International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Prediction of Humidity Depending on Temperature with Low Error Rate using regression model

Pinki Sagar, Prinima Gupta, Indu Kashyap

***Abstract:** This paper depicts prediction algorithm based on linear equations to forecast sequence trends in the future for time series data sets, prediction of humidity that dependent on temperature which is found in every 10 minutes. In this paper we analyze error rates during the prediction by using linear regression based algorithms. Two algorithms are discussed in this paper one for one dimensional stream data and second for two dimensional stream data. Both algorithms are applied on time series multivariate data sets and analysis of errors is to be done.*

***Keywords:** Data Mining, Prediction, Frequent Data sets, Linear Regression, Non Linear Regression.*

I. INTRODUCTION

In today information age, information leads to importance and hit, and gratitude to refined technologies of computers. Absurd amounts of information have been collected from various sources. Initially, with the start of computers, huge digital storage, collecting and storing all type of data, counting on the power of computers to help sort through this mixture of information. Consequently, *data mining* consists of collecting and managing data and extracts the relevant information; it also consists of analysis and prediction. Before applying any data mining techniques first the data should be preprocessed through the data integration, data cleaning, data transformation etc .in figure 1 it is explained first the data is collected from the various types of resources like repositories, data bases, and organization etc. in the next phase binning method is used to smooth the data and data cleansing is a method to remove or detect the inaccurate or corrupted data from the records of data sets. in data transformation data is consolidated in the appropriate forms for the data mining techniques, after the preprocessing of data patterns(in the form of rules, patterns, sequences etc.) are identified so mining strategies can be applied. In the end of data mining process results are identifies or analysis is to be done.

Manuscript published on 30 April 2019.
***** Correspondence Author (s)
Ms. Pinki Sagar, Assistant Professor, Department of CSE, faculty of engineering and technology, manav rachna Faridabad. (Haryana), India.
Dr. Prinima Gupta, Associate Professor, Department of CST, at Manav Rachna University, Faridabad. (Haryana), India.
Dr. Indu Kashyap, Department of CST, at Manav Rachna University, Faridabad. (Haryana), India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

A. Feature Selection: It is the process where important feature are selected on the basis of automatic or manually, theses feature highly contribute to the interested prediction **variable** or output. If data set have not relevant features that can cause the decreasing of the accuracy rate of algorithm. Which is designed on the basis of inaccurate features and noisy data sets.

B. significance of Feature Selection

It is very easy to work on machine learning – if work is done on noisy or corrupted data in result we must get noisy and disturbed results: rule of machine learning In the feature selection it is very important we must use useful features of data sets so that we can get efficient results from sample data sets.It is not necessary that all features will be used in the algorithms. We can help our algorithm by suckling in only significantcharacteristics that is really important. In the industrial applications it is very useful and in research area as well. Motives to use feature selection are:

- It helps to train the data very fast for thedata mining algorithm.
- It marks it easier to understand and reduces the complexity of a model.
- If right selection of feature is done then accuracy of model can be improved.
- It helps in reduces overfitting.

In the Next, here multiple methodologies and techniques are discussed so subset the feature data space could be used and help to the models for performing better and efficiently.

II . FILTER METHODS

It is not depend on any machine learning or data mining algorithms. Feature selection is done on the basis of many statistical results and result of correlation. If correlation between two attribute of data set is high they can be combined in to single attribute by doing summation.

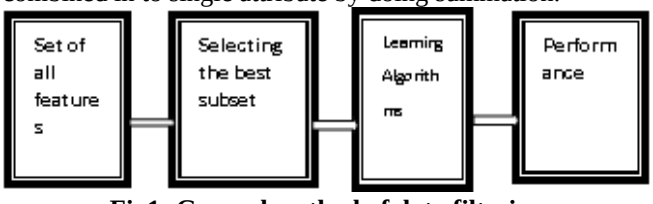


Fig1: General method of data filtering

2.1 **Pearson's Correlation**: in this method we measure quantitative linear dependencies between continuous data like X and Y. Its value lies between -1 to +1." Pearson's correlation" is defined as:

$pX, Y = cov(X, Y)/\sigma X \sigma Y$---------- (1)

Linear discriminate analysis (LDA) is used for findings of a linear combination of structures that describes or separates the data set in to two or more modules (or levels) of a categorical variable.

A. **ANOVA-Analysis of variance**: It is more similar to LDA except only for the one fact that it is operated more than one categorical autonomous characteristics and one continuous reliant *feature. It* generates a statistical outcomesthat means of many groups or categories are equal or not.

B. Chi-Square: It is applicable to the categorical data or it is defined as a statistical approach for the data like 0, 1 .to evaluate correlation and association between different variable of data set by using their frequency distribution. When chi square method is applied it should be clear that filter method which is applied on the dataset do not remove multi co linearity Coefficient of regression X on Y:

III. REGRESSION COFFICIENTS OF X ON Y

It is represented by the symbol bxy that measures the change in X for the unit change in Y". independent variable is responsible for changing in value of dependent variable. it can be represented as:

$$\boldsymbol{b_{xy} = r\frac{\sigma x}{\sigma y}}\text{----------(2)}$$

$$\sigma x = Standard\ Deviation\ of\ x$$

$$\sigma y = Standard\ Deviation\ of\ y$$

In equation 3 value is calculated by using the deviations which is taken from the actual means of variables: X and Y:

$$\mathbf{b_{xy} = \frac{\sum XY}{\sum yY}}\text{----------(3)}$$

If mean is assumed and standard deviation is obtained from assumed mean following equation is used:

$$\boldsymbol{b_{xy} = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dy * y - (\sum dy)(\sum dy)}}\ \text{----------(4}$$

IV. REGRESSION COFFICIENTS OF Y ON X

The symbol byx is used that measures the change in Y corresponding to the unit change in X. Representatively, it can be denoted as:

$$\boldsymbol{byx = r\frac{\sigma y}{\sigma x}}\text{----------(5)}$$

In equation 6, the deviations are taken from the actual means; the following formula is used:

$$\boldsymbol{b_{yx} = \frac{\sum XY}{\sum X * X}} - - - - - - - - - - (6)$$

The by b_{yx} can be calculated by using the following formula when the deviations are taken from the assumed means:

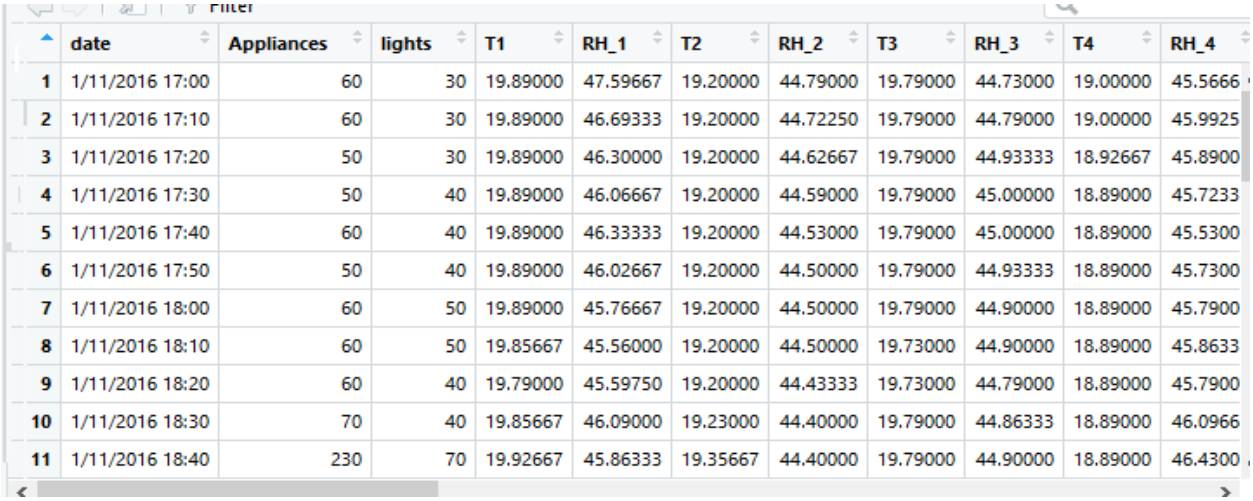
$$\boldsymbol{b_{yx} = \frac{N\sum dxdy - \sum dx \sum dy}{N\sum dx * x - (\sum dx)(\sum dx)}}\text{----------(7)}$$

The slope line of regression model is identified by regression coefficients. These also known as Regression slope coefficients. Regression Coefficient is also called as a slope coefficient because it determines the slope of the line i.e.in regression model two basic variables are: dependent variable and independent variable .changing in dependent variable can cause the changes in dependent variable the change in the independent variable for each and every unit of data.

V. TIME SERIES DATA SETS

A time series is a sequence of data points that is collected indexed (or listed or graphed) in a regular time order. Most frequently, a time series is a sequence that is obtained at successive equally spaced points in time. Time series forecasting is the use of a model to predict future values based on previously observed values. Quantities of data that is represented or traced the values collected by a variable over a period such as a month, quarter, or year. Time series data occurs wherever the same measurements are recorded on a regular basis. Multivariate Time Series Data. Often, the first step in creating a multiple time series model is to obtain data. There are two types of multiple time series data: Response data.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.



	date	Appliances	lights	T1	RH_1	T2	RH_2	T3	RH_3	T4	RH_4
1	1/11/2016 17:00	60	30	19.89000	47.59667	19.20000	44.79000	19.79000	44.73000	19.00000	45.5666
2	1/11/2016 17:10	60	30	19.89000	46.69333	19.20000	44.72250	19.79000	44.79000	19.00000	45.9925
3	1/11/2016 17:20	50	30	19.89000	46.30000	19.20000	44.62667	19.79000	44.93333	18.92667	45.8900
4	1/11/2016 17:30	50	40	19.89000	46.06667	19.20000	44.59000	19.79000	45.00000	18.89000	45.7233
5	1/11/2016 17:40	60	40	19.89000	46.33333	19.20000	44.53000	19.79000	45.00000	18.89000	45.5300
6	1/11/2016 17:50	50	40	19.89000	46.02667	19.20000	44.50000	19.79000	44.93333	18.89000	45.7300
7	1/11/2016 18:00	60	50	19.89000	45.76667	19.20000	44.50000	19.79000	44.90000	18.89000	45.7900
8	1/11/2016 18:10	60	50	19.85667	45.56000	19.20000	44.50000	19.73000	44.90000	18.89000	45.8633
9	1/11/2016 18:20	60	40	19.79000	45.59750	19.20000	44.43333	19.73000	44.79000	18.89000	45.7900
10	1/11/2016 18:30	70	40	19.85667	46.09000	19.23000	44.40000	19.79000	44.86333	18.89000	46.0966
11	1/11/2016 18:40	230	70	19.92667	45.86333	19.35667	44.40000	19.79000	44.90000	18.89000	46.4300

Fig2: Data set collected from UCI repository

Response data corresponds to in the multiple time series models defined in Types of Multivariate Time Series Models.

VI. DATA SET

The data set is collected at every 10 min for about 4 and half months. The monitoring home temperature and umidity conditions were done with a ZigBee wireless sensor network. T: temperature in Celsius,

T1-T7: Temperature recorded in different rooms in building details are in table

Table 1: detail of temperature attributes of a house

Attribute	Detail
T1	Temperature of kitchen area
T2	Temperature of living room
T3	Temperature of laundry room area
T4	Temperature of office room
T5	Temperature of bathroom
T6	Temperature outside the building (north side)
T7	Temperature of ironing room

R_h1 to R_h7: humidity based on temperature from different rooms in building,

Table 2: Detail of humidity attributes of a house

Attribute	Detail
R_h1	Humidity of kitchen area
R_h2	Humidity of living room
R_h3	Humidity of laundry room area
R_h4	Humidity of office room
R_h5	Humidity of bathroom
R_h6	Humidity outside the building (north side)
R_h7	Humidity of ironing room

VII. REGRESSION ALGORITHM FOR TWO DIMENSIONAL STREAM DATA SET

Step1: For specific sliding window ids are identified on which data sequence appeared.

Step2: Findings of the support for the sequence of stream data

Support = number of ids /total number of ids.

Step3: Dependent variable: Support for each sliding window independent variable: Ending time of sliding window

Step4: Apply the Linear regression method and Predicted support is calculated.

Linear Model is y= a+bX

Attributes are (ts, ∑tf, ∑f, ∑f2)

Stt = ∑t2-(∑t) 2 /n

Sff = ∑f2-(∑f) 2/n

Stf = ∑tf-((∑t)(∑f)/n)

b= Stf/ Stt

a= (∑f/n)-b*(∑t/n)

VIII. REGRESSION ALGORITHM FOR ONE DIMENSIONAL STREAM DATA SET

y=b0+b1X

b0= mean of y-b1 (mean of x)

b1=∑(x-mean of x)(y-mean of y)/∑(x-mean of x)2

y=b0+b1X

b0= mean of y-b1 (mean of x)

b1=∑(x-mean of x)(y-mean of y)/∑(x-mean of x)2

Yi = b0 + b1 xi +error

Correlation with temperature and humidity:

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

```
> cor(x = data$T1, y = data$T2)
[1] 0.9955426
> cor(x = data$T3, y = data$T4)
[1] 0.6879202
> cor(x = data$T4, y = data$T2)
[1] 0.7750046
> cor(x = data$T1, y = data$T8)
[1] 0.9546742
> cor(x=data$RH_1,y=data$RH_2)
[1] 0.8461243
> cor(x=data$RH_5,y=data$RH_6)
[1] 0.1494702
> cor(x = data$T1, y = data$T2)
[1] 0.9955426
> cor(x = data$T3, y = data$T4)
[1] 0.6879202
> cor(x = data$T4, y = data$T2)
[1] 0.7750046
> cor(x = data$T1, y = data$T8)
[1] 0.9546742
> cor(x=data$RH_1,y=data$RH_2)
[1] 0.8461243
> cor(x=data$RH_5,y=data$RH_6)
[1] 0.1494702
```

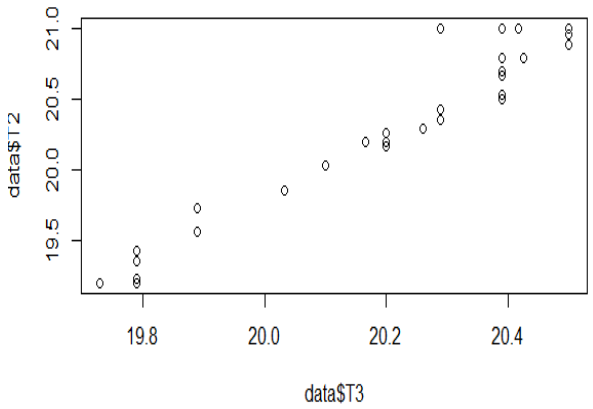


Fig 3: Graph for Positive correlation

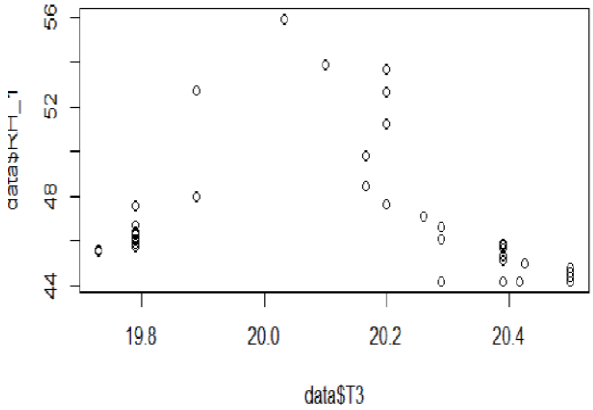


Fig 4: Graph with negative correlation

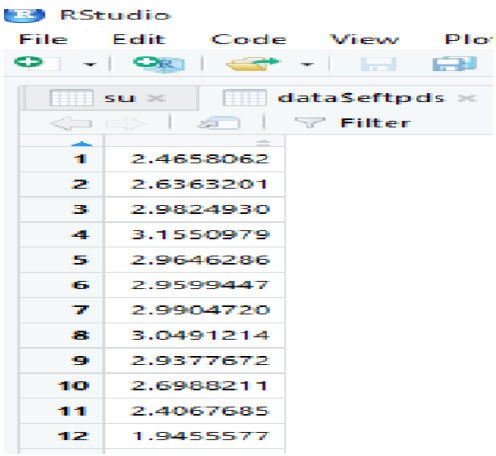


Table 3: Error generated through prediction method for two dimensional stream data

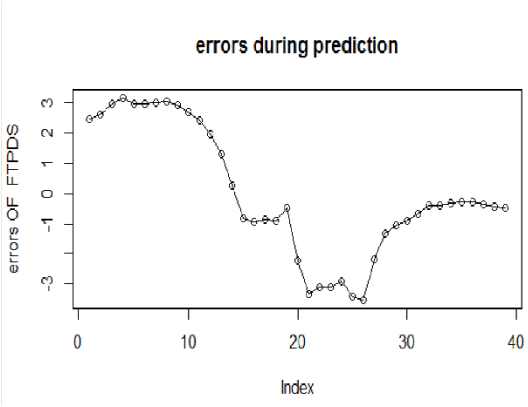
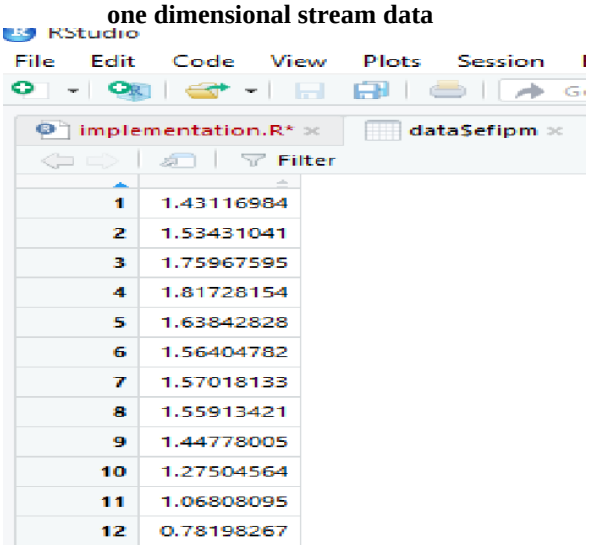


Fig5: Error rate graph during the prediction using FTPDS

Table 4: Error generated through prediction method for one dimensional stream data



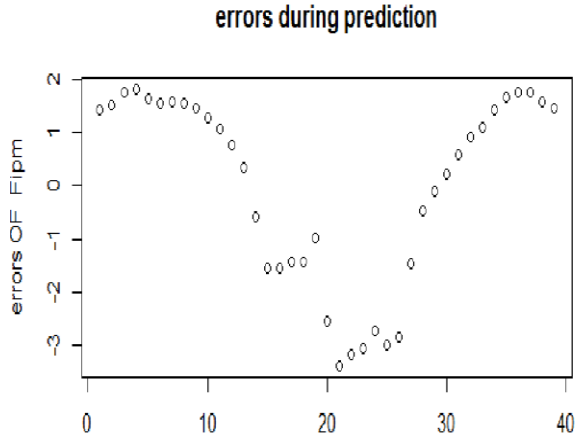


Fig 6: Error rate graph during the prediction using FIPM

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

IX. RESULT AND CONCLUSION

Regression Algorithms are implemented in r studio 3.3.2 on multivariate data sets which is collected on daily basis in every 10 minutes. x axis is error rate during the prediction of humidity on the basis of temperature of different rooms of a building. By using algorithm for two dimensional stream data, error rate during the prediction is higher in comparison of algorithm for one dimensional stream data. For multivariate data set algorithm for two dimensional stream data gives high error rate during prediction in comparison of one prediction algorithm for one dimensional stream data. Prediction of two dimensional data give the better result in continuous data are frequent accrued data. It will help in prediction of network congestion during transferring of data, weather forecasting etc.

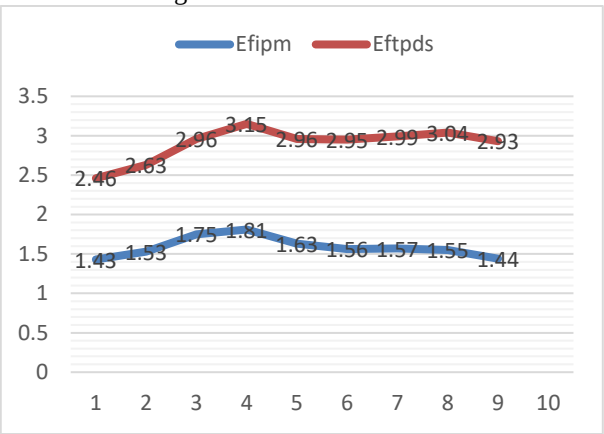


Fig 7: analysis of error rates during prediction

REFERENCES

1. XiaoboZhou,Kuang-Yu Liua, b,Stephen T.C. Wonga, "Biomedical Machine Learning Cancer classification and prediction using logistic regression with Bayesian gene selection,"Volume 37, Issue 4, August 2004.
2. FENG ZHAO,QING HUA LI, " A plane regression-based sequence forecast algorithm for stream data", Published in: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on18-21 Aug. 2005 IEEE.
3. Duck Jin Chai EunHee Kim "Prediction of Frequent Items to One Dimensional Stream Data Fifth International Conference on Computational Science and Applications" 0-7695-2945-3/07 © 2007 IEEE.
4. Andreas Hecker, Thomas K¨urner," Application of Classification and Regression" Trees for Paging Traffic Prediction in LAC Planning 1550-2252/$25.00 ©2007 IEEE.
5. Qingshan Ni1, Zhengzhi Wang1, Qingjuan Han2, Gangguo Li1, Xiaomin Wang1, GuangyunWang,"Using logistic regression method to predict protein function from protein-protein interaction" data 978-1-4244-2902-8/09/$25.00 ©2009 IEEE.
6. Alberto Landi, Paolo Piaggi Artificial Neural Networks for Nonlinear Regression and Classification 978-1-4244-8136-1/10/$26.00_c 2010 IEEE..
7. Krisztian Buza, AlexandrosNanopoulos, Lars Schmidt-Time-Series Classification based on Individualised Error Prediction Thieme 2010 13th IEEE International Conference on Coputational Science and Engineering 978-0-7695-4323-9/10 $26.00 © 2010 IEEE.
8. Umesh Kumar Pandey, Saurabh Pal Data Mining : A prediction of performer or underperformer using classification (IJCSIT) International Journal of Computer Science and Information Technology, Vol. 2(2), 2011.
9. NidhiBhatla KiranJyoti An Analysis of Heart Disease Prediction using Different DataMining Techniques International Journal of Engineering Research & Technology (IJERT) vol. 1 Issue 8, October – 2012.
10. M.A.NisharaBanuGomathy Disease Forecasting System Using Data Mining Methods 2014 International Conference on Intelligent Computing Applications 978-1-4799-3966-4/14 $31.00 © 2014 IEEE.
11. Aida Mustapha et al. International Journal of Engineering and Technology (IJET) A Regression Approach for Forecasting Vendor Revenue in Telecommunication Industries Vol 6 No 6 Dec 2014-Jan 2015.
12. GustiAyuPutriSaptawati, GustiNgurah Mega Nata Knowledge Discovery on Drilling Data to Predict Potential Gold Deposit, 2015 International Conference on Data and Software Engineering, 978-1-4673-8430-8/15/$31.00 ©2015 IEEE.
13. Swati Gupta, A Multiple Regression Technique in Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 126 – No.5, September 2015
14. SheelaGole Frequent Itemset Mining for Big Data in social media using ClustBigFIM algorithm International Conference on Pervasive Computing (ICPC) -1-4799-6272- 3/15/$31.00(c)2015 IEEE
15. Swati Gupta, A Regression Modeling Technique on Data Mining, International Journal of Computer Applications (0975 – 8887) Volume 116 – No. 9, April 2015
16. Ilayaraja M*, MeyyappanT , Efficient Data Mining Method to Predict the Risk of Heart Diseases through Frequent Itemsets4t h International Conference on Eco-friendly Computing and Communication Systems, ICECCS 2015 1877-0509 © 2015. Published byElsevier B.V.
17. JaeKwonBaeandJinhwaKim,"A Personal Credit Rating Prediction Model Using DataMining in Smart Ubiquitous Environments, RESEARCH ARTICLE PUBLISHED IN FEBRUARY,2015
18. Farhan Khan, Dariush Kari, IlyasAlperKaratepe, and Suleyman S. Kozat "UniversalNonlinear Regression on High Dimensional Data Using Adaptive Hierarchical Trees 2016 IEEE.

AUTHORS PROFILE

First Author: Ms. Pinki Sagar Working as an assistant Professor in department of CSE , faculty of engineering and technology , manav rachna international institute of research and studies. And research scholar(PH.D) of CST department of manav rachna university. Ms. Pinki Sagar has published 15 papers in National & International level journals and presented 7 papers in National & International Conferences and also attended various workshops & FDPs.

Second Author :Dr. Prinima Gupta is presently working as Associate Professor in CST Department, at Manav Rachna University, Faridabad. She has over 13.6 years of experience including academics & research.She has done her Ph D in the area of Ad-hoc Networks in 2013. She has studied the performance of various routing protocols of MANET using different simulation tools. She has also completed projects for implementing Information security at various level using different security algorithms. Dr. Prinima has published 18 papers in National & International level journals and presented 13 papers in National & International Conferences and also attended various workshops & FDPs. She has also been associated with many conferences as a reviewer of the paper and a member of technical committee.

Third Author Dr. Indu Kashyap has more than eleven years of experience in teaching. She has done M.Tech and Ph.D in Computer Science and Engineering. She has guided many M. Tech projects, Dissertations and Ph.Dscholars. She has several publications to her credit in various leading International and National Journals in the various areas like, Wireless Networking, Databases, Cloud Computing etc.Currently, she is working as an Associate Professor in the Faculty of Engineering and Technology (FET), MRIIRS and also acting as a Ph.D coordinator for Engineering Programme. She is a member of many technical committees

© Copyright: All rights reserved.