# Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector

**Dr. Y. Jeevan Nagendra Kumar**[1]

[1]Professor and Dean TIC, Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, JNTUH, Hyderabad, India
jeevannagendra@griet.ac.in , 9010180199

**V. Spandana**[2], **V.S. Vaishnavi**[3], **K. Neha**[4], **V.G.R.R. Devi**[5]

[2 3 4 5]Department of Information Technology
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India.

*Abstract:* **Machine learning (ML) is a crucial perspective for acquiring real-world and operative solution for crop yield issue. From a given set of predictors, ML can predict a target/outcome by using Supervised Learning. To get the desired outputs need to generate a suitable function by set of some variables which will map the input variable to the aim output. Crop yield prediction incorporates forecasting the yield of the crop from past historical data which includes factors such as temperature, humidity, ph, rainfall, crop name. It gives us an idea for the finest predicted crop which will be cultivate in the field weather conditions. These predictions can be done by a machine learning algorithm called Random Forest. It will attain the crop prediction with best accurate value. The algorithm random forest is used to give the best crop yield model by considering least number of models. It is very useful to predict the yield of the crop in agriculture sector**

***Keywords—Supervised Learning, Naïve Bayes Algorithm, Regression, Decision Trees, Plots etc.,***

## I. INTRODUCTION

Agriculture is an important sector for Indian economy and also human future. It is first and foremost work which is essential for life. It also contributes a large portion of employment. As the time passes the need for production has been increasingly exponentially. In order to produce in mass quantity people are using technology in a wrong way. New kinds of hybrid varieties are produced day by day. However, these varieties do not provide the essential contents as naturally produced crop. These unnatural techniques spoil the soil. It all leads to further environmental harm. Most of these Unnatural techniques are used to avoid losses. But when the producers of these crops know the accurate information on the crop yield it minimizes the loss. To achieve this project is made. Using past information on weather, temperature and several other factors the information is given.

Data mining is machine learning tool which is used to view data in all possible ways and analyze. After analyzing the data, it is used to predict for future purposes. It can be used in several fields. These patterns provide information about crop.the aim of the project is to results to increase the yield and profit for producers. The proposed system concentrates on yield, weather predictions and crop type.the dataset is taken on agriculture statistics. This dataset is used as an experimental basis. After the data processing it is divided into training and testing.

## II. LITERATURE SURVEY

[1] Sujatha describes how the old farming data can be utilized to depict the future expectation of harvests and yield. It likewise proposes the ranchers about what kind of yield can be developed utilizing the climate station data and gives the appropriate data to incline toward the precise season for greatness cultivating.

[2] Amritha describes the forecast of yield utilizing IOT with the reasonable climatic conditions and the potential outcomes of progress and its application. They have utilized the Hadoop record framework. To manufacture an expectation framework for crops and to distinguish the nuisances the characterization, investigation and forecast calculation is utilized.

[3] Fathima describe the diverse mining procedures to consider crops that are quantitative and relationship them for interseason development. Grouping enormous information is a test, so k implies calculation is utilized to oversee huge information. Proper calculation is utilized to decide the harvests are chosen as incessant thing set. What's more, they centre around the administration strategies and the trimming practices of edges.

[4] Nirupama describes significant job that performed by data mining strategies in farming field. They have introduced the distinctive ML calculations, for example, k implies, SVM, ANN and so on. The harvests were anticipated for the most part dependent on climatic highlights which gives exactness score of about 95% with the C4.5 calculation.

[5] Sajitha describes the different components that associated with ecological parameters which impact

978-1-7281-5371-1/20/$31.00 ©2020 IEEE

the yield of the harvest are Area under Cultivation, Annual Rainfall, and Food Price Index and set up the relationship among these parameters. The curse on the harvest yield is broke down by utilizing different ecological elements and Regression Analysis (RA), Linear Regression (LR) Algorithm.

[6] Raorane describes by utilizing the various data mining strategies how to improve the harvest production. And the procedures they have utilized for order, for example, ANN, SVM and k implies and so on.

## III. METHODOLOGY

Crop yield is a very useful information for farmers. It is very beneficial to know the yield which results in reduction in loss. In the past the yield prediction is done by experienced farmers. The proposed system also works in a similar way. It takes the previous information and uses it to predict the future yield. The crop yield mainly depends on weather and pesticides. This prediction is proportional to the accuracy on information provided. Therefore, the proposed system predicts the yield and decreases the loss [11]

The anticipated system acts as experienced farmer. But, with more accuracy and considers many other factors. Factors like soil condition, weather prediction, yield. The more increase in accuracy results in more profit in crop yield. To increase accuracy the data has to be perfect. With all the information provided the proposed system process all the data using data mining methods and predicts the harvest yield. With this forecast the farmer will be able to know his requirements [9]

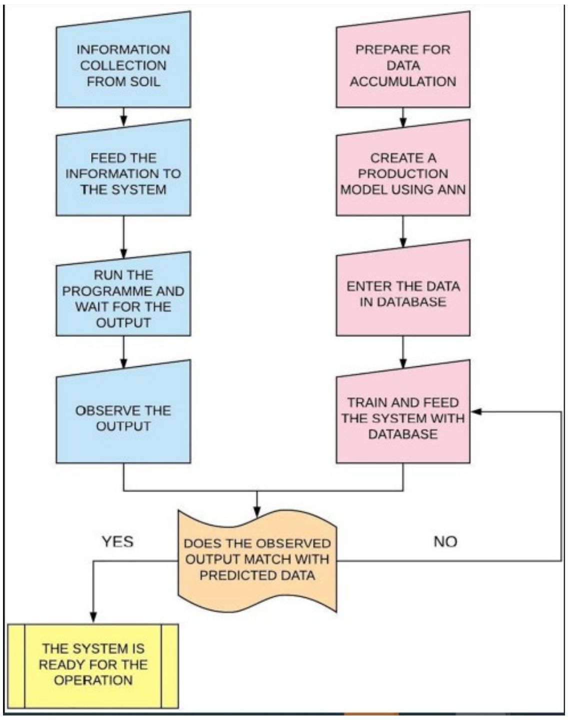

INFORMATION COLLECTION FROM SOIL
FEED THE INFORMATION TO THE SYSTEM
RUN THE PROGRAMME AND WAIT FOR THE OUTPUT
OBSERVE THE OUTPUT
PREPARE FOR DATA ACCUMULATION
CREATE A PRODUCTION MODEL USING ANN
ENTER THE DATA IN DATABASE
TRAIN AND FEED THE SYSTEM WITH DATABASE
DOES THE OBSERVED OUTPUT MATCH WITH PREDICTED DATA
YES
NO
THE SYSTEM IS READY FOR THE OPERATION

Fig 1: Workflow of proposed system

**Random Forest Classifier:**

At training situation multitude decision trees are made and the output will be divided based on number of classes i.e., classification, prediction of class i.e., regression. The number of trees is proportional to accuracy in prediction. The dataset includes factors like rainfall, perception, temperature and production. These factors in dataset is used for training. Only two-third of the dataset is considered. Remaining dataset is used for experimental basis.

### A. Datasets

The dataset consists of factors like temperature, rainfall, humidity, ph. The datasets have been obtained from the Kaggle website [10]

The data set has 3101 instance or data that have taken from the past historic data. It includes 5 parameters or features like the temperature, ph., humidity, rainfall and crop name.

**Random forest algorithm**

Random Forest is a ML algorithm. At training situation multitude decision trees are made and the output will be divided based on number of classes i.e., classification, prediction of class i.e., regression. The number of trees is proportional to accuracy in prediction. The dataset includes factors like rainfall, perception, temperature and production. These factors in dataset is used for training. Only two-third of the dataset is considered. Remaining dataset is used for experimental basis.

The algorithm random forest has 3 parameters like: n tree which describes the n number of trees which need to grow, m try - mentions how many variables need to be taken at a node split. Node size - In terminal nodes it suggest us the number of observation need to take. [7]

### B. Decision Tree

Decision tree classifiers utilizes greedy methodology henceforth a feature picks from the start move cannot be consumed any longer, that gives us best grouping whenever utilized in further advances. Likewise, it over fit the preparation information which can give poor outcomes for inconspicuous information. In this way, to beat this confinement gathering model is utilized. In gathering model outcomes from various models are consolidated. [8]

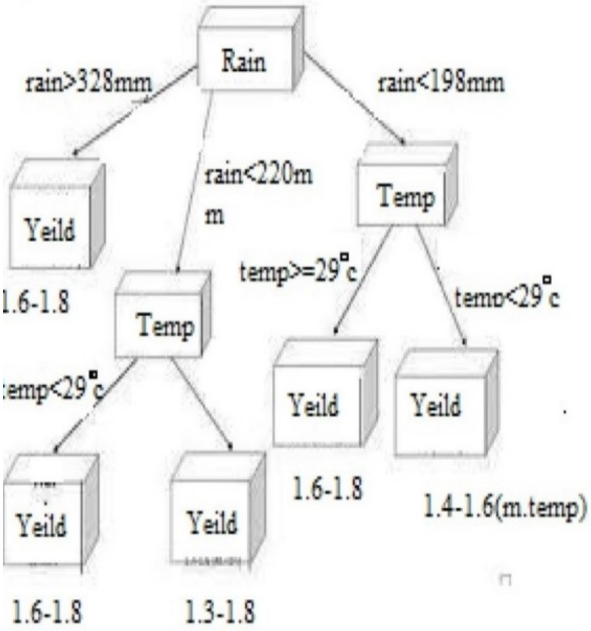


Fig 2: Decision tree

### C. Procedure

Dataset consists of few important attributes such as temperature, rainfall, humidity, ph., applied random forest algorithm for classification and regression tasks.

Tried to train the model with decision trees but found that random forest algorithm reduced the overfitting problem and also improved accuracy and used SVR, Random forest and random forest got more accuracy.

The dataset which used is imported from Kaggle repository. From the dataset, used 80% of data for training the model and 20% of data for testing the results to obtain better results and trained the model by applying random forest algorithm. Then, compared the predicted result with the original data set. Later, estimated the accuracy of the model using test samples.
Likewise, predicted the accuracy of the model with different algorithms. Then finally concluded that random forest algorithm gives us more accuracy. Hence, used random forest algorithm to train the model.

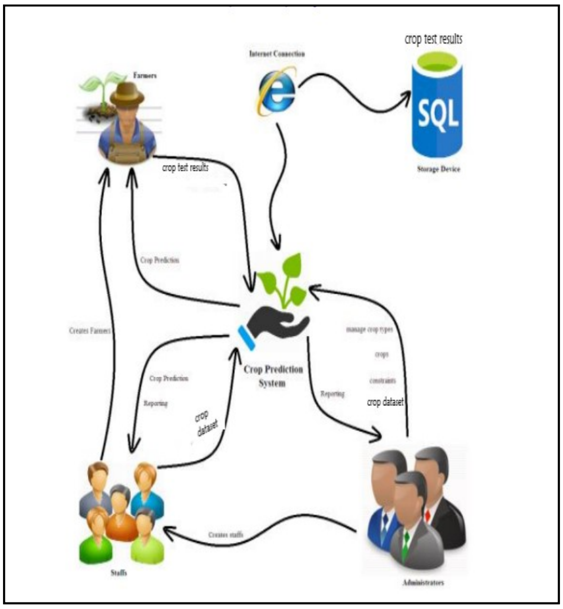


Fig 3: Procedure

## IV. RESULTS

Data visualization is the representation of understanding data by showing it in a graphical context, so that the designs, inclinations and connexions can be detected and exposed. [14]

The few prevalent plotting collections:

- Matplotlib: Small level, provides lots of liberty to user.
- Pandas Visualization: comfortable to use this boundary. It can be constructed on Matplotlib
- Plotly: Can create interactive plot for visualization [15]

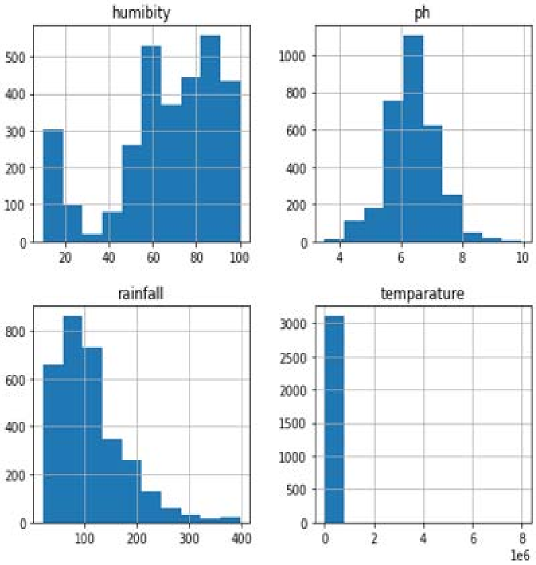


Fig 4: Hist Diagram

The attributes rainfall land humidity may have an exponential distribution. The attribute temperature is easy to notice the distribution is skewed very much to left. The attribute ph. has a Gaussian or nearly Gaussian distribution. [13]

Density plot comes under univariate plots. It is also another technique for getting the distribution of each attributes in dataset. A density pot is smoothed, extended version of a histogram. Density plot shows the distribution of a numerical variable. It takes input as the only set of numerical values.

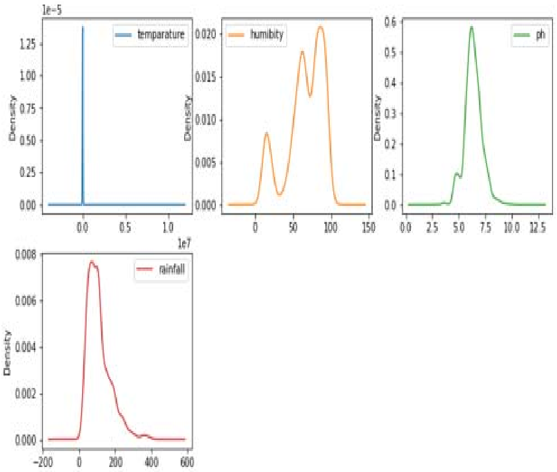


Fig 5: Density Diagram

Box plot also comes under univariate plot. Its functionality is similar to density plot. It visualizes the distribution of each attributes. [16]

- The green line represents median of the attribute
- The rainfall attribute has appear nearly skewed towards smaller values
- The humidity attribute has skewed towards Larger values
- The ph attribute is skewed neither towards smaller values nor larger values

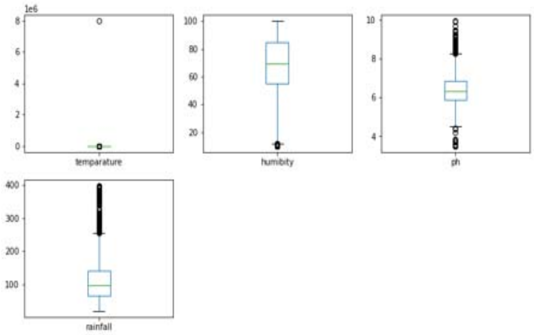
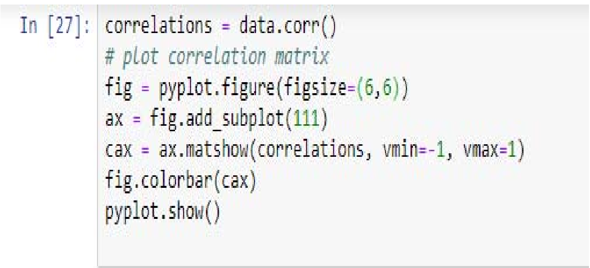


Fig 6: Box Diagram

- Correlation between the attributes helps us to figure out how strong or how weakly they are related to each other. The numeric values 1 represents the positive relationship exists between the variables. The numeric values 0 with darker colour gets the more negative relationship exists between the variables.



```
In [27]: correlations = data.corr()
         # plot correlation matrix
         fig = pyplot.figure(figsize=(6,6))
         ax = fig.add_subplot(111)
         cax = ax.matshow(correlations, vmin=-1, vmax=1)
         fig.colorbar(cax)
         pyplot.show()
```

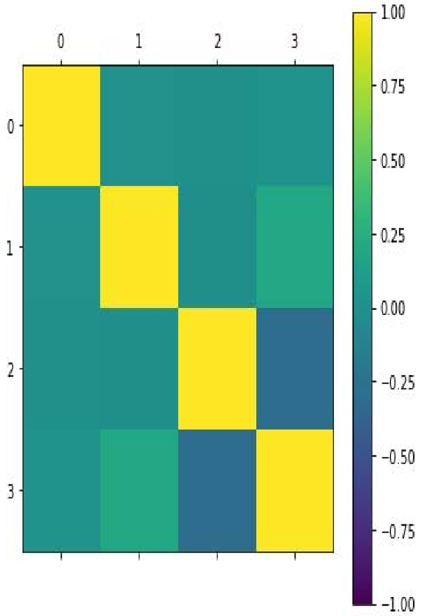


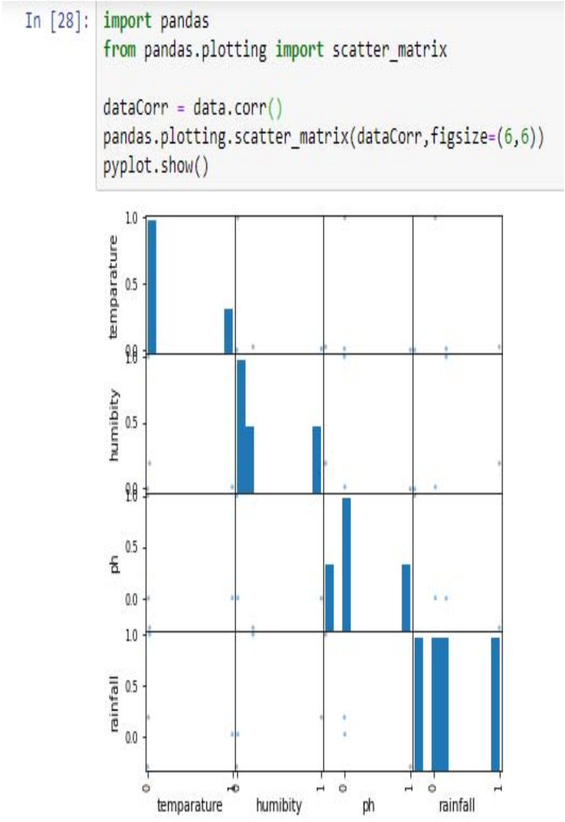
Fig 7: Correlation diagram

- A correlation plot matrix can be formed for a collection of variables with each other variables will be plotted against each other. Here have four columns where normally distributed with random values and column names are: temperature, humidity and rainfall.

- The attributes rainfall land humidity may have an exponential distribution.

- The attribute temperature is easy to notice the distribution is skewed very much to left.

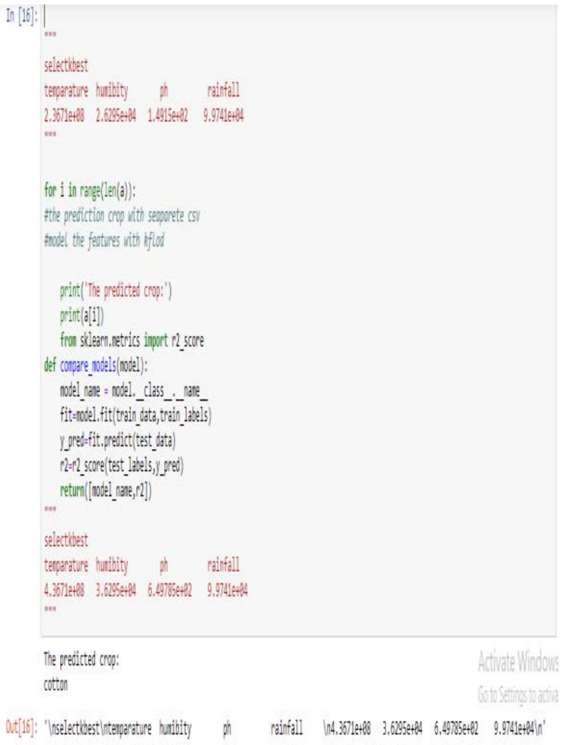
- The attribute ph. has a Gaussian or nearly Gaussian distribution.



```
In [28]: import pandas
         from pandas.plotting import scatter_matrix

         dataCorr = data.corr()
         pandas.plotting.scatter_matrix(dataCorr,figsize=(6,6))
         pyplot.show()
```

Fig 8: Correlation Matrix plot diagram



```
In [16]: |
         """
         selectkbest
         temparature  humibity     ph        rainfall
         2.3671e+08   2.6295e+04   1.4915e+02   9.9741e+04
         """


         for i in range(len(a)):
         #the prediction crop with seaparete csv
         #model the features with kflod


            print('The predicted crop:')
            print(a[i])
            from sklearn.metrics import r2_score
         def compare_models(model):
            model_name = model._class_._name_
            fit=model.fit(train_data,train_labels)
            y_pred=fit.predict(test_data)
            r2=r2_score(test_labels,y_pred)
            return([model_name,r2])
         """
         selectkbest
         temparature  humibity     ph        rainfall
         4.3671e+08   3.6295e+04   6.49785e+02   9.9741e+04
         """

         The predicted crop:
         cotton

Out[16]: '\nselectkbest\ntemparature  humibity      ph        rainfall    \n4.3671e+08   3.6295e+04   6.49785e+02   9.9741e+04\n'
```

Fig 9: Prediction snapshot

## V. Conclusion

Implement a system to predict crop production from the collection of past data. Using data mining techniques crop yield is predicted. Here, using Random Forest algorithm for predicting the best crop yield as output. In agriculture field, the crop yield prediction is mostly appropriate. The more increase in accuracy results in more profit to the crop yield. The proposed technique helps farmers to acquire apprehension in the requirement and price of different crops. It helps farmers in decision making of which crop to cultivate in the field. The more increase in accuracy results in more profit to the crop yield. This work is employed to search out the gain knowledge about the crop that can be deployed to make an efficient and useful harvesting. Under this system, maximum types of crops will be covered. The accurate prediction of different specified crops across different districts will help farmers of India.

## VI. Future work

This research work can be enhancing to the high level by building a recommender system of agriculture production and distribution for farmer. By which farmers can make their own decision like which season which crop should sow so that they can get better profit. This system works for structured dataset or database. In coming years try applying data independent system also that meanas the format may be whatever, our system should work with same accuracy.

## References

[1] https://en.wikipedia.org/wiki/Agriculture
[2] https://en.wikipedia.org/wiki/Data_analysis
[3] JeetendraShenoy, YogeshPingle, "IOT in agriculture", 2016 IEEE.
[4] M.R. Bendre, R.C. Thool, V.R. Thool, "Big Data in Precision agriculture", Sept, 2015 NGCT.
[5] Monali Paul, Santosh K. Vishwakarma, Ashok Verma, "Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining approach", 2015
[6] International Conference on Computational Intelligence and Communication Networks.
[7] N. Heemageetha, "A survey on Application of Data Mining Techniques to Analyze the soil for agricultural purpose", 2016IEEE.
[8] https://en.wikipedia.org/wiki/Linear_regression
[9] Y. Jeevan Nagendra Kumar, Dr. T. V. Rajini Kanth, "GIS-MAP Based Spatial Analysis of Rainfall Data of Andhra Pradesh and Telangana States Using R", International Journal of Electrical and Computer Engineering (IJECE), Vol 7, No 1, February 2017, Scopus Indexed Journal, ISSN: 2088-8708
[10]B Sankara Babu, A Suneetha, G Charles Babu,Y.Jeevan Nagendra Kumar, G Karuna, "Medical Disease Prediction using Grey Wolf optimization and Auto Encoder based Recurrent Neural Network", Periodicals of Engineering and Natural Sciences, June 2018 ISSN 2303-4521 Vol.6, No.1, pp. 229~240
[11]Dr. Y. Jeevan Nagendra Kumar, Guntreddi Sai Kiran, Partapu Preetham, Chila Lohith, Guntha Sai Roshik, G. Vijendar Reddy, "A Data Science View on Effects of Agriculture &

Industry Sector on the GDP of India" International Journal of Recent Technology and Engineering, Volume-8, Issue-1, May 2019, ISSN: 2277-3878

[12] Y. Jeevan Nagendra Kumar, N. Kameswari Shalini, P.K. Abhilash, K. Sandeep, D. Indira, “Prediction of Diabetes using Machine Learning” International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, Volume-8 Issue-7 May 2019

[13] Y. Jeevan Nagendra Kumar, B. Mani Sai, Varagiri Shailaja, Singanamalli Renuka, Bharathi Panduri, “Python NLTK Sentiment Inspection using Naïve Bayes Classifier” International Journal of Recent Technology and Engineering, ISSN: 2277-3878, Volume-8, Issue-2S11, Sep 2019

[14] D. Srinivasa Rao, Ch. Ramesh Babu, Y. J. Nagendra Kumar, N. Rajasekhar, T. Ravi, “Medical Image Fusion Using Transform Based Fusion Techniques”, International Journal of Recent Technology and Engineering, Volume-8 Issue-2 ISSN: 2277-3878

[15] Srikanth Bethu, V Sowmya, B Sankara Babu, G Charles Babu, Y. Jeevan Nagendra Kumar, “Data Science: Identifying influencers in Social Networks”, Periodicals of Engineering and Natural Sciences, ISSN 2303-4521 Vol.6, No.1, pp. 215~228

[16] Raj, J. S., & Ananthi, J. V. (2019). RECURRENT NEURAL NETWORKS AND NONLINEAR PREDICTION IN SUPPORT VECTOR MACHINES. Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40.

## AUTHOR PROFILE

Dr. Y. Jeevan Nagendra Kumar, obtained his Ph.D. in Computer Science and Engineering from Acharya Nagarjuna University, Guntur, AP in 2017 and MTech Computer Science Technology from Andhra University in 2005. He is working as Professor and Dean - Technology and Innovation Cell in GRIET since 2005.

He has about 16 Research Papers in International / National Conferences and Journals and also attended many FDP Programs to enhance his knowledge. With his technical knowledge he guided the students in developing the useful Web applications and data mining related products. As B O S member was able to introduce new subjects, topics in UG/ PG Courses. Students are encouraged to work on research projects, engineering projects as well as for industrial training.

He was acted as Coordinator for 3 International Conferences and Technical Committee member for several International Conferences. He is Coordinator for J Lab under J Hub JNTUH and Robotic Club. Also, Coordinator for NBA and NAAC at College Level.

Currently acting as Convener and Vice-President for MHRD IIC (Institution’s Innovation Cell) GRIET.