A Disparate Univariate Regression Model for Sugarcane Crop Forecasting to Meet the Global Demand

Dr.M.Janaki
Associate Professor, Department of Computer Science, Dr. UmayalRamanathan College for Women, Karaikudi, India.
Email:mjanaki81@gmail.com

ABSTRACT: The art of predicting the yield of crop before harvesting is termed as crop forecasting. Crop forecasting helps in economic planning and also ensures global food security. Crop forecasting depends on various factors such as crop management, plant physiology, meteorology, soil science and statistical models etc. The dataset used for this study contains one independent variable hectares of land in which sugarcane is cultivated and one dependent variable yield in tones. This paper proposed a Disparate Univariate Regression Model (DURM) which uses simple linear, polynomial, Decision Tree, Random Forest and Support Vector Regression to predict the yield of sugarcane crop with respect to the hectares of land used for sugarcane cultivation. The Regression Models were built and the Equation was framed with x and y intercepts values and the future yield of sugarcane was predicted. The correlation between the hectares of the land cultivated and sugarcane produced in tones was calculated using the correlation metrics and the performance of the various regression models was tested with the error metrics such as (RMSE, MAE, MSE, MAPE) and the results were compared. The extracted knowledge through thisstudy helps the farmers to take decision related to their business activities and it also helps to improve food security.

Keywords:*Crop Forecasting, Machine Learning, Univariate Regression Models,*

1. Introduction

Agriculture is a major contributor to Indian Economy by providing employment opportunities to sixty percent of the total population. Agricultural products are mainly used as raw materials for lot of agro based industry and plays a vital role in developing the country's capital formation[1]. The extensive growth of human population in developing countries like India gives rise to a great demand for food production. A country should assure food security to its people by providing access to their daily food by physically, socially and economically at all time to meet their dietary needs for a healthy life. The global rise in population, changing in the global climatic environment, decrease in the geographical area of cultivated land and rising of price for food products have a significant impact on global food security. Proper care should be taken by governments to handle the water resources, increasing the geographical area of agricultural land, maintaining global warming, use of advanced technology for crop production and promotion of sustainable food preserving technologies to ensure food security. Crop forecasting helps in increasing food security by predicting the amount of crop before the harvest actually take place. The early prediction of crop yield helps in reduction of risk associated with production, processing, transport and consumption of food[2]. Machine Learning is a recent research area used in agriculture for developing crop forecasting models. The factors which have influence in crop production such as soil fertility, water level, moisture content in atmosphere were measured with the help of sensors and feed as input to the machine learning algorithm. The algorithm after processing the input data gives as output which called as models in machine learning. These developed models helps to predict the yield of the crop in advance before the actual harvest take place. The major contribution of this paper includes:

- A Disparate Univariate Regression Model is proposed to predict the yield of sugarcane varieties (CoC(Sc)22, CoC90063, TNAUSCSi7, TNAUSCSi8) with respect to the geographical area of cultivated land.
- Correlation Metrics were used to find the correlation between the hectares of land cultivated and production of sugarcane in tones.
- The general equation to predict the future yield of sugarcane was constructed with the input variable value and beta coefficients of the various regression models.
- The dataset is plotted in Hyperplane using the obtained equation of the various regression models.

2. Literature Review

Thomas van Klompenburga et al[3] in their paper made a detailed review on machine learning algorithms used in crop yield prediction. The author have reviewed 567 papers related to crop yield prediction and point outs the research gap present in the existing methods of crop prediction. The author also clearly give suggestion to the upcoming researchers about the areas to focus on the crop yielding research domain. The author reviewed papers which uses Random Forest, Support Vector Machine, Linear Regression and Neural Network for crop forecasting. The author finds that Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) performed well in crop forecasting by providing the accurate result for the selected crops. Amaury Dubois et al[4] in their paper used supervised machine learning algorithms to predict the soil moisture content in the soil used for potato cultivation. The authors used low cost sensors and tensiometers to measure the water level content in the soil and construct a dataset having parameters such as types of soil, climatic conditions etc. The author used feature selection to select the significant features and analyze the dataset with Neural Network, Random Forest and Support Vector Machine. The soil water potential at three layer depth is analyzed with the proposed model and the discovered knowledge is given to farmers for decision making in ecological and economical point of view.M.Kalimuthu et al[5] presented a paper and in it they used Guassian Naïve Bayes classifier to predict the crop yield. The author also designed an android mobile application which allows farmers to feed the information about the parameters of the crop such as temperature, humidity, moisture, rainfall etc and analyse the data using the Guassian Naïve Bayes algorithm and provide the extracted information to the farmer automatically about the crop yield prediction through their developed mobile application. Rai A. Schwalbertaet al[6] presented a paper and in it they used machine learning techniques such as Neural Network having Long Short Term Memory, multivariate OLS Regression Model and Random forest to predict the yield of Soyabean in southern Brazil. The author used satellite images of 80 municipalities in Brazil and mask the images and store them in a table format. The image data is analyzed with the machine learning and the parameters were tuned. Error metrics were used to analyze the performance of each classifier and the final extracted knowledge is used to forecast the yield of Soyabean. Wanie M. Ridwan et al[7] presented a paper and in it they used Regression Techniques such as Bayesian Linear, Decision Forest, Boosted Decision Tree and Neural Network Regressions to predict the amount of rainfall in TasikKenyir, Terengganu region. The results of each regression model was compared with the help of error metrics. In order to accurately predict the rainfall the author additionally used two methods namely Autocorrelation Function (ACF) and Projected Error method and find the correlation between the amount of rainfall with respect to different time zone such as daily, weekly, 10 days and monthly interval.

3. Materials and Methods Used

3.1 Data Collection

The data needed for this research was collected from Private Consultancies, District Agricultural Headquarters, and farmers cultivating sugarcane of Cuddalore and Villupuram Districts of Tamilnadu in India. The database contains totally two attributes namely Hectares of land and Sugarcane Yield in tones for the year 2018. Hectare is taken as the independent variable and Sugarcane yield is taken as Dependent variable. The total number of records present in the dataset is 10000. The variety of sugarcanes used in this study is CoC(Sc)22, CoC90063, TNAUSCSi7, TNAUSCSi8.

3.1. Simple Linear Regression

Regression is a statistical technique used to identify the relationship between the dependent variable that is the input variable and it is impact on the independent variable that is the output variable[6]. Linear Regression is a machine learning techniques that has one input variable and one output variable. The general equation for simple linear regression is given in Equation (1).

$$Y = a + bX \qquad \text{Equation (1)}$$

In the above equation, X represents the independent attribute and Y represents the dependent attribute. The independent attribute is always plotted on the X-axis and the dependent attribute is always plotted on the Y-axis. The y-intercept is denoted by a, and it is calculated by the Equation (2). The slope of the straight line formed by the linear equation is denoted by b and is calculated by Equation (3).

$$a=\frac{(\sum y)(\sum x^2)-(\sum y)(\sum xy)}{n(\sum x^2)-(\sum x)^2} \qquad \text{Equation (2)}$$

b=$\frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2}$ Equation (3)

3.2. Polynomial Regression

Polynomial Regression is used where there is a high positive correlation between the input variable and output variable but the relationship between the two variables is non-linear[7]. Polynomial equation model fit very well do the dataset having large observations. The general equation for polynomial regression is shown in equation (4).

yi=$\beta_0+\beta_1 x_i+\beta_2 x_i^2+\beta_3 x_i^3+\epsilon$i Equation (4)

Here in the above equation Y is the response variable, X is the input variable and β0, β1, β2 are called linear regression coefficients and ϵ is called residual error.

3.3. Decision Tree Regression

The decision tree regression model is built in the same procedure followed to build a decision tree which was proposed by J.R.Quinlan in 1986. This model partitions the dataset into smaller subset having homogenous values in the dataset. The decision tree regression model use standard deviation or mean squared error as the attribute selection measure for partitioning the dataset. The equation for calculating mean square error is shown in equation (5).

Mean Squared Error (MSE) =$\frac{1}{n}\sum(y-y_{pred})^2$ Equation (5)

In this research the mean squared error is used as the attribute selection measure. The mean squared error of the target variable with respect to the input variable is calculated and the attribute having highest means square error is chosen as the root node and the dataset is divided into subsets. The output of the decision tree regression model is a decision tree having root node of the attribute having highest mean square error and the leaf node is the average value of the input attribute having homogenous value.

3.4. Random Forest Regression

Random Forest Regression is an Ensemble method that makes prediction using decision tree algorithm and combines the result to produce the overall output. The number of estimators that is the number of sub trees used to generate was given initially to the random forest regression method. Bootstrap Aggregation or Bagging is used to select the subset of the dataset at random basis and apply the decision tree algorithm and produces an output.Then the subset of the dataset is replaced again by new subset of dataset and a decision tree is generated for the new dataset[8]. The result produced by each decision tree for the subset of the dataset is aggregated and the finally a random forest is constructed.

3.5. Support Vector Regression

Support Vector Regression is used to build linear and nonlinear models by defining a boundary for the regression line with a tolerance value called Epsilon denoted by ϵ. The value of the input features are allowed to plotted in an n-dimensional space and classes are separated by a hyper-plane which accurately classifies the classes without any outliers. The kernel function used to transform the data into a higher dimension space and form the correlation matrix is given in equation (6).

k(x^i-y^j)=exp(-$\gamma[\![x^{(i)}-x^{(j)}]\!]^2$), $\epsilon\delta_{ij}>0$ Equation (6).

To predict the new value of the target variables is done using the following equation.

y=$\sum_{i=1}^{N}(\alpha_i-\alpha_i^*).\langle\varphi(x_i),\varphi(x)\rangle+b.$ Equation (7).

3.6. Finding the Relationship of the Dependent and Independent Variable using Correlation Method

Correlation is a statistical method that is used to measure the relationship of two continuous variables and its value must be between -1 to +1. If the correlation value is in minus or zero then there is a weak relationship or otherwise the two variables are not related with one another. If the correlation value is positive then there is a strong relationship between the two continuous variables. This proposed model used four types of correlation methods namely Spearman Correlation, Pearson Correlation, Kendall Correlation and R^2 or otherwise called as Coefficient of determination.

The Spearman Correlation is calculated using the following Equation:

$$\rho=1-\frac{6\sum d_i^2}{n(n^2-1)}$$ Equation (8).

The Kendall Correlation is calculated using the following Equation:

$$\tau=\frac{n_c-n_d}{\frac{1}{2}n(n-1)}$$ Equation (9).

Pearson correlation is calculated using the following Equation

$$r_{xy}=\frac{n\sum x_i y_i-\sum x_i \sum y_i}{\sqrt{n\sum x_i^2-(x_i)^2}\sqrt{n\sum y_i^2-(y_i)^2}}$$ Equation (10).

The R2 or Coefficient of determination is calculated by squaring the value of Pearson correlation or it may be represented as

$$R^2=1-\frac{Variance\ \ Explained\ \ by\ the\ model}{Total\ \ Variance}$$ Equation (11)

The hypothesis framed for this model is

H0: There is no relationship between the hectares of land cultivated and yield of sugarcane produced in tones.

H1: There is a strong relationship between the hectares of land cultivated and yield of sugarcane in tones.

3.7. Error Metrics used for Evaluating the Performance of the Model

Once the regression module has successfully build and the dataset is tested with the model then it is necessary to check the performance of the model with the error metrics available in machine learning so as to make sure about the correctness of the predicted result. The yield of sugarcane with respect to the hectares of land cultivated was predicted with the help of the regression analysis and the following error metrics were used to test the performance of the regression model. The root mean square error (RMSE) which is the squared difference between the predicted value by the build model and the target value is calculated using the formula

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(y_{pred}-y)}{Total\ \ Number\ \ of\ Records}}$$ Equation (12)

Here $y_{pred}$is the yield of sugarcane predicted by the regression model and y is the actual value of yield of sugarcane in tones. The Mean Absolute Error (MAE) is absolute difference between the predicted yield of tons of sugarcane by the built model and the actual value of the yield of sugarcane produced in tones and it is calculated using the formula

$$MAE=\frac{1}{n}\sum|y-y_{pred}|$$ Equation (13)

The average of the squared difference between the predicted yields of sugarcane in tones with respect to the actual yield of sugarcane in tones is represented as Mean Squared Error and it is measured using the formula

$$MSE=\frac{1}{n}\sum\left(y-y_{pred}\right)^2$$ Equation (14)

The accuracy of the regression model is calculated by its Mean Absolute Percentage Error using the formula

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y-y_{pred}}{y}\right|$$ Equation (15)

4. PROPOSED DISPARATE UNIVARITE REGRESSION MODELFOR SUGARCANE PRODUCTION

The architecture of the proposed **Disparate Univariate Regression Model (DURM)** for Sugarcane Production is shown below in Figure 1. The proposed regression model uses regression analysis such as simple linear, polynomial, decision tree, random forest and support vector regression to build the regression model. The collected data is preprocessed and the dataset is divided into Training and Testing set. The regression model was built using the Training set and the future yield of sugarcane is predicted with the current values. The equation for each regression models was framed with x and y intercepts values. The correlation between the input variable hectares of land cultivated and the output variable yield of sugarcane were analyzed with the correlation metrics and the performance of the model was evaluated using the error metrics.

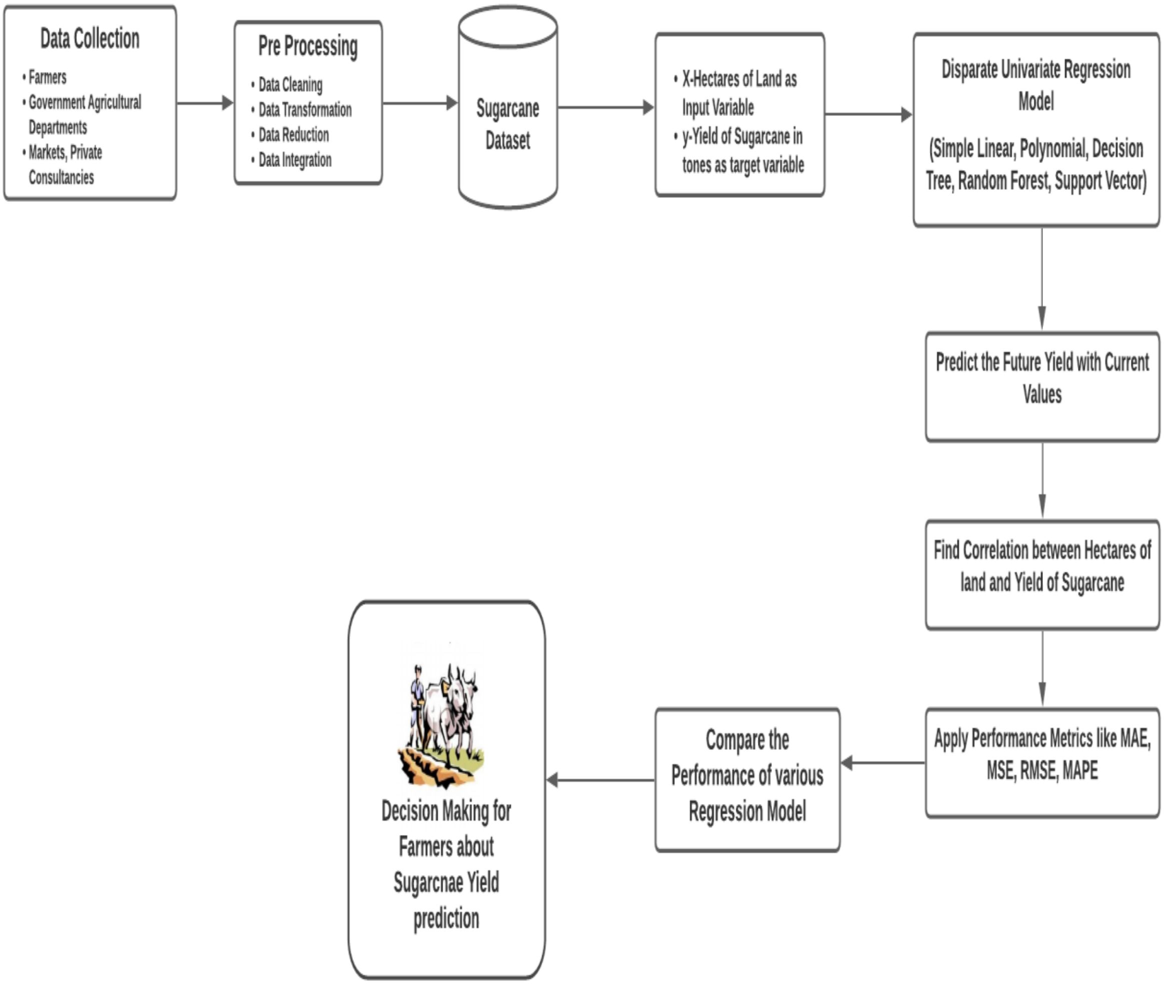

Data Collection
• Farmers
• Government Agricultural Departments
• Markets, Private Consultancies
Pre Processing
• Data Cleaning
• Data Transformation
• Data Reduction
• Data Integration
Sugarcane Dataset
• X-Hectares of Land as Input Variable
• y-Yield of Sugarcane in tones as target variable
Disparate Univariate Regression Model
(Simple Linear, Polynomial, Decision Tree, Random Forest, Support Vector)
Predict the Future Yield with Current Values
Find Correlation between Hectares of land and Yield of Sugarcane
Apply Performance Metrics like MAE, MSE, RMSE, MAPE
Compare the Performance of various Regression Model
Decision Making for Farmers about Sugarcnae Yield prediction

Figure 1: Architecture of the Proposed Disparate Univariate Regression Model (DURM)

5. Results and Discussion

The proposed model is implemented using Anaconda Navigator 3 with the help of Spyder as code editor. Python 3.8 version was used for coding. The machine learning package Scikit learn or Sklearn model was used for building the various Univariate regression models. The variable hectare is taken as the input variable and yield in tones is taken as the output variable for implementing the various regression models and graph obtained for various regression models is given above. The graph obtained from the various regression model shows that the predicted value of yield of sugarcane by the model is closely related to the actual value of the model. The regression curve in Figure 2 says that the relationship between the hectares of land and production in tones is linear. Figure 3 shows that the nonlinear data in the sugarcane dataset is clearly covered by the polynomial regression model and it accurately predict the yield of sugarcane for polynomial degree 4.The Decision Tree Regression model shown in Figure 4 and Random Forest Regression shown in Figure 5 predict the yield in the same manner, the only difference is Random Forest Regression selects the training and testing set randomly at each state whereas in Decision Tree Regression Model the training and testing data selected is not changed during the entire process. The SVR regression model built using RBF Kernel plots the predicted yield in its maximum margin shown in Figure 6 than the Linear Kernel shown in Figure 7. The nonlinear data present in the sugarcane dataset is properly covered by the polynomial and RBF kernel SVR model.

Figure 2: Predicted yield of sugarcane using Simple Linear Regression

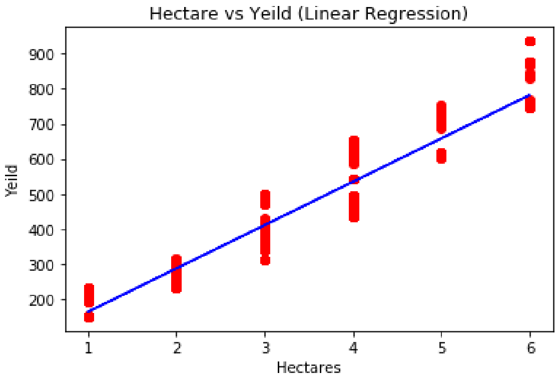


Figure 3: Predicted yield of sugarcane using Polynomial Regression

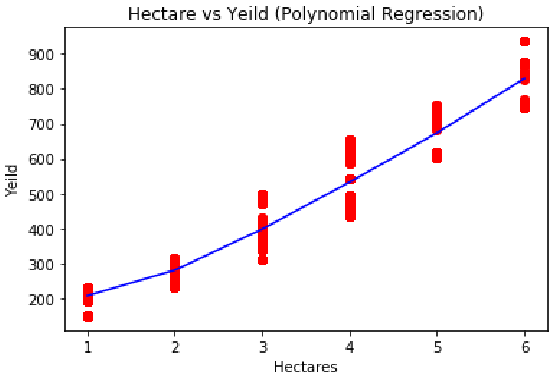


Figure 4: Predicted yield of sugarcane using Decision Tree Regression

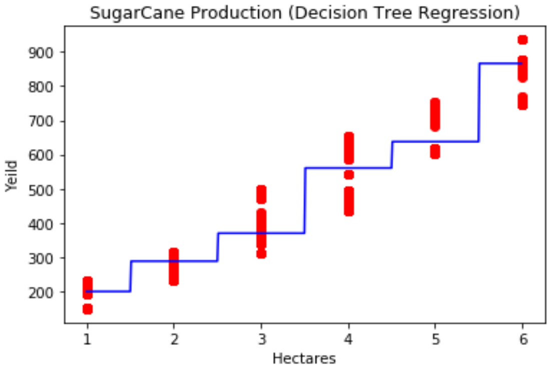


Figure 5: Predicted yield of sugarcane using Random Forest Regression

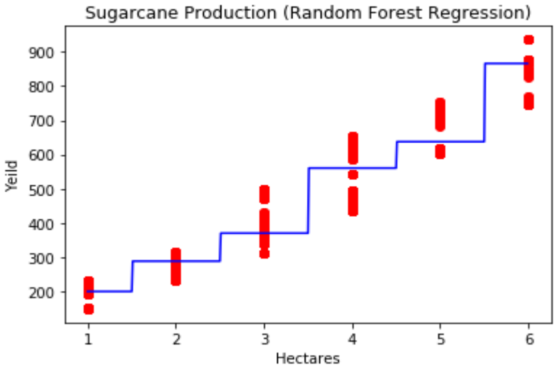


Figure 6: Predicted yield of sugarcane using SVR regression(RBF Kernel)

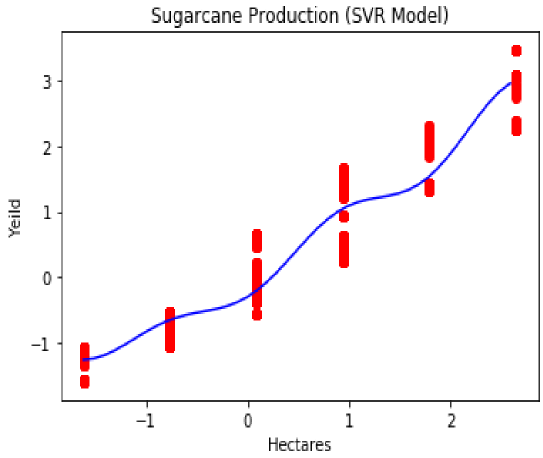
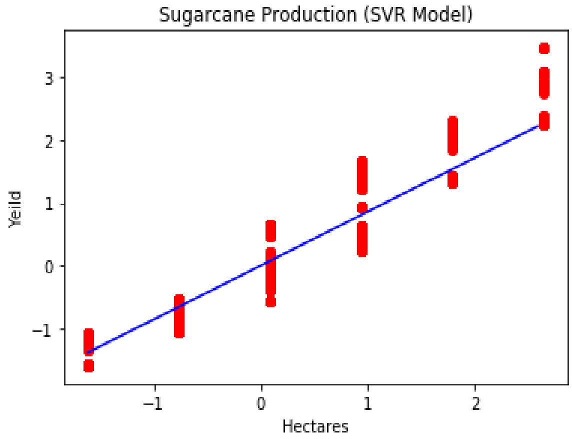


Figure 7: Predicted yield of sugarcane using SVR regression(Linear Kernel)



The various Correlation methods used to determine the relationship between the hectares of land cultivated and yield of sugarcane in tones was given in Table1 given below.

Table 1: Results of the Correlation Metrics for Hectare and Yield

S.No	Correlation Methods	Coefficient of Correlation	Interpretation
1	Spearman Correlation	0.9109	High Positive Correlation
2	Kendall Correlation	0.8091	High Positive Correlation
3	Pearson correlation	0.9414	High Positive Correlation
4	R^2 or Coefficient of Determination	0.8876	High Positive Correlation

The Correlation metrics clearly shows that there is *a high positive correlation between the input variable hectares of land cultivated and yield of sugarcane in tones*. In this situation the Alternate hypothesis has to accept; that is the yield of sugarcane highly depends on the hectares of the land it is cultivated. The various error metrics used to measure the performance of the various regression model is shown in Table2 given below.

Table 2: Results of the Error Metrics of Various Regression Models

S.No.	Error Metrics	Linear Regression	Polynomial Regression	Decision Tree Regression	Random Forest Regression	Support Vector Regression (Linear Kernel)
1	y-intercept	39.75	228.02	Nil	Nil	0.0002
2	Beta Coefficient	123.69	0,-86.13, 78.05, -12.44,0.76	Nil	Nil	0.85
3	Mean Absolute Error	42.33	40.86	34.44	34.42	0.26
4	Mean Squared Error	2657.89	2465.03	1963.83	1963.85	0.12
5	Root Mean Squared Error	51.55	49.64	44.31	44.31	0.34
6	Mean Absolute Percentage Error	11.49	10.62	9.18	45.20	234.29

From the error metrics shown in the Table 2 it was clear that the Decision Tree Regression and Random Forest Regression almost perform in a same manner and also produce the same result. When compared to linear regression the polynomial regression performs better. While considering the overall performance of the entire regression model the Support Vector Regression performs better and produces good results with best accuracy and least error. The various equation formed by the various regression models were listed in the Table 3.

Table 3: Equation of the Regression Models

S.No	Regression Models	Equation framed by the Model
1	Simple Linear	**y=123.69+39.75x**
2	Polynomial Regression (Degree=4)	**$y=0.76x^4-12.4x^3+78.05x^2-86.13x+0$**
3	Decision Tree Regression	There is no y-intercept and Beta Coefficients for decision tree algorithm, so equation can't be framed for this model
4	Random Forest Regression	There is no y-intercept and Beta Coefficients for random forest regression, so the equation can't be framed for this model
5	Support Vector Regression	**y=0.85x+0.0002**
In the above equations x is the input variable hectares, and y is the output variable yield.		

6. Conclusion

This research used machine learning technique called regression to predict the yield of sugarcane. The results of the proposed Disparate Univariate Regression Model (DURM) shows that there is a high positive correlation between the hectares of land cultivated and the yield of sugarcane produced in tones. That is the production of sugarcane increases with increase in hectares of land cultivated. The performance of each regression model is compared with the performance metrics; it shows that polynomial regression performs better than single linear regression. The performance of decision tree regression and random forest regression resembles same; the only difference is decision tree regression uses the same dataset for the entire model. But in random forest the samples of dataset was changed with Bootstrap Aggregation or Bagging method. While comparing all the regression technique used in Disparate Univariate Regression Model the support vector regression performs better, the regression model constructed by SVR has high accuracy and low error. Thus the machine technique implemented in the field of sugarcane farming predicts the yield of sugarcane. In future these proposed techniques can be applied to analyze various crops and can be applied in various domains such as Social Psychology, Forensic psychology, Educational Psychology etc.

References

Van Klompenburg, Thomas, AyalewKassahun, and CagatayCatal. "Crop yield prediction using machine learning: A systematic literature review." *Computers and Electronics in Agriculture, Vol.*177, pp: 105709, 2020.

Dubois, Amaury, Fabien Teytaud, and SébastienVerel. "Short term soil moisture forecasts for potato crop farming: A machine learning approach." *Computers and Electronics in Agriculture*, Vol.180, pp: 105902, 2021.

Kalimuthu, M., P. Vaishnavi, and M. Kishore."Crop Prediction using Machine Learning." *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE, pp:926-932, 2020.

Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V., &Ciampitti, I. A, "Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil", *Agricultural and Forest Meteorology*, Vol.*284*, pp:107886, Vol.2020.

Ridwan, W. M., Sapitang, M., Aziz, A., Kushiar, K. F., Ahmed, A. N., & El-Shafie, A, Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. *Ain Shams Engineering Journal*, pp:1-13, 2020.

Szymanski, Piotr, and Tomasz Kajdanowicz. "Scikit-multilearn: a scikit-based Python environment for performing multi-label classification." *The Journal of Machine Learning Research, Vol.*20, Issue-1, pp: 209-230, 2019.

Vasiloudis, Theodore, Gianmarco De Francisci Morales, and HenrikBoström."Quantifying Uncertainty in Online Regression Forests." *Journal of machine learning research, Vol.*20, Issue-155, pp: 1-35, 2019.

Ao, Y., Li, H., Zhu, L., Ali, S. and Yang, Z., “The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling”, *Journal of Petroleum Science and Engineering*, Vol.*174*, pp.776-789, 2019.