Team Members - TEAM PT4 :

Shravani Swaroop Urala (19BPS1019), Dharaneesh (19BPS1031), A L Vishwanath (19BPS1014)

Sr.No.	Title	Journal/year of publication	Data set used	Methodologies used	Metrics used	Interpretation of Results
1.	Extemporize Agriculture Yield with Predictions Based on Water and Soil Properties using Multivariate Analytics and Machine Learning Algorithm Link : https://www.ijeat.org/wp-content/uploads/papers/v8i6/F8137088619.pdf	International Journal of Engineering and Advanced Technology (IJEAT) ,August 2019	agriculture soil Testing Laboratory, Cuddalore District, Tamilnadu, India	Principal Component Analysis (PCA), Partial Least Squares regression (PLS), Support Vector Machines (SVM).	Growth of the crops depending upon its mixing ratio with soil, maximum yield from soil status, yield from water content dependencies.	Dataset has 13 attributes such village name as Ec, pH, N, P, K, Zn, Cu, Fe, Mn, B, Ca, Mg, and S. Analysis of the soil, water and field crop data using classification techniques and prediction techniques to predict the status for maximized yield. They have reported comprehensive study of various classification Algorithms with the Principal component Analysis, Partial Least Squares Regression, Descriptive statistics must perform efficiently.

2.	Data Science and Analytic Technology in Agriculture Link : https://www.ijcaonline.org/archives/volume179/number37/29283-2018916850	2018	USDA website for the state of Iowa. The yield and weather datasets contained 423 observations of harvested corn yield.	Support vector machine, random forest, multivariate polynomial regression	Seed Used per hectare, Source of Seed and Production Obtained, Quantity of Manure used per hectare, Quantity of Chemical Fertilizer per hectare,	The algorithms used are compared on basis of predicted yield, RMSE, MAE, median absolute error, and R-squared values. MAE does not give as much penalty to outliers as RMSE, so it is a better metric if outliers are few. MAE is also quite robust to outliers and will ignore outliers completely as it chooses only the median as compared to the mean in MAE. A lower value in RMSE, MAE, and MAE indicates better performance. Here, SVM performs better than the other two models in all three metrics. SVM performs better than the other models based on the R-squared value too.

3.	An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning Link : https://link.springer.com/article/10.1007/s11119-018-09628-4	2019	Several large farms in Western Australia were used as a case study, and yield monitor data from wheat, barley and canola crops from three different seasons (2013, 2014 and 2015) that covered ~ 11 000 to ~ 17 000 hectares in each year were used	Random forest	amount of yield monitor data for each crop, Total annual rainfall (mm) , soil properties i.e. moisture, texture , electrical conductivity per area	Predictions at the field resolution had a Lin's concordance correlation coefficient (LCCC) ranging from 0.19 to 0.27 for the leave-one-field year-out cross-validation (LOFOCV) technique, and ranging from 0.89 to 0.92 for the LOFYOCV technique. As the season progressed, the models performed slightly better, with the September models possessing the lowest RMSE, and the highest LCCC (Table 4). The significantly improved predictions of the LOFYOCV technique show the important benefit of including prior yield information for a particular field.

4.	Ensemble data mining approaches to forecast regional sugarcane crop production Link : https://www.sciencedirect.com/science/article/abs/pii/S0168192308003043	28 October 2008	Cane yields from 1976 to 2003 were obtained for Ayr (_198340, 1478240) (Fig. 1). With more than 70,000 ha of cane land, Ayr is a major coastal sugarcane growing region in Queensland, that produces approximately 20% of the total Australian sugarcane production.	3 Ensembles of different model settings were used, The first ensemble prediction was the simple average of the 840 biomass models. The second ensemble prediction was the mean of the reduced subset of the biomass models. These models were selected by the forward stagewise algorithm but instead of using the weights generated by the algorithm, equal weights were applied. The third ensemble also used the forward stagewise Algorithm to subset the biomass models but instead of choosing equal weights, standardized weights supplied by the forward stagewise algorithm were applied to the subsetted models	A predictive correlation represents a measure of how well a regression model predicts as opposed to fits a response. To invoke the prediction scheme, each observational unit (the measurements associated with each year) is left out one-by-one (also called leave-one-out cross-validation), and the independent variable (simulated yield) is regressed against yield. The ensemble technique demonstrated in this paper has provided a viable solution for overcoming obstacle of hetrogenity. We do stress however that with any statistical approach there is always the risk of artificially producing a good result and thus it is important to measure this risk.	if only the good models are included (those selected by the forward stagewise algorithm), the ensemble is more accurate than the single best model. The reduced average ensemble (rcv = 0.67) appreciably outperforms the single best model. Weighting models within the reduced subset, further improves the predictive correlation (rcv = 0.71)

5.	Prediction of Crop Production in India Using Data Mining Techniques Link : https://ieeexplore.ieee.org/document/8697446	2018	All the datasets were collected from the publicly available records of the Indian government for the duration of 64 years from 1950 to 2013. It consists of monthly rainfall, monthly mean temperature, area under irrigation, area, production and yield for the (1) Rice-Kharif season (June to December) and Rabi Season (January to June) (2) wheat-Rabi Season (October to May) (3) Maize- Kharif season (July to October) and Rabi season (October to April).	Multiple Linear Regression, Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) using scikitlearn and py-earth.	Mean Squared Error (MSE) and Root mean squared Error (RMSE) are the evaluation metrics used in regression analysis.	The experimental results showed that the performance of Multivariate Adaptive Regression Splines (Earth) was better compared to Multiple Linear Regression and Random Forest Regression on the Rice and Wheat dataset and the performance of Multiple Linear Regression was better compared to Random Forest Regression and Multivariate Adaptive Regression Splines (Earth) on the maize dataset

6.	Predicting Crop Diseases Using Data Mining Approaches: Classification Link : https://ieeexplore.ieee.org/document/8384523	2018	Dataset that was used has 155 records, 8 features. It has four classes named as: low, average, high and very-high in the dataset.	To analyse the damages, very well-known data mining techniques will be applied such as Decision Tree (DT), Random Forest (RF), Neural Networks (NN), Gaussian Naïve Bayes (GNB), Support Vector Machines (SVM) and KNearest Neighbors (KNN)	Accuracy, mean validation accuracy, precision, recall and F1-score are our criteria to evaluate the prediction model	It can be easily analysed that NN, RF and GNB produced better results as compared to other classifiers. Ensemble models with different combination of classifiers have been used to improve the accuracy of weak classifiers and these models also produced significant results.

7	Analysis of agriculture data using data mining techniques: application of big data Link : https://journalofbigdata.springeropen.com/articles/10.1186/s40537-017-0077-4	2017	Input dataset consist of 6 year data with following parameters namely: year, State-Karnataka (28 districts), District, crop (cotton, groundnut, jowar, rice and wheat.), season (kharif, rabi, summer), area (in hectares), production (in tonnes), average temperature (°C), average rainfall (mm), soil, PH value, soil type, major fertilizers, nitrogen (kg/Ha), phosphorus (Kg/Ha),Potassium(Kg/Ha), minimum rainfall required, minimum temperature required.	Modified approach of DBSCAN method is used to cluster the data based on districts which are having similar temperature, rain fall and soil type. PAM and CLARA are used to cluster the data based on the districts which are producing maximum crop production. Multiple linear regression method is used to forecast the annual crop yield.	Cluster performance was analysed using purity, homogeneity, completeness, V-measure, precision, recall, F-measure and Random index.	Various data mining techniques are implemented on the input data to assess the best performance yielding method. The paper used data mining techniques PAM, CLARA and DBSCAN to obtain the optimal climate requirement of wheat like optimal range of best temperature, worst temperature and rain fall to achieve higher production of wheat crop. Clustering methods were compared using quality metrics. It has been observed that DBSCAN gives the better clustering quality than PAM and CLARA, CLARA gives the better clustering quality than the PAM for the dataset

8	Agriculture Data Analytics in Crop Yield Estimation: A Critical Review Link :https://www.researchgate.net/publication/329467349_Agriculture_Data_Analytics_in_Crop_Yield_Estimation_A_Critical_Review	2018	The data was collected related to the principle rice crop yield influencing parameters such as different atmospheric conditions and various harvest parameters i.e Precipitation rate, minimum, average, maximum and most extreme temperature	Data mining technologies like Neural Networks, Support Vector Machine, Big Data analysis and soft computing in the assessment of agriculture field based on weather conditions. It also uses Regression Analysis to find the relationship between the explanatory variables and the crop yield which is considered as the response variable	Regression analysis parameters like MSE, MAE, R-squared value.	The paper discusses various technologies available for Researchers to use and predict the crop yield in their area. It presents various articles that have used different types of data mining and data analytic techniques.

9	Analysis of Crop Production Dataset using R Tool Link : https://www.ijeat.org/wp-content/uploads/papers/v9i1s4/A11001291S419.pdf	2019	The dataset is from Tamil Nadu agriculture dataset. This agriculture dataset includes 13,547 records which describes the crop production details of 31 districts of Tamil Nadu from 1997 to 2013. This data set contains Crop Data which is collected from different districts of Tamil Nadu. The dataset includes State_Name, District_Name, Crop_Year, Season, Crop, Area, Production details.	The data is analysed using Regression in various aspects such as a) crop produced in various season b) Production details of various crops c) crop produced in different years from 1997 to 2013 . It is used to find the crops that are frequently produced and are rarely produced	The paper uses various plots between the explanatory variables for each of the crops and the response variable (crop yield) to analyse and predict which crop is produced the most and which crop is produced rarely	The results and plots help the farmers and stake holders understand the crops where they can make the highest profits depending on the one that has the highest yield in Tamil Nadu.