Rainfall Prediction for Udaipur, Rajasthan Using Machine Learning Models Based on Temperature, Vapour Pressure and Relative Humidity

Jitendra Shreemali, Praveen Galav, Gaurav Kumawat, Pankaj Chittora

Abstract: The study aims at Rainfall prediction using Machine Learning models using the minimum of features. The prediction here is based on temperature, vapour pressure and relative humidity. Numerous studies carried out earlier used more features than this study. A training-test split of 75-25 was used. The best results were obtained by combining the best of the candidate models into an ensemble model to identify that predictor importance of vapour pressure was 0.89 while that of relative humidity was 0.11 with temperature not seen as a significant predictor for rainfall though the high correlation of temperature (°C) with vapour pressure (Torr) and relative humidity (Percentage) suggests that the two predictor variables subsume the impact of temperature.

Keywords: Rainfall prediction, Neural Network, Ensemble model, CHAID, Random Forest

I. INTRODUCTION

Predicting weather patterns especially rainfall has been a challenge that mankind has grappled with since times immemorial. With agricultural production being directly or indirectly dependent on rainfall, it continues to draw immense interest among scientists as well as agriculturists across the globe. Rainfall also plays an important role in ensuring heat balance in the atmosphere on account of its impact on atmospheric circulation across the world. Katsaros and Buettner (1969) carried out an experimental study using a salt water tank to estimate impact of falling rain drops on salinity as well as temperature to understand how rain could affect oceans. They report that while smaller drops created a very stable surface, the larger ones led to increased mixing. Thus the impact of rainfall extends well beyond agriculture to multiple aspects of human existence.

Rainfall itself is the result of multiple and closely integrated natural processes that makes simulating the process model a very challenging task.

Revised Manuscript Received on March 5, 2020.

Jitendra Shreemali, Techno India NJR Institute of Technology, Udaipur,**jitendrapshreemali@gmail.com***

Praveen Galav, Techno India NJR Institute of Technology, Udaipur

Gaurav Kumawat, Techno India NJR Institute of Technology, Udaipur

Pankaj Chittora, Techno India NJR Institute of Technology, Udaipur

However, the advancement of machine learning tools for predictive modeling preset an opportunity to fine tune the prediction accuracy but presents the challenge for researchers to identify appropriate model and relevant features that have a bearing on rainfall levels.

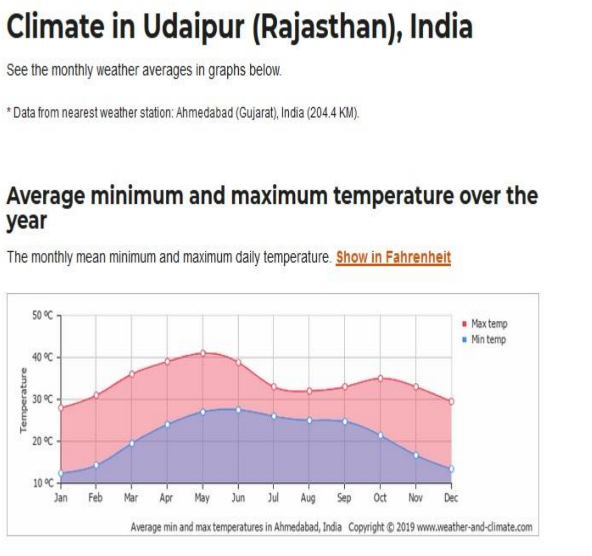
II. LITERATURE REVIEW

A comparison of different machine learning algorithms by Singh and Kumar (2019) finds that while adaptive boosting algorithm gave an F-score of 0.9726 on the test data while K-nearest, Neural Network and SVM gave F-scores of 0.8754, 0.7946, and 0.8045 respectively suggesting a superiority of adaptive boosting algorithm over the others listed here based on F-score without feature selection.

Hong (2008) described using a hybrid model of RNN and SVM for forecasting rainfall depth values. The parameters of a SVR model (referred to as RSVR model) were chosen using the chaotic particle swarm optimization algorithm (CPSO). And applied to rainfall values during typhoon periods from Northern Taiwan. Their study provides a forecasting performance making the RSVRCPSO model an alternative worth considering for forecasting rainfall values. Based primarily on classification in terms of high-low-average Mohd, Butt and Baba (2018) considered the following parameters as candidate features for predicting rainfall: date, temperature in °C, Dew point in °C, Humidity as percentage, sea level pressure in hPa, visibility in KMs, wind speed in KM/h and precipitation in mm with the events of interest being the rainfall (snow, thunderstorm, fog) but used the average temperature, humidity, sea level pressure and wind speed to predict rainfall. They reported accuracy ranging from 82.56% to 87.76% and corresponding precision from 0.815 to 0.874. Janbandhu, Meshram and Gedam (2017) report using Bayesian Model to predict rainfall (mm) based on the following features: Temperature (°C), Station Level Pressure (hpa), Mean Sea Level Pressure (hpa), Relative humidity (percentage), Vapour Pressure (hpa) and Wind speed (Km/hour).

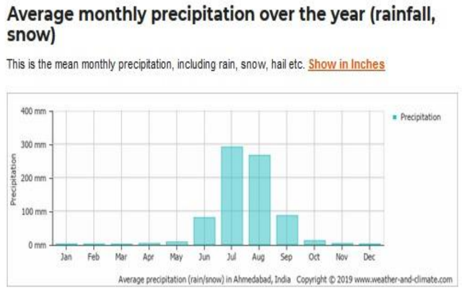
The results in use of monsoon season data across three cities, namely, Pune, Mumbai and Delhi are seen to be above 90% in all three cases. Swapna and Sudhakar (2018) report using the Long Short Term Memory (LSTM) deep Learning Model for rainfall prediction in coastal Andhra Pradesh with the features including Maximum as well as, Minimum temperature (both in °C), Wind, Pressure and Visibility. For inputs they used parameterized data inputs to predict rainfall. Using classification algorithms like SVM (Support Vector Machines), 2 layered ANN (Artificial Neural Networks) and logistic regression, Tarun et. al (2019) carried out qualitative analysis on data from the hydrological department of Rajasthan using 12 features and reported an accuracy of greater than 85% for the test data besides a precision value of 96% and recall of 91.4% for the logistic regression. Aftab et.al (2018) list the features used in different machine learning models aimed at predicting rainfall as including: polarity, quantity of rainfall, maximum and minimum temperature, humidity levels, wind speed etc. with the quality of prediction being a function of multiple factors including training algorithm, climatic attributes used as features and pre-processing techniques besides others. Oswal (2019) reported using a dataset with 23 features including location, minimum and maximum temperature, rainfall for the day in milimetres, evaporation, sunshine, wind gust direction, wind gust speed, wind direction at 9 AM and 3 PM, wind speed at 9 AM and 3 PM, humidity at 9 AM and 3 PM, atmospheric pressure at 9 AM and 3 PM, fraction of sky obscured by clouds at 9 AM and 3 PM and precipitation. While a large number of features does increase the chances of greater accuracy, it inevitably increases data collection and computation costs. To predict rainfall through reduced features thereby reducing data collection costs and computation efforts was the primary motivating factor behind this study.

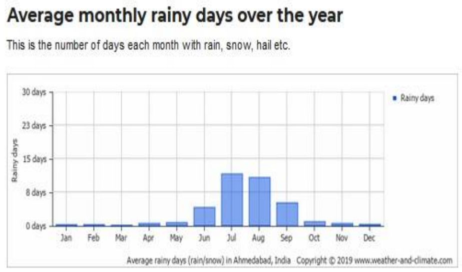
Udaipur, a city also referred to as the City of Lakes is a popular tourist spot for people all over the world. The city was founded in 1558 by Maharana Udai Singh at a very scenic spot on the south slope of Aravalli Ranges in Rajasthan, India. Wikipedia quotes James Tod on Udaipur being "…the most romantic spot on the Continent of India". It's natural beauty and attractiveness for tourists depends almost entirely on the rainfall in the area. Udaipur's weather in terms of temperature and rainfall are given below:



Source: https://weather-and-climate.com/average-monthly-Rainfall-Temperature-Sunshine,udaipur,India

Rainfall data for Udaipur is summarized below:





Source: https://weather-and-climate.com/average-monthly-Rainfall-Temperature-Sunshine,udaipur,India

Figure 1: Udaipur Weather (Temperature and Rainfall)

III. DATA COLLECTION AND PREPARATION

Data on rainfall and temperature was collected from government site for data (https://data.gov.in) while data for Relative Humidity and Vapour Pressure was collected from the NASA site (https://search.earthdata.nasa.gov). Temperature, relative humidity and vapour pressure during the month from 1983 to 1990 were used as features to predict monthly rainfall. Since all possible values were not available available, average values were treated as estimator for Temperature, relative humidity and vapour pressure during the month bringing an element of error into the system because these values are not consistent during the entire month.

Table 1: Weather Data Elements

Sl.No.	Attribute	Data Type	Units
1	Temperature	Continuous	°C
2	Vapour Pressure	Continuous	Torr
3	Relative Humidity	Continuous	Percent
4	Rainfall (Target variable)	Continuous	MM

IV. MODEL TRAINING AND EVALUATION

IBM SPSS Modeler was used to build and train the models. Given the fact that all variables are continuous in nature, following twelve models were considered for model building. These are: (i) Regression; (ii) Generalized regression; (iii) Linear-AS; (iv) LSVM; (v) Random Trees; (vi) Tree-AS; (vii) XGBoost Linear; (viii) XGBoost Tree; (ix) Linear; (x) CHAID; (xi) Random Forest; (xii) Neural Net. A Partitioning of 75% for training and 25% for testing was used. A representation of these is presented below:

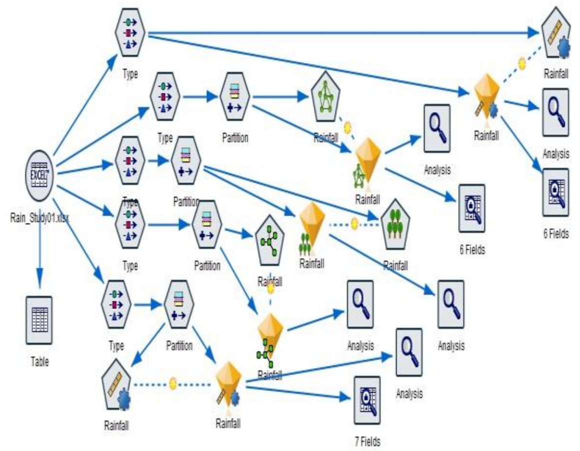


Figure 2: Schematic of Various Models Built

Multiple models were attempted to improve the accuracy.

V. RESULTS AND DISCUSSIONS

Use of multiple machine learning algorithm increases the chances of getting better results when no single algorithm is seen to produce optimal results under all conditions. In this specific case, the following six models producing the best results (in descending order): (i) Linear; (ii) Neural Net; (iii) Regression; (iv) Generalized Linear; (v) CHAID; and (vi) Random Forest. The result reported relate to the ensemble model that helps address limitations of individual models and also enhances accuracy. These results are presented below as Figure 3:

Sort by: Use | Ascending | Descending | Delete Unused Models | View: Testing set

Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		Linear 1	2	0.909	2	0.584
☑		Neural Net 1	2	0.905	3	0.380
☑		Regression 1	2	0.886	3	0.708
☑		Generalized Linear 1	2	0.886	3	0.708
☑		CHAID 1	2	0.868	1	0.833
☑		Random Forest 1	2	0.839	3	0.389

Figure 3: Summary of ML Models

Considering the variation in results across different machine learning models, the results of each model in terms of error, standard deviation and correlation (eg. linear correlation values falling between 0.909 to 0.636) are presented below:

'Partition'	1_Training	2_Testing
Minimum Error	-155.103	-141.909
Maximum Error	169.68	43.401
Mean Error	1.755	-5.087
Mean Absolute Error	34.053	25.764
Standard Deviation	54.603	39.135
Linear Correlation	0.852	0.889
Occurrences	58	24

Figure 4: Ensemble Model Results

'Partition'	1_Training	2_Testing
Minimum Error	-113.205	-100.264
Maximum Error	179.717	68.166
Mean Error	6.42	-5.132
Mean Absolute Error	30.507	13.104
Standard Deviation	53.598	28.802
Linear Correlation	0.86	0.905
Occurrences	58	24

Figure 5: Neural Network Results

'Partition'	1_Training	2_Testing
Minimum Error	-11.278	-235.442
Maximum Error	66.558	107.626
Mean Error	5.067	-8.391
Mean Absolute Error	5.913	23.543
Standard Deviation	12.985	57.064
Linear Correlation	0.996	0.636
Occurrences	58	24

Figure 6: Random Trees Results

'Partition'	1_Training	2_Testing
Minimum Error	-86.8	-47.2
Maximum Error	82.0	106.52
Mean Error	0.522	-1.969
Mean Absolute Error	16.027	13.871
Standard Deviation	29.496	27.496
Linear Correlation	0.964	0.823
Occurrences	58	24

Figure 7: Random Forest Model Results

Given below is a very brief description of these models listed above:

Regression Models: The Linear (regression) model for machine learning is built around the equation:

$\boxed{Y_{Pred} = B_0 + B_1*X_1}$ with B_0 signifying the intercept (bias) and B_1 signifying the slope (weight) of the parameter. Depending upon the number of features the prediction can be a line, a plane or a hyperplane. Machine learning uses different kinds of regression models where the key difference lies in the underlying equation, how parameters are learnt during training and how model complexity is kept under control to avoid overfitting.

Generalized linear (regression) models extend the simple method above to cater to added complexity brought in by multiple impacting variables. This equation takes the general form:

$\boxed{Y_{Pred} = b_0 + b_1*X_1 + b_2*X_2 + b_3*X_3 + \ldots}$ with the model estimating all b_i's.

Where linearity does not hold, models would go in for non-linear functions (eg. the sigmoid function).

Neural Net: Neural networks are algorithms modeled around the human brain with the nodes representing neurons and layers in the network representing clustering, classification or regression algorithms to manage the data fed at the input layer.

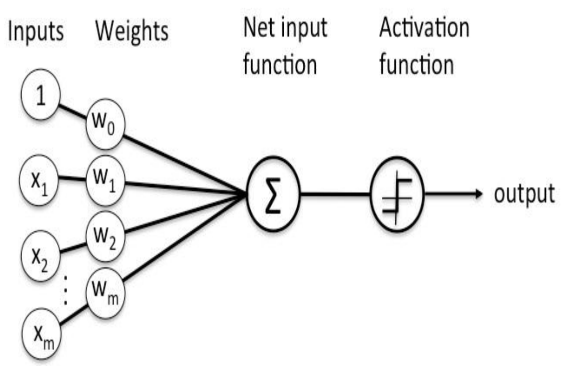


Figure 8: Schematic of a Perceptron

Source: https://pathmind.com/wiki/neural-network

Since one layer as well as linearity are often insufficient excepting for the simplest of purposes, multiple layers and non linear activation functions are needed for accurate prediction, classification or clustering making multiple layers (i.e. deep learning) a necessity that can not be done away with.

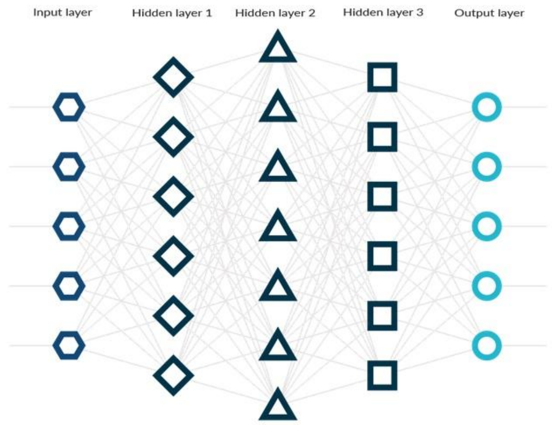


Figure 9: Schematic of a deep learning model

Source: https://www.ayima.com/blog/artificial-intelligence-neural-networks-deep-learning.html

Neural networks, like different regression models aim at estimating the weight parameters to minimize the error (or loss/cost). Starting with the artificial neural networks, the subsequent developments of convolutional neural networks and recurrent neural networks at doing this for while also optimizing computation load and improving accuracy of prediction, clustering or classification as the case be.

CHAID Model: Chi-Squared Automatic Interaction Detector model is an evolution of earlier AID and THAID (Theta AID) models. The analysis aims at building a predictive model and helps develop an understanding of how variables best merge to help estimate a given dependent variable. The popularity of CHAID model lies partly in its ability to allow use of nominal, ordinal as well as continuous data. CHAID creates multiple cross tabulations for each categorical predictor to arrive at the best outcome. A systematic approach is adopted to build the decision tree starting with the dependent variable forming the root node. The constituents of this root node are split into two or more categories based on mathematical indices to arrive at the best ordered distribution that explains the relationships between variables with the highest accuracy even though these variables may not necessarily be normally distributed.

Random Forrest: Random Forest is a supervised learning algorithm used more for the purpose of classification though it can well be used for regression too. The algorithm utilizes a voting mechanism among the largest possible (decision) trees built from data samples to arrive at the best solution for predicting the outcome. The large number of relatively uncorrelated trees arriving at the prediction through a voting mechanism greatly enhances the accuracy while also reducing chances of over-fitting.

Based on the analysis carried out and using the feature of IBM SPSS Modeler that combines the best of candidate models into a single ensemble model with the best results (accuracy) shows that Vapour Pressure and Relative humidity are the key parameters for predicting rainfall with a relative importance of 0.89 and 0.11 with Temperature not being a statistically significant factor as shown in the Figure 6 below.

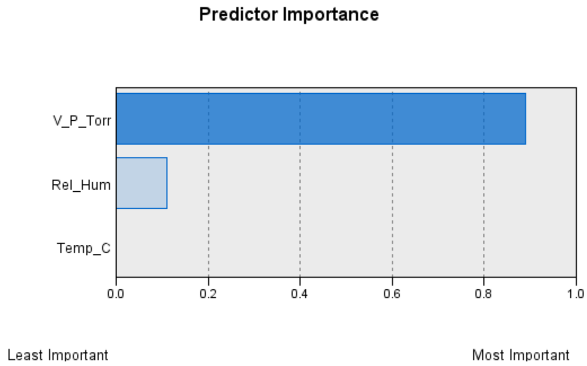


Figure 10: Predictor Importance

The low importance of temperature could be due to a high coreelation between the variables as seen by the correlation matrix below:

	Temp	Vap. Pres.	Rel.Hum
Temp	1		
Vap. Pr.	0.648847	1	
Rel.Hum	0.726633	0.515658	1

Figure 11: Correlation Matrix for predictors

To visually examine the quality of prediction, a plot of predicted (on X-Axis) versuus actual rainfall (on Y Axis) is shown in Figure 7 below:

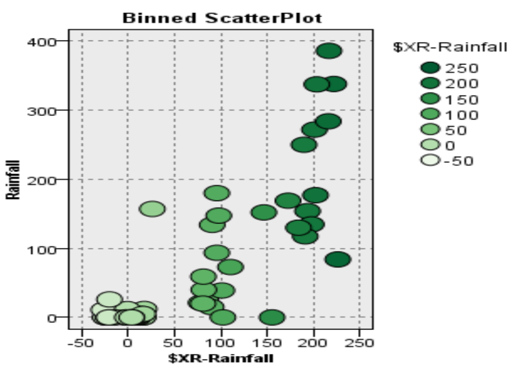


Figure 12: Predicted versus Actual Rainfall

The graphical representation indicates greater spread (error) in the middle as compared to the extremities.

Audit Report of the Model is shows in Figure 13 below:

Field	Graph	Measurement	Min	Max	Mean	Correlation
Rainfall		Continuous	0.000	386.000	49.755	--
Temp_C		Continuous	16.808	33.525	25.819	0.214
V_P_Torr		Continuous	6.334	29.861	16.128	0.681
Rel_Hum		Continuous	12.000	85.833	43.669	0.178
Partition		Nominal	--	--	--	--
$XR-Rainfall		Continuous	-26.023	225.909	50.002	0.834
$XRE-Rainfall		Continuous	0.925	26.782	9.100	0.526

VI. LIMITATIONS OF THE STUDY AND SCOPE FOR FURTHER STUDY

The key limitation of the study stems from the use of average values of the features (temperature, vapour pressure and relative humidity) for a given month. Since there can be significant variation over a month, the accuracy of estimation is expected to increase significantly if the model granularity is improved so that the time intervals considered are smaller, say, a period of a few days. Also the relatively small dataset suggests the need for a deeper study.

VII. CONCLUSION

The study suggests that machine leaning techniques make it possible to predict rainfall based on vapour pressure, relative humidity and temperature. This opens the possibility of significant variable reduction for predicting rainfall but care must be exercised since the duration of data was less than a decade. Applying the model over an extended period in multiple geographic locations would be required to add to its reliability.

REFERENCES

1. **Aftab, S., Ahmad, M, Hameed, N., Bashir, M.S., Ali, I and Nawaz, Z. (2018)**. Rainfall Prediction using Data Mining Techniques:A Systematic Literature Review. International Journal of Advanced Computer Science and Applications (IJACSA),Vol. 9, No. 5, 2018. Retrieved from https://thesai.org/Downloads/Volume9No5/Paper_18-Rainfall_Prediction_using_Data_Mining_Techniques.pdf.
2. **Hong W-C (2008)**. Rainfall forecasting by technological machine learning models. Applied Mathematics and Computation 200 (2008), 41–57.
3. **Janbandhu, C.C., Meshram, P.D. and Gedam, M.N. (2017)**. Modelling Rainfall Prediction Using Data Mining Method - A Bayesian Approach. IJFRCSCE, 3(11), November 2017, pp 472-474. Retrieved from: http://www.ijfrcsce.org/download/browse/Volume_3/November_17_Volume_3_Issue_11/1512370802_04-12-2017.pdf
4. **Katsaros, K. and Buettner, K.J.K. (1969)**. Influence of Rainfall on Temperature and Salinity of the Ocean Surface. Jouranl of Aplied Metrology, Vol. 8, February 1969, pp 15-18. Retrieved from: https://journals.ametsoc.org/doi/pdf/10.1175/1520-0450%281969%29008%3C0015%3AIOROTA%3E2.0.CO%3B2
5. **Mohd., R, Butt M.A. and Baba, M.Z. (2018)**. Comparative Study of Rainfall Prediction Modeling Techniques (A Case Study on Srinagar, J&K, India), Asian Journal of Computer Science and Technology, 7 (3), 2018, 13-19. Retrieved from http://www.trp.org.in/wp-content/uploads/2018/11/AJCST-Vol.7-No.3-Oct-Dec-2018-pp.13-19.pdf
6. **Oswal, N. (2019)**. Predicting Rainfall Using Machine Learning Techniques. Retrieved from https://arxiv.org/pdf/1910.13827.pdf.
7. **Singh, G. & Kumar, D. (2019)**. Hybrid Prediction Models for Rainfall Forecasting. 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence),392-396. Available at https://ieeexplore.ieee.org/document/8776885.
8. **Swapna, M. and Sudhakar, N. (2018)**. A Hybrid Model for Rainfall Prediction Using Both Parametrizedand Time Series Models. International Journal of Pure and Applied Mathematics, 119 (4), 2018 2018, pp 1549-1556. Retrieved from https://acadpubl.eu/hub/2018-119-14/articles/3/27.pdf
9. **Tarun, G.Bala Sai, Sriram, J.V., Sairam, K., Sreenivas, K.Teja, Santhi, M.V.B.T (2019)**. Rainfall prediction using Machine Learning Techniques, International Journal of Innovative Technology and Exploring Engineering (IJITEE)ISSN: 2278-3075, Volume-8 Issue-7, May, 2019. Retrieved from https://www.ijitee.org/wp-content/uploads/papers/v8i7/G5295058719.pdf.
10. Wikipedia page on Udaipur https://en.wikipedia.org/wiki/Udaipur

AUTHORS PROFILE

Prof (Dr.) Jitendra Shreemali is a graduate from IIT Madras with post graduate from IIM Bangalore. He is working as Professor of the Department of Computer Science and Engineering of Techno India NJR Institute of Technology Udaipur. He has worked in reputed companies in India & abroad for about a decade and half followed by about two decades of academic/ research/ training experience. He has taught a very wide variety of subjects/courses including operations management, research methodology, and data science besides others. His areas of work include data science, optimization, mathematical modeling and machine learning.
Email: jitendrapshreemali@gmail.com

Dr. Praveen Galav has done PhD in Physics in January 2012 and has completed a research project as a principle investigator. The research project was sanctioned to him under DST young scientist scheme by Department of Science and Technology, Government of India, New Delhi for a period of 3 years from August 2012 to July 2015. He has to tal 15 publications (13 International and 2 national to his credit) and is engaged in research related to effect of space weather events on the varion of ionospheric electron density as well as also involved the study of variation of atmospheric pollutant gases.
Email: praveen.galav@gmail.com

Mr. Gaurav Kumawat is working as an Assistant Professor in the Department of Computer Science and Engineering at Techno India NJR Institute of Technology, Udaipur. He has done M. Tech from Pacific University and has a total experience of 14 Years as a faculty in the CSE Department.
His expertise lies in the areas of C Programming, Object-Oriented Programming using C++, Core Java, Advance Java, Unix Programming, Web Programming, Operating Systems, and Computer Architecture etc. His Major Area of research includes Information Security, Programming Analysis and Data Analysis.
Email: Gaurav.kumawat@technonjr.org

Mr. Pankaj Chittora is an Assistant Professor in the Department of Computer Science and Engineering at Techno India NJR Institute of Technology Udaipur, He received his B. Tech from NIT Durgapur in 2010 and M. Tech in 2012. He has 10 years of experience as a faculty in the CSE department. He has delivered various subject including Data Structure, Algorithms Analysis, Theory of Computation, Compiler construction, and others. His major areas of research include data science, database, programming, information security and analysis of algorithms.
Email: pankaj.chittora@technonjr.org