QALAAI ZANIST JOURNAL
A Scientific Quarterly Refereed Journal Issued by Lebanese French University – Erbil – Kurdistan – Iraq
Vol. (2), No. (2), April 2017
Speical Issue : The 1st International Conference on Information Technology (ICoIT'17)
ISSN 2518-6566 (Online) - ISSN 2518-6558 (Print)

The Estimation of Wind Velocity Using Data Mining Techniques

Sattar Nabee Rasool
Software Engineering, Technology Faculty, Firat University-Turkey
star81sat@gmail.com
Ahmet Koca
Mechatronics Engineering, Technology Faculty, Firat University-Turkey
akoca@firat.edu.tr
Karwan Hussein Qader
School of Computing, Faculty of Technology, University of Portsmouth- United Kingdom
Karwan.qader@port.ac.uk

ARTICLE INFO

Article History:

Received: 10 March 2017

Accepted: 1 April 2017

Published: 10 April 2017

DOI:

10.25212/lfu.qzj.2.2.44

Keywords: *Climate parameter estimation, data mining, predictive modeling, clustering, classification*

ABSTRACT

Estimation of wind velocity in real time is very essential as it can provide valuable information to people of different domains such as agriculture, aviation and tourism to mention few. Since climate data is growing exponentially it is hard to analyze it manually. Therefore, machine-learning techniques such as unsupervised and supervised learning methods are used to mine voluminous data and discover valuable knowledge. Predictive modeling in data mining is required to estimate climate parameters. In this paper, we proposed a framework that exploits data mining techniques such as J48, KNN, Neural Networks, SVM and Linear Regression. The framework takes climate dataset as input, completes training phase and makes different models using data mining algorithms. Finally, it ends by exploiting linear regression, which models the relationship between a dependent variable and an exploratory variable. The framework results in estimating wind velocity and finding prediction error rate. A prototype application is built based on Weka, which is used to demonstrate proof of the concept. The empirical results applied on all data are obtained from the Turkish Government Meteorology Services for summer months of 2013. It revealed that the proposed framework is useful to have a predictive model with respect to estimation of climate parameters.

INTRODUCTION

Data mining techniques became indispensable in the contemporary business world where enterprises need to analyze huge amount of data. Data mining provides both predictive and descriptive models. The predictive models include techniques pertaining to classification, regression, time series analysis and prediction. Descriptive models include clustering, summarization, association rules, and sequence discovery. Clustering is an unsupervised learning method used to group similar objects while classification is known as supervised learning method, which needs training samples for prediction of class labels for some test objects. Similarity measures like Euclidean distance are widely used to know similarity between two objects in the given dataset. Predictive methods are used in many real time applications. For instance weather forecasting, forecasting of stock values, and prediction of possible increase or decrease of values of certain items to mention few.

In this paper different data mining algorithms are analyzed, trained, and models are created to have utility in wind velocity of given dataset. After experimenting with different models, five different datamining tools is used in order to predict wind velocity. An application is built based on Weka tool in order to demonstrate proof of the concept. The remainder of the paper is structured as follows. Section 2 reviews related literature. Section 3 presents the proposed framework for data mining. Section 4 presents experimental results while section 5 presents conclusions and recommendations for future work.

RELATED WORKS

Aziz and Yosof [1] proposed a classification model to analyze graduates employment. They employed different algorithms like J48, k-Nearest Neighbor (KNN), Multilayer perception, Logistic Regression and Naive Bayes. They found that logistic regression resulted in highest accuracy 95.2%. Balle and Lisboa [2] focused on machine learning methods and models of data mining. They include monograms, rule induction, graphical models, and data visualization. Minh, Hue, Dzung, and Toan [3] made a simulation study of equity valuation model. Langone et al. [4] proposed a machine learning method known as Least Squares Support Vector Machine (LS-SVM) for fault detection online with industrial machine data. It is supervised learning method, which helped to know early detection and classification of faults.

Claesen et al. [5] proposed a supervised learning model using SVM. The model makes use of positives and unlabeled objects. It actually makes use of an ensemble of SVM models in order to have an intuitive approach for appropriate learning model. It has features like supervised learning, positive and unlabeled learning, and learning with false positives. Frandi et al. [6] propose another SVM classification approach along with an algorithm for large-scale benchmark data classification. Alaiz et al. [7] proposed a graph based classification method for robust classification of data presented in the form of graphs. Chae et al. [8] proposed a data mining approach known as feature selection for intrusion detection using NSL-KDD dataset. They used the concept of information gain and gain ratio. Chen et al. [9] emphasized the need for considering big data and big data eco-system for many real time enterprise applications.

Huang et al. [10] proposed an SVM based classifier known as pin-SVM for finding pinball loss in SVM. Mehrkanoon et al. [11] proposed a semi-supervised algorithm that is based on kernel spectral clustering. Kernel here holds prior knowledge about labels. Drumetz et al. [12] focused on dimensionality estimation algorithms for finding intrinsic dimensionality in datasets.

Olaiya and Adeyemo [13] investigated the utility of data mining techniques for predicting weather information like wind speed, evaporation, rainfall and maximum temperature. They used two algorithms namely decision tree and artificial neural network for achieving this. They found the usefulness of data mining techniques for weather forecasting and study of climate changes. Krishna [14] reviewed many weather forecasting models available. They include models based on classification techniques, clustering techniques, and artificial neural networks. Schmidhuber [15] made a review of data mining techniques that are used for deep supervised and unsupervised learning. Especially they focused on neural networks. Cortest et al. [16] proposed a framework known as AdaNet for analyzing artificial neural networks. The framework followed adaptive structural learning. Siddiqui and Muhammad [17] performed clustering based data mining operations on KDD Cup 99 dataset. They used k-Means clustering algorithm provided by Oracle Data Mining tool.

Schubert et al. [18] proposed a methodology for clustering uncertain data. Uncertain data is the data which has some noise and the data is not precise. They used a general framework to cater to the needs of analysis of uncertain data. They named the framework as ELKI which has many implementations of algorithms that made use of distance measures, pruning techniques, visualization components and evaluation measures. Agrawal and Gupta [19] explored usage of C4.5 algorithm with certain optimization for healthcare domain with additional rules. They studied the difference between classification functionality of original C4.5 and an optimized C4.5 in terms of efficiency and complexity. Denoeux et al. [20] explored a rule known as evidential K-nearest neighbor rule for making a clustering procedure. Their procedure is named as EK-NNclus, which takes scale parameter and k of neighbors. In addition, it does not need the number of clusters to be generated. It was able to get both model-based and density-based procedures with having knowledge of number of clusters priori. Sanghavi et al. [21] explored three kinds of algorithms related to feature selection. They are known as Filter, Wrapper and Embedded. They built logistic regression algorithm in order to identify disease with medical data mining. In this paper a framework is proposed to explore data mining algorithms and produce a prediction model ultimately using linear regression technique. The proposed model shows that such climate can be predicted using the data mining techniques considering the error rate calculations. Therefore, this work filled those gaps, which it has been conducted to gather fifth different tools into a model includes cluster, classification and regression techniques. The proposed model shows the comparison evaluation results that obtain by the model execution. It is simply give clear results of all used techniques and guide the user, which one will be the most suitable tool according to the nature of work.

PROPOSED METHODOLOGY

In this paper, we proposed a methodology for the exploration of data mining algorithms for prediction of wind velocity parameter of climate dataset. In other words, the methodology throws light into building various models using different data mining techniques. The methodology takes climate dataset as input and performs training phase, so as to get training data ready for exploration of different classification algorithms. The data mining algorithms used in the exploration of the models include kNN, Neural Networks, J48, SVM, and linear regression. The methodology is illustrated in Figure 1.

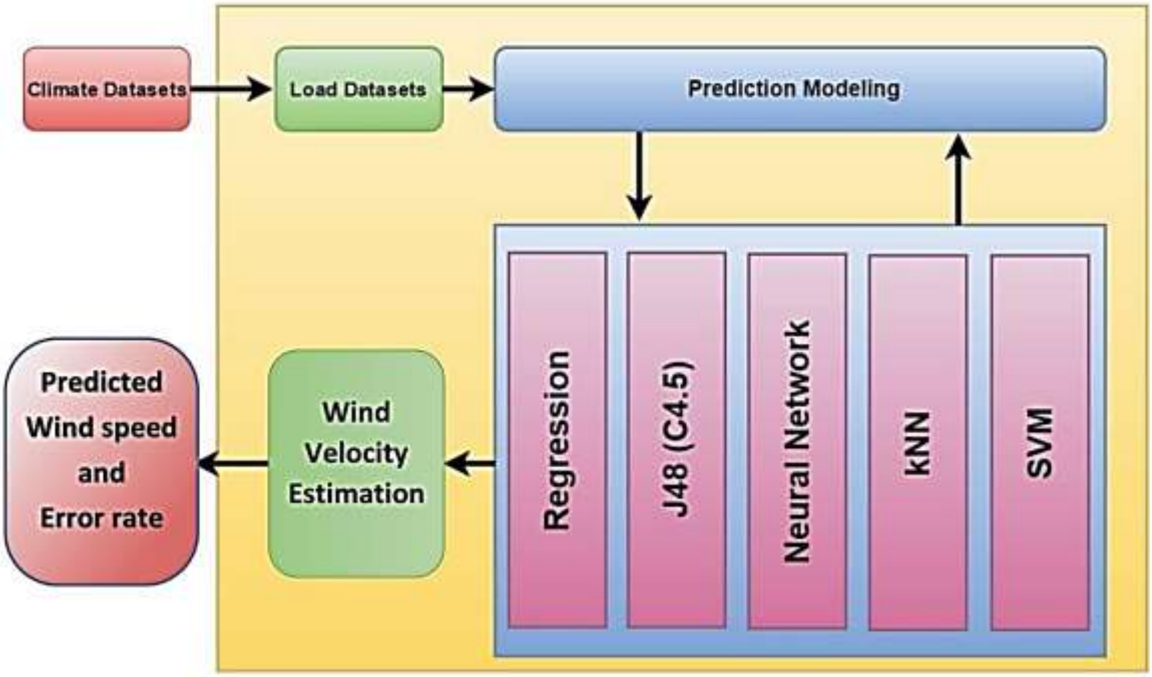

FIGURE 1. Proposed methodology

KNN is meant for pattern recognition method meant for regression and classification. Neural networks are the networks used in computer science that model human brain and nervous system. It is a computational approach, which is based on collection of artificial neurons that mimic how biological neurons work. J48 is a decision tree algorithm based on ID3. SVM is a supervised learning model which can have associated learning algorithms for both regression and classification tasks.

Linear regression is a data mining approach to modeling the relationship between a dependent variable and one or more independent variables. Simple linear regression makes use of one independent variable. A linear model is built using linear regression approach which models relationship between model parameters linear predictor functions.

In this paper, the kNN, J48, SVM, neural networks and linear regression model are used as a prediction models. The individual methods are employed to have wind velocity prediction. Then the results pertaining to actual wind velocity and predicted wind velocity are presented for all algorithms in the next section. Error rate is computed for all algorithms to know prediction accuracy of different algorithms.

EXPERIMENTAL RESULTS

A Dataset that contains weather data with attributes is taken as a case study for this research work. First column of the file is station number, every station represents a Turkish city which is five stations. These cities chosen by their geographical position and include INEBOLU, SINOP, AMASRA, SILE and AKCAKOCA. Every station place is near the sea. So that their wind potential is higher than other places. All data are obtained from the Turkish Government Meteorology Services for summer months of 2013.

The wind potential of a place is related with geographical position, humidity, atmospheric pressure and the ambient temperature. All of these are datasets parameters. And it has 456 instances covering data of five stations for summer months. An excerpt of the dataset is shown in Table 1. Weka is the environment used to explore a climate dataset that is very useful and the mining of such data provides useful information to various domains in the world. Weka tool is used to exploit all existing methods. Apart from Weka, the environment includes JDK 1.7 (Java programming language), Net Beans 7.2 (Integrated Development Environment) are used in Windows platform. JDK is used to have basic tools like compiler, application launcher and Java language support. Net Beans is used to have rapid application development as it supports developing applications faster. Weka is integrated with Net Beans in order to reuse existing methods and other support functionality like loading files, viewing data and results. Swing API in JDK is used to have Graphical User Interface (GUI) while java.io package is used to work with input and output files. Different kinds of data mining techniques as presented in Figure 1 are used to perform mining on given dataset. The dataset is related to climate data that were captured in a one of the region in Turkey within first fifteenth days of the June. The attribute of pressure, Humidity and temperature and actual wind speed resented in the datasets. To predict the wind velocity using all datamining techniques the data inputted into the proposed model. In supervised simulation the predefined classes achieved through a training phases of the existing datasets. In NN phase, the backward propagation used based on excerpt of dataset is presented in Table 1 (except the last column which is the predicted wind velocity using Linear Regression), which is used as input to the proposed framework.

As shown in Table 1, it is evident that the input dataset contains different attributes like station number month, day, and pressure, humidity, and temperature. The proposed framework is applied to have training samples and different models. All these experiments are done using Weka, which is one of the widely used data mining tools.

TABLE 1: An excerpt from generated output of Regression

Station Number	Month	Day	Pressure (hPa)	Humidity (%)	Temperature (ºC)	Wind Velocity (m/s)
1	6	1	1008.7	67.7	19.7	19.52310330013237
1	6	2	1005	80.3	19.6	19.60674819378439
1	6	3	1010.1	69.3	17.3	19.69039308743642
1	6	4	1011.1	72	19.4	19.77403798108844
1	6	5	1013.4	82.7	17.4	19.85768287474047
1	6	6	1014.6	80.7	18.8	19.941327768392494
1	6	7	1013.6	72	19.5	20.02497266204452
1	6	8	1010.1	77.7	20.4	20.108617555696544
1	6	9	1012.3	81.3	19.2	20.19226244934857
1	6	10	1012.9	77.7	19.2	20.275907343000593
1	6	11	1008.4	75.7	20.6	20.35955223665262
1	6	12	1005.3	80.3	21.2	20.443197130304643
1	6	13	1006.5	80.3	21.9	20.526842023956668
1	6	14	1009.8	88.7	20	20.610486917608696

1	6	15	1010.7	85.3	20.8	20.694131811260718

As shown in Table 1, the outcome of the proposed methodology is presented in the last column. It is wind velocity, which is predicted using linear regression model. This kind of prediction models has impact on many applications in the real world such as weather forecasting, prediction of dependent variables for making well-informed decisions. Same scenario applied on the datasets based on all other techniques and then Error rate is computed for all by comparing predicted wind velocity and actual wind velocity.

TABLE 2: Error rate for Regression method

Station Number	Month	Day	Actual Wind Velocity	Predicated Wind Velocity (m/s)	Error Rate
1	6	1	17.4	19.52310330013237	12.2017
1	6	2	15.2	19.60674819378439	28.9918
1	6	3	15.5	19.69039308743642	27.0348
1	6	4	15.5	19.77403798108844	27.5744
1	6	5	15.8	19.85768287474047	25.6815
1	6	6	18.2	19.941327768392494	9.56773
1	6	7	16.6	20.02497266204452	20.6324
1	6	8	15.5	20.108617555696544	29.733
1	6	9	19.4	20.19226244934857	4.08383
1	6	10	19.1	20.275907343000593	6.15658
1	6	11	18.4	20.35955223665262	10.6497
1	6	12	17.2	20.443197130304643	18.8558
1	6	13	15.9	20.526842023956668	29.0996

1	6	14	16.2	20.610486917608696	27.2252
1	6	15	18.4	20.694131811260718	12.4681

As presented in Table 2, the results show the actual wind velocity and predicted wind velocity. Error rate is computed as follows.

Error Rate= (Predicted Wind Velocity-Actual Wind Velocity)*100/Predicted Wind Velocity (or)

Error Rate= (Actual Wind Velocity - Predicted Wind Velocity)*100/Predicted Wind Velocity

Error rate is presented in the last column of the Table 2, which reflects the difference between actual wind velocity and predicted wind velocity. In the same fashion, the results of other algorithms are presented as follows.

TABLE 3: Error rate for kNN

Station Number	Month	Day	Actual Wind Velocity	Predicated Wind Velocity (m/s)	Error Rate
1	6	1	17.4	17.566	0.95402
1	6	2	15.2	17.725	16.6118
1	6	3	15.5	18.112	16.8516
1	6	4	15.5	18.345	18.3548
1	6	5	15.8	18.365	16.2342
1	6	6	18.2	18.75	3.02198
1	6	7	16.6	19.12	15.1807
1	6	8	15.5	17.56	13.2903
1	6	9	19.4	18.25	5.9278
1	6	10	19.1	18.28	4.2931

1	6	11	18.4	18.56	0.86957
1	6	12	17.2	18.85	9.59302
1	6	13	15.9	18.88	18.7421
1	6	14	16.2	18.54	14.4444
1	6	15	18.4	18.56	0.86957

TABLE 4: Error rate for J48

Station Number	Month	Day	Actual Wind Velocity	Predicated Wind Velocity (m/s)	Error Rate
1	6	1	17.4	15.466	11.11494
1	6	2	15.2	15.723	3.44079
1	6	3	15.5	16.12	4
1	6	4	15.5	16.35	5.48387
1	6	5	15.8	16.58	4.93671
1	6	6	18.2	15.95	12.36264
1	6	7	16.6	17.26	3.9759
1	6	8	15.5	16.58	6.96774
1	6	9	19.4	16.95	12.62887
1	6	10	19.1	17.56	8.062827
1	6	11	18.4	17.85	2.98913
1	6	12	17.2	17.86	3.83721
1	6	13	15.9	18.12	13.9623
1	6	14	16.2	18.2	12.3457
1	6	15	18.4	18.15	1.358696

TABLE 5: Error rate for SVM

Station Number	Month	Day	Actual Wind Velocity	Predicated Wind Velocity (m/s)	Error Rate

1	6	1	17.4	13.46	22.64368
1	6	2	15.2	13.72	9.736842
1	6	3	15.5	14.25	8.064516
1	6	4	15.5	14.35	7.419355
1	6	5	15.8	14.58	7.721519
1	6	6	18.2	14.35	21.15385
1	6	7	16.6	15.25	8.13253
1	6	8	15.5	14.68	5.290323
1	6	9	19.4	15.28	21.23711
1	6	10	19.1	15.45	19.10995
1	6	11	18.4	15.85	13.8587
1	6	12	17.2	16.12	6.27907
1	6	13	15.9	16.05	0.9434
1	6	14	16.2	16.35	0.92593
1	6	15	18.4	16.15	12.22826

TABLE 6: Error rate for Neural Networks (BPNN)

Station Number	Month	Day	Actual Wind Velocity	Predicated Wind Velocity (m/s)	Error Rate
1	6	1	17.4	13.56	22.06897
1	6	2	15.2	12.78	15.92105
1	6	3	15.5	12.25	20.96774
1	6	4	15.5	12.35	20.32258
1	6	5	15.8	12.58	20.37975
1	6	6	18.2	14.12	22.41758
1	6	7	16.6	13.54	18.43373
1	6	8	15.5	13.24	14.58065
1	6	9	19.4	15.86	18.24742
1	6	10	19.1	14.75	22.77487
1	6	11	18.4	14.76	19.78261
1	6	12	17.2	13.89	19.24419
1	6	13	15.9	13.95	12.26415

1	6	14	16.2	13.98	13.7037
1	6	15	18.4	14.86	19.23913

As shown in Table 3 through Table 6, the results of different algorithms such as kNN, J48, SVM and neural networks are presented along with the error rate in wind velocity prediction.

TABLE 7: Error rate for all methods

Method Name	**Average Error Rate**
KNN	10.349264
J48	7.1644882
SVM	10.98300233
Neural Networks	18.68987467
Linear Regression	19.33040933

In addition, Table 7 shows the average error rate of all instances for each method is computed. The results revealed the performance of different methods for predicting wind velocity. The results stated that J48 obtained the lowest error rate while Linear Regression got the biggest average of the error rate. Likewise, the error rate of other used techniques are shown in the table 7.

CONCLUSIONS AND FUTURE WORK

This paper focuses on data mining techniques used to build different modes using training samples. The paper explored different data mining techniques using Weka environment. The techniques employed include J48, kNN, Neural Networks, SVM and Linear Regression. Climate dataset is considered as input. A framework is proposed to have data mining applied on the given dataset. Different algorithms aforementioned are applied for making different models. The framework functionality ends with application of linear regression model, which predicts wind velocity. The dependent variable considered as wind velocity, which is empty in the given dataset. A prototype application is built to demonstrate the proof of the concept. The experimental results revealed the utility of the proposed framework in predicting wind velocity. The error rate of Linear Regression is between 4.08383 and 29.0996. The error rate of KNN is between 0.86 957 to 18.7421. The error rate of J48 is between 1.358696 and 13.9623. The Error rate of SVM is between 0.92593 and 22.64368 while the error rate of neural networks is between 12.26415 and 22.77487. J48 shows low average error rate. In future, this research can be extended to ensemble methods to have more quality predictions for datasets from different domains.

REFERENCES

[1] Mohd Tajul Rizal Ab Aziz and Yuhanis Yusof , “Graduates Employment Classification using Data Mining Approach", *ICAST*, pp. 1-8, 2016.

[2] Vanya Van Belle and Paulo Lisboa. “ Research directions in interpretable machine learning models”. *ESANN*, pp. 533-541, 2013.

[3] Nguyen Dang Minh and Nguyen Thi Minh Hue , ”Equity valuation model of vietnamese companies in a foreign securities market a simulation approach, *IEEE*, vol. 66, pp. 1155-11, 2012.

[4] RoccoLangone , CarlosAlzate , BartDeKetelaere , JonasVlasselaer , WannesMeert and JohaA.K.Suykens,“LS-SVMbasedspectralclusteringandregressionforpredicting maintenanceofindustrialmachines”, *Elsevier Ltd*, vol. 37 , pp. 268-278, 2014.

[5] Marc Claesen, Frank De Smet, Johan A. K. Suykens and Bart De Moor, “Arobust ensemble approach to learn from positive and unlabeled data using svm base models. *neurocomputing, ”SI on Advances in Learning with Label Noise*, pp. 1-34, 2014.

[6] Emanuele Frandi, Ricardo Ñanculef and Johan A. K. Suykens, ”A partan-accelerated frank-wolfe algorithm for large-scale svm classification”, pp. 1-19, 2015.

[7] Carlos M. Alaz, Micha • el Fanuel and Johan A. K. Suykens. “Robust classification of graph-based data” , pp. 1-8, 2016.

[8] Hee-su Chae, Byung-oh Jo, Sang-Hyun Choi and Twae-kyung Park, “ Feature selection for intrusion detection using nsl-kdd”, *isbn*, pp. 184-187, 2013.

[9] Min Chen ,” Shiwen Mao and Yunhao Liu. big data,” *A Survey.Springer*, vol. 19 , pp. 171-209, 2014.

[10] Xiaolin Huang, Lei Shi, and Johan A. K. Suykens. “ Solution path for pin-svm classifiers with positive and negative values,” pp. 1-10, 2012.

[11] Siamak Mehrkanoon and Johan A. K. Suykens. “Non-parallel semi-supervised classification based on kernel spectral clustering,” pp. 1-8, 2014.

[12] Lucas Drumetz, Miguel Angel Veganzones, Ruben Marrero, Guillaume Tochon, Mauro Dalla Mura, Giorgio Licciardi, Christian Jutten and Jocelyn Chanussot, “Hyperspectral Local intrinsic dimensionality”, *HAl*, pp. 1-16. 2016.

[13] Folorunsho Olaiya and Adesesan Barnabas Adeyemo . (2012). Application of Data Mining Techniques in Weather Prediction and Climate Change Studies . Information Engineering and Electronic Business, p51-59.

[14] G.Vamsi Krishna (2015) A Review of Weather Forecasting Models-Based on Data Mining and Artificial Neural Networks 06 , p214-22.

[15] Jurgen Schmidhuber,“Deep learning in neural networks: an overview”, pp. 1-88, 2014.

[16] Corinna Cortes, Xavi Gonzalvo and Vitaly Kuznetsov, ” Adanet: adaptive structural learning of artificial neural networks”, pp. 1-20, 2016.

[17] Mohammad Khubeb Siddiqui and Shams Naahid, ” Analysis of kdd cup 99 dataset using clustering based data mining”**,** *International Journal of Database Theory and Application*, vol. 6, (1-5), pp. 23-34, 2013.

[18] Erich Schubert, Alexander Koos and Tobias Emrich ,” A framework for clustering uncertain data” , pp. 1976-1979, 2012.

[19] Gaurav L. Agrawal and Hitesh Gupta, ” Optimization of c4.5 decision tree algorithm for data mining application*” International Journal of Emerging Technology and Advanced Engineering***,** vol. 3, (1-3), pp. 341-345, 2013.

[20] Thierry Denoeux, Orakanya Kanjanatarakul and Songsak Sriboonchitta, ” Ek-nnclus: A clustering procedure based on the evidential k-nearest neighbor rule,” pp. 1-39, 2016.

[21] Dhaval Sanghavi, Hitarth Patel and Sindhu Nair, “Logistic regression in data mining and its application in identification ofdisease***”,*** *International Journal of Current Engineering and Technology* , vol.4, (1-6), pp. 3837-3839, 2014.