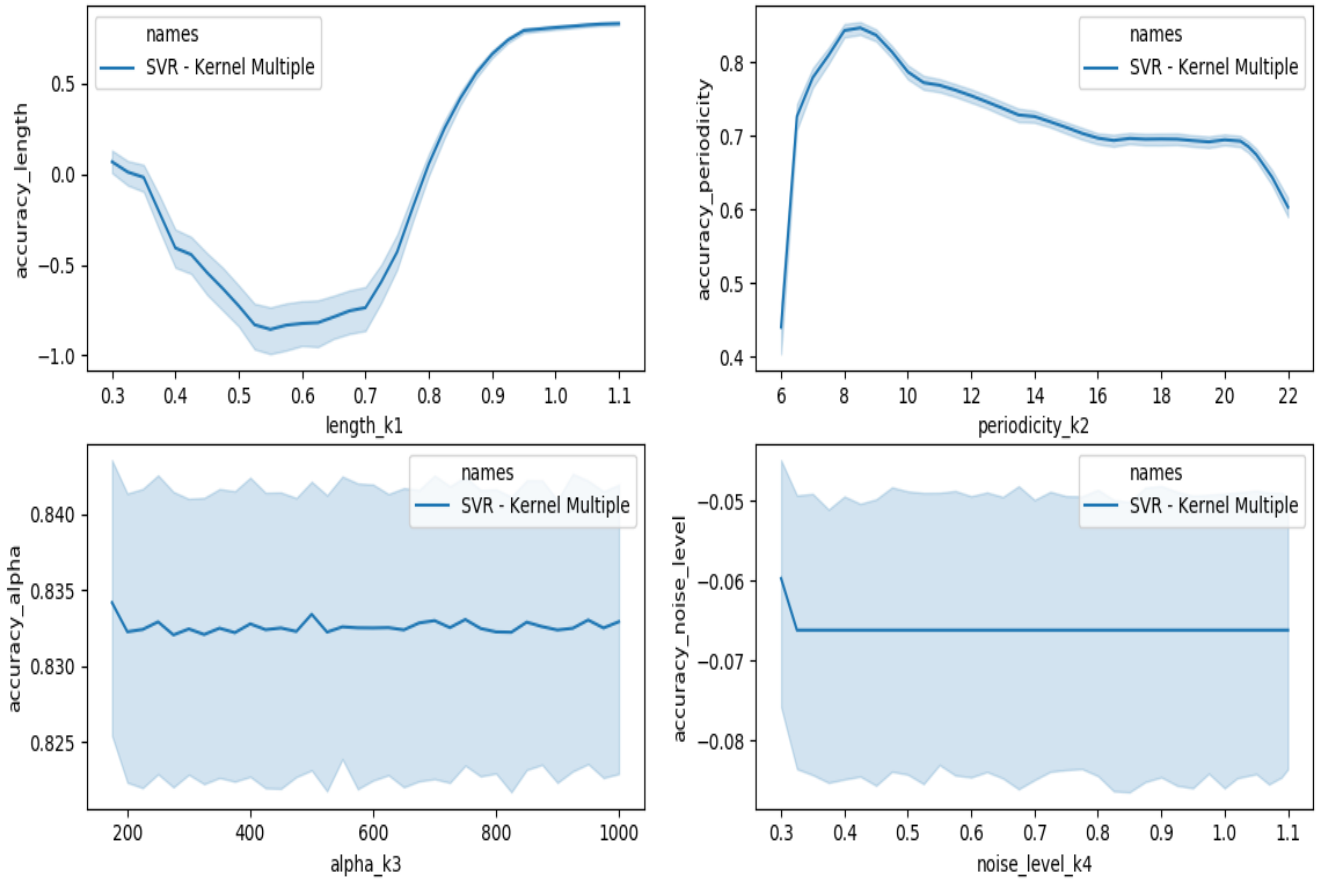

Fig.6. Graph displaying the values of various kernel parameters such as length scale, periodicity, alpha and noise level.

Fig.6 justifies the value selection for various Kernel Parameters such as ***length_scale, periodicity,*** and ***noise_level*** which have been used in the Custom-Kernel. The RBF kernel used in the above equation makes use of ***length_scale*** as a parameter. Fig.6 helps in justifying the value selection for ***length_scale***. We made use of values ranging from 0.3 to 1.1 and we observed that 0.3 to 0.7 was a negative trend and affected the performance of the Custom-Kernel. Values from 0.7 to 0.9 show a positive trend after which the curve stabilizes from 0.9 onwards, thus we have selected ***length_scale***=1.1.

The ExpSineSquared kernel makes use of ***length_scale*** and ***periodicity*** as parameters. As ***length_scale*** gives a positive trend of 0.9 onwards, we make use of ***length_scale=0.9.*** An experiment was conducted where the ***periodicity*** value was ranged from 6 to 22. The results of this experiment helped us selecting the value for ***periodicity***. Fig.6 shows that ***periodicity=9.0*** hits the peak point after which a negative trend begins in the graph thus making us choose ***periodicity =9.0***. Fig.6 explains the parameter selection for RationalQuadratic Kernel. The RationalQuadratic Kernel makes use of ***length_scale*** and ***alpha*** as parameters.

As ***length_scale*** gives a positive trend of 0.9 onwards, we make use of ***length_scale =10.0.*** An experiment was conducted where the ***alpha*** value was ranged from 200 to 1000. We have selected ***alpha=1000*** as the majority of the values selected gave similar results where the trend did not variate significantly. Fig.6 also explains about WhiteKernel and its parameter selection. The WhiteKernel makes use of ***noise_level*** as a parameter. An experiment was conducted where the ***noise_level*** value was ranged from 0.1 to 1.1. The results of this experiment helped us selecting the value for ***noise_level***. The experiment gives us a clear view that even when we change the value of ***noise_level*** from 0.3 to 1.1, it gives a stable trend with no change in the curve whatsoever, thus making us select the value of ***noise_level as 0.1.***

The custom-kernel used is:

k4=RBF(length_scale=1.1) + ExpSineSquared(length_scale=0.9,periodicity=9.0) + RationalQuadratic(length_scale=10.0,alpha=1000.0) + WhiteKernel(noise_level=0.12, noise_level_bounds=(1,np.inf))**

V. EVALUATION RESULTS AND ANALYSIS

Earlier researchers have worked on predicting sugarcane yield for small regions where weather conditions and sowing start times were the same. In this research, we have successfully modeled sugarcane yield prediction considering different sowing start period under different conditions in the India region. The implemented model has been evaluated by running experiments with various test and train sizes for SVM-Kernel Multiple algorithm and comparing with Lasso, GPR, SVM-RBF, and Kernel Ridge-RBF algorithms. The application of the Custom-Kernel gives considerable boost in accuracy with SVM-Kernel Multiple leading the accuracy score with **86.31%** of accuracy, SVM-RBF kernel in second with an accuracy of 83.40%, GPR producing an accuracy score of 81.75%, Lasso giving an accuracy score of 26.81% and Kernel Ridge-RBF with the least accuracy score of 21.46%. The difference in the score when Custom-kernel is applied is evident and the graphs depict the same. The below graph represents the Box and Whisker Plot and the Line Graph for accuracies for different sample size into consideration.

Table-I: Performance of various algorithms

Sample Size	SVM-RBF	SVM-Kernel Multiple	GPR	Kernel Ridge -RBF	Lasso
100	0.7551	0.8400	0.7770	0.2166	0.2597
110	0.8166	0.8544	0.7933	0.1703	0.2356
120	0.8340	0.8631	0.8128	0.1968	0.2703
130	0.8091	0.8349	0.7839	0.2052	0.2497
140	0.7873	0.8354	0.7598	0.2032	0.2333
150	0.7551	0.7956	0.7770	0.2166	0.2597

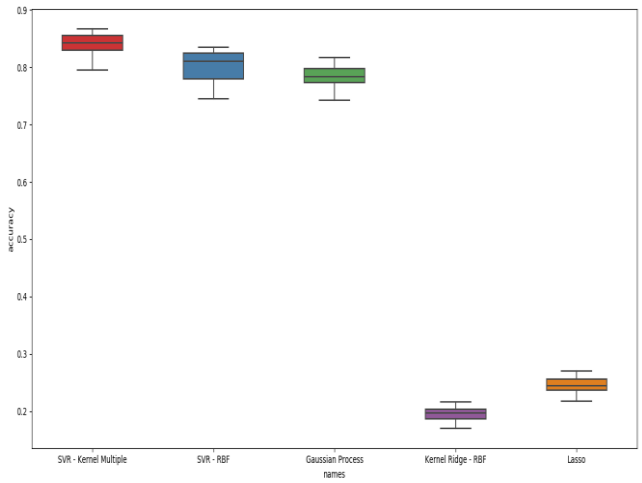

Fig. 7. Box and Whisker Plot.

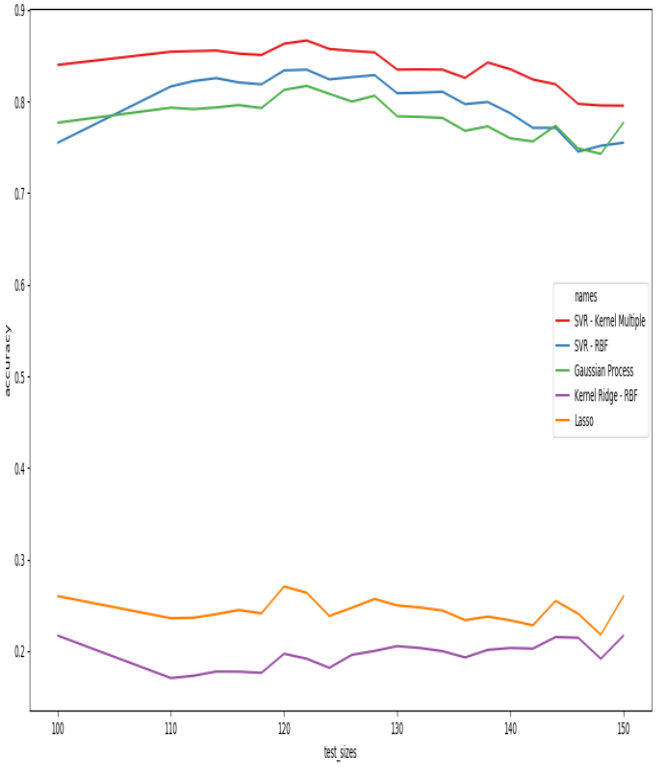

Fig. 8. Line Graph.

VI. CONCLUSION

Earlier researchers have worked on predicting sugarcane yield for small regions where weather conditions and sowing start times were the same. In this research, we have successfully modeled sugarcane yield prediction considering different sowing start period under different conditions in the India region. In this paper, we have successfully modeled sugarcane yield forecasting using weather and soil attributes and NDVI considering 12 months growth period. The application of the Custom-Kernel gives us a considerable boost in accuracy with SVM-Kernel Multiple leading the accuracy score with **86.31%** of accuracy, SVM-RBF kernel in second with an accuracy of 83.40%, GPR producing an accuracy score of 81.75%, Lasso giving an accuracy score of 26.81% and Kernel Ridge-RBF with the least accuracy score of 21.46% for final yield prediction.

REFERENCES

1. Sorjamaa A, Hao J, Reyhani N, Ji Y, Lendasse A, "Methodology for long-term prediction of time series, Neurocomputing", 2007, pp. 2861-2869. doi: 10.1016/j.neucom.2006.06.015.
2. Aghighi H, Azadbakht M, Ashourloo D, Shahrabi H, Radiom S, "Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018, pp. 4563-4577.
3. Jayawardhana W, Chathurange V, "Extraction of Agricultural Phenological Parameters of Sri Lanka Using MODIS, NDVI Time Series Data", Procedia Food Science, 2016, pp. 235-241. doi:10.1016/j.profoo.2016.02.027.
4. Lai Y, Pringle M, Kopittke P, Menzies N, Orton T, Dang Y, " An empirical model for the prediction of wheat yield using time-integrated Landsat NDVI. International Journal of Applied Earth Observation and Geoinformation", 2018, pp. 99-108. doi:10.1016/j.jag.2018.07.013.
5. Saeed, Umer, Dempewolf, Jan, Becker-Reshef, Inbal, Khan, Ahmad, Ahmad, Ashfaq, et al, "Forecasting wheat yield from weather data and MODIS NDVI using Random Forests for Punjab province, Pakistan", International Journal of Remote Sensing, 2017, pp. 4831-4854. doi:10.1080/01431161.2017.1323282.
6. Mkhabela M, Mkhabela M, Mashinini N, "Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR", Agricultural and Forest Meteorology, 2005, pp. 1-9.
7. doi:10.1016/j.agrformet.2004.12.006.
8. Prasad, Anup, Singh, Tare, Kafatos, Menas, "Use of vegetation index and meteorological parameters for the prediction of crop yield in India", International Journal of Remote Sensing, 2007, pp. 5207-5235. doi:10.1080/01431160601105843.
9. Rüping, Stefan. SVM kernels for time series analysis. Technical Report. SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, No. 2001,43, SFB 475, Universität Dortmund, Dortmund.
10. Smits G, Jordaan E, "Improved SVM regression using mixtures of kernels", *International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, Honolulu, HI, USA, 2002, pp. 2785-2790. doi: 10.1109/IJCNN.2002.1007589.
11. Yekkehkhany B, Safari A, Homayouni S, Hasanlou M, "A Comparison Study Of Different Kernel Functions For SVM-Based Classification Of Multi-Temporal Polarimetry Sar Data", 1st ISPRS International Conference on Geospatial Information Research, 2014, pp. 281-285. doi:10.5194/isprsarchives-XL-2-W3-281-2014.
11. Zhong Z, Carr T, "Application of mixed kernels function (MKF) based support vector regression model (SVR) for CO2 – Reservoir oil minimum miscibility pressure prediction, Fuel", 2016, pp. 590-603. doi:10.1016/j.fuel.2016.07.030.
12. https://www.meteoblue.com/en/weather/forecast/week/16.246N74.737E.

AUTHORS PROFILE

Ramesh Medar, Assistant Professor, Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, Karnataka, India. Pursuing Ph.D. in the domain Machine learning, Data mining. Completed M.Tech. in the year 2011 from KLS Gogte Institute of Technology. Total teaching experience of 12.5 years, 5 years of research experience. Published a few papers in national, international journals. Presented papers in national, international conferences. Member of LMISTE, CSTA.

Dr. Vijay S Rajpurohit, working as Professor in the Department of Computer Science and Engg at Gogte Institute of Technology, Belagavi, Karnataka, India. Completed B.E. in Computer Science and Engg. from Karnataka University Dharwad, M.Tech. at N.I.T.K Surathkal and Ph.D. from Manipal University, Manipal in 2009. His research areas include Image Processing, Cloud Computing, and Data Analytics. He has published a good number of papers in Journals, International and National conferences. Dr. V. S. Rajpurohit is the reviewer for a few international journals and conferences. He is the associate editor for two international journals and Senior Member of the International Association of CS and IT. He is also the life member of SSI, ISC and ISTE associations.