Prediction of Time Serise Data Using Data Mining Techniques

4.Dr.SAI
Computer Science Engineering
Prestige Institute of Engineering Management & Research
Indore, India

5.Dr.ECCLESTON
Computer Science Engineering
Prestige Institute of Engineering Management & Research
Indore, India

***Abstract*—the data mining techniques are basically used for handling the data in a database. That kind of databases is static in nature. But in some applications the data bases are transactional in nature and the data is continuously increasing. This kind of databases is also known as time series data. In this paper the time series dataset is used for finding efficient and accurate data model. In this context the proposed work involves the five popular machine learning algorithms. Additionally, on different dataset sizes the experiments are carried out. The experimental data is collected and based on mean performance the back propagation neural network is found in best in accuracy and error parameter.Additionally, the regression model found for memory usages and the C4.5 decision tree consumes the less amount of time for training. The selected most effective classifier is further used with the real world application for predicting the time series values.**

Keywords-BPN, SVM, regression analysis, decision tree, hidden makov model, implementation, performance study

I. INTRODUCTION

Data mining is a technique of analyzing data using the computational algorithms. The algorithms evaluate the data for recovering the essential patterns by which the prediction or categorization of data objects become feasible and accurate. In data mining there are two key types of data mining techniques are used supervised and unsupervised learning [1]. In this work the supervised learning technique is studied for predicting the next values of a solar still plant performance prediction. During this research work we need to find some accurate prediction algorithm. That accepts the different solar still plant parameters and produces the next possible value of water production. Therefore, using the prepared data set using the real world application the dataset is prepared. That dataset is a time series dataset which growing with the amount of time. Therefore, first a set of experimental data is prepared with the size of 100, 500, 1000, 2000 and 5000 samples. This data samples are used with the implemented data mining model and tried to predict next 5 min performance of water solar still. Here we have used all five supervised learning models. The implementation of the proposed model is carried out in JAVA technology, additionally the implementation usages some classes of WEKA library as JAR file in our project. After implementation the system is evaluated for their performance. The performance of the algorithms in terms of memory, time, accuracy and error rate is evaluated. Additionally, using the concluded experimental outcomes, the algorithms are selected for future application development and prediction.

II. PROPOSED WORK

The main aim of the proposed work is to perform the comparative performance study among various predictive algorithms of data mining.The basic issue is to manage and predict the time series data. Additionally, we need to predict the data values for one step ahead. In order to process and predict the values of the model is demonstrated in figure 3.1.

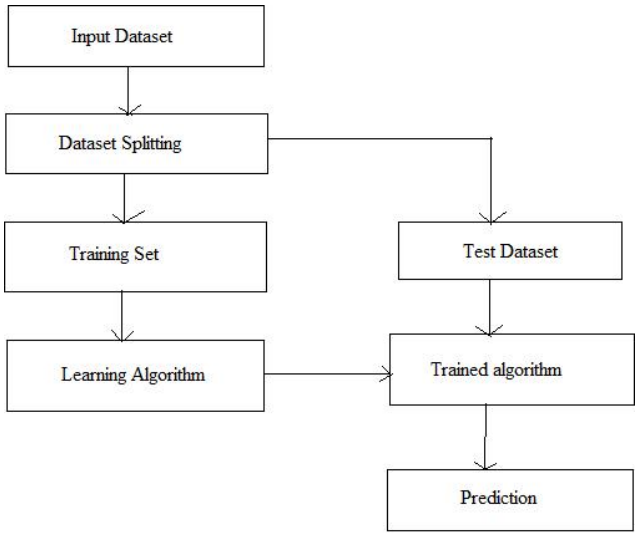

Figure 3.1 proposed model

The dataset of different size obtained from the previous experiments on solar still water plant performance is applied. Additionally, using the data, it is tried to predict the sill plant performance. The dataset is split in 70-30% ratio. The 70% of the randomly selected data from entire dataset is selected first as the training sample. Additionally, again 30% random samples are extracted here as the test dataset. There are five popular data mining algorithms namely support vector machine (SVM), back propagation neural network (BPN), C4.5 decision tree algorithm, Regression analysis and the hidden Markov model is implemented. The experimenter can select one of the implemented algorithms, additionally trained on the historical data. After that over the trained

model the test dataset is applied and their performance is evaluated.

III. ALGORITHMS

This section includes the overview of algorithms which are used for prediction:

A. SVM (Support Vector Machine)

In machine learning research, support vector machine (SVM) areone of the essential technologies.SVM is offering most robust and accurate results with respect to other well-known machine learning algorithms. It is a sound technique for learn by example task, and suitable forhigher dimensional data.Fast training execution of SVM algorithmis also available. The aim of SVM is to find the best classification function for binary.

The concept of "best" classification function ismeasured geometrically. To classify a linear dataset, a linear classification function corresponding to a hyper plane $f(x)$ is created that passes through the middle of separable classes. Once function is determined, data instance x_n can be classified simply applying the function ( ) .In these conditions belongs to the positive class when ( ) > 0. Because there are various linear hyper planes is available using which the SVM guaranteesto maximize the margin between the two classes.The margin is defined as the space available for separating two classes. The margin is corresponding to the shortest distance between the closest data points. This allows us to explore and maximize the margin, though an infinite number of hyper-planes [2].

B. BPN(Backpropagation Neural Network)

Back propagation neural network (BPN) is a popular technique for complex multivariable data analysis and non-linear relations development technique between input and output. Therefore, it worksas the biological neural system. And similar to brain that receives the information, interprets and produces output [3].A basic architecture of BPN, involve a large number of processing elements such as nodes or neurons, highly interconnected with each other as network [4]. The architecture of BPN is given in figure 2.1.

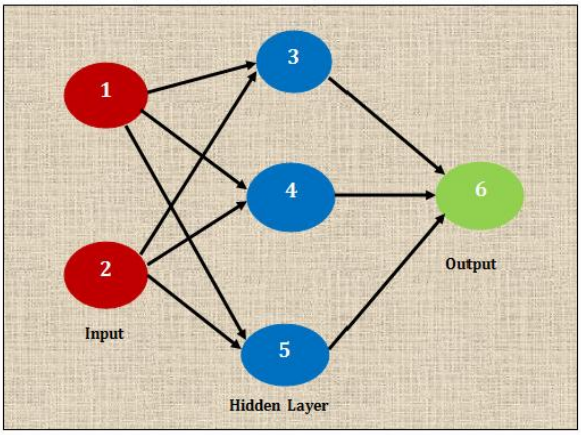

Figure 2.1 a neural network

The features include its ability to interpret dataprecisely andrecognize the patterns.Additionally, the speed and accuracy is also higher. When BPN is usedthen it is not required to have the previous knowledge therefore it is referred as black box model. Thus BPN can accept all kinds of variables or parameters [5] [6].

Advantages of NN include their high tolerance to noise, and ability to classify on which they not trained. A main concern of the training is to focus on the weights of the neurons.That is used according to the transactions used in the learning process. For each training transaction, the neural network receives in addition the expected output [7].

C. C4.5

The C4.5 or J48 decision tree decision tree is an extension of ID3 decision tree. It uses the concept of entropy and information gain (IG). Theattribute with highest IG is selected to create tree. The C4.5 algorithm recourses sub-list to create a complete tree. The algorithm considers the following constraints.

1. If Datasetcontainsingleclass then it creates a leaf node.
2. If IG computation is not feasible then it creates a node higher up then tree using the expected value.
3. If previously-unseen class found, algorithm creates a decision node using the target value.

To discuss entropy let's assume that resultant decision tree classifies data into two classes, i.e. P (positive) and N (negative). The entropy S based on this classification is:

$$E(S) = -P(Pos)log_2P(Pos) - P(neg)log_2P(neg)$$

P (pos): ratio of positive samples, P (neg): ratio of negative samples

To reduce depth of tree, while traversing, selection of best possible characteristic is usedto split tree, it is showsattribute with minimum entropy. The IG can be termed as required drop in entropy in relation with an

attribute during splitting.The IG, Gain (E, A) of an attribute A can be computed using,

$$Gain(E, A) = Entropy(s) - \sum_{=1}^{v} XEntropy(E\ )$$

The gain is used to decide positions of attributes in decision tree.Every node is positioned with maximum gain that is not considered in path yet.Because:

1. To generate small size tree.
2. To attain desired level of unfussiness.

C4.5 tree is developed by Quinlan[8]. The following steps can be used for generating decision tree

INPUT: A dataset (D).

OUTPUT: A decision tree T.

1) A node (X) is created;
2) If instance in same class.
3) Make node (X) as leaf node and assignlabel C;
4) If attribute list is empty,
5) Make node(X) a leaf node and assign a class label of most frequent class;
6) An attribute has highest IG, and then marked as test;
7) If X in role of test-attribute;
8) Generate a new branch of tree
9) If Bi is NULL,
10) Add a new leaf node, of common class**;**
11) ELSE
12) Add a leaf node and return.

D. Rigression analysis

Regression analysis (RA) is a technique used to find equations that fit the data. Once we prepare the equation, it is use this to predict values accurateaccording to available samples. The correlation coefficient demonstrates, the data is likely to be predict consequencesbased on a scatter plot that creating a straight line [9].Herewe use a simple linear regression method. The equation for a line is

$$= \quad +$$

To calculate linear regression following equation can be used.

$$y' = a + bx$$

Linear regression is a method for modeling relationship between two variables. The equationis:

$$= \quad +$$

Where, is a dependent variable, and is the independent variable, and is the slope and is the y-intercept. The variables and can be calculated as:

$$= \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x^2)}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x^2)}$$

E. HMM (Hidden Markov Model)

HMM (Hidden Markov Model) is a double implanted stochastic process with two levels. It is used to model complex processes. In a specific state, an observation can be generated according to an associated probability distribution. It is only the observation and not the state that is visible to an external observer. An HMM can be characterized by the following [10]:

1. N is the number of states in the model. The set of states' S = {S_1; S_2;..., S_N}, where S_i, i= 1;2;...;N is an individual state. The state at time instant t is denoted by qt.
2. M is the number of unique observation symbols. The set of symbols V = {V_1; V_2; ...V_M}, where V_i, I = 1;2;...; M is an individual symbol.
3. The state transition probability matrix A = [a_{ij}], where

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), 1 \le i \le N, 1 \le j \le N; t = 1,2 \ldots$$

Here $a_{ij} > 0$ for all i, j. Also,

$$\sum_{j=1} a_{ij} = 1\ , 1 \le i \le N$$

4. The observation probability matrix B = {$b_j(k)$}, where

$$b_j(k) = P(V_k | S_j), 1 \le j \le N, 1 \le k \le Mand$$

$$\sum_{k=1} b_j(k) = 1\ , 1 \le j \le N$$

5. The initial state probability vector r= , where

$$\pi_i = P(q_1 = S_i), 1 \le i \le N$$

Such that

$$\sum_{=1} \pi_i = 1$$

6. The sequence O = O_1; O_2; O_3; ...O_R, where each observationO_t isa symbols from V, and R is the number of observation.

A complete specification of HMM needs the computation of two parameters, N and M, and three possibility distributions A, B, and . The notation =(A; B;) to specify the complete set of parameters, where A, B implicitly contain N and M.

An observation sequence O, can be generated by many possible states. Consider one of them Q = q_1; q_2; ...; q_R; where 1 is the initial state. The probability that O is generated from this state sequence is given by

$$(| ,) = \prod_{=1} (| ,)$$

The independence of observations is expanded as

$$P(O|Q,\lambda) = b_{q1}(O_1)b_{q2}(O_2) \dots \dots b_{qR}(O_R)$$

The probability of the state sequence Q is given as

$$P(Q|\lambda) = \pi_{q1}a_{q1q2}a_{q2q3} \dots \dots a_{qR-1qR}$$

The probability of the observation sequence O specified as follows:

$$P(O|\lambda) = \sum_{Q} P(O|Q,\lambda)\, P(Q|\lambda)$$

The value of (|)is calculated using the direct definition of computationally intensive. Hence, a procedure Forward-Backward is used.

IV. RESULTS ANALYSIS

This section provides the understanding about the performance of implemented algorithm for time series data prediction.Thus different parameters and their descriptions are provided in this section.

A. Accuracy

The accuracy of the data mining model for time series prediction is given in table 3.1. The table contains the observations of experiments during the dataset classification. The accuracy is computed here in terms of percentage. To compute the accuracy the following formula is used:

$$= \frac{100}{}$$

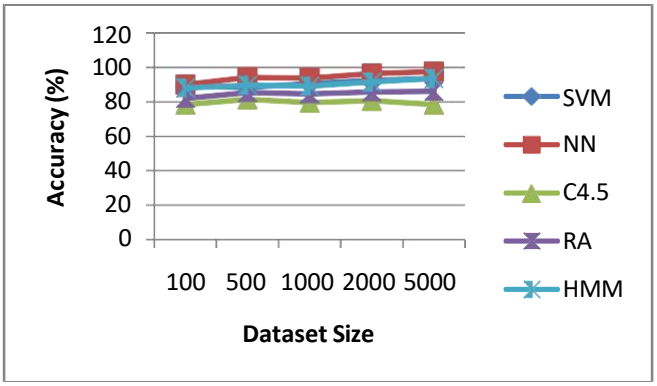

Figure 3.1 accuracy

The figure contains accuracy of the algorithms used in prediction of time series data. The accuracy described here is also evaluated for finding the mean performance. Thus the following function for mean computation is used.

$$Mean = \frac{1}{} \sum Accuracy$$

=1

Dataset size	SVM	NN	C4.5	RA	HMM
100	89	90	78.4	82	88
500	88.3	94.2	81.3	85.3	89.6
1000	90.6	93.7	79.5	84.6	89.1
2000	92.4	96.3	80.5	85.8	91.4
5000	93.6	97.5	78.4	86.1	93.5

Table 3.1 accuracy

The mean performance of the implemented algorithms is reported in figure 3.2. In this diagram the X axis shows the algorithms and Y axis shows the mean performance of the algorithms.

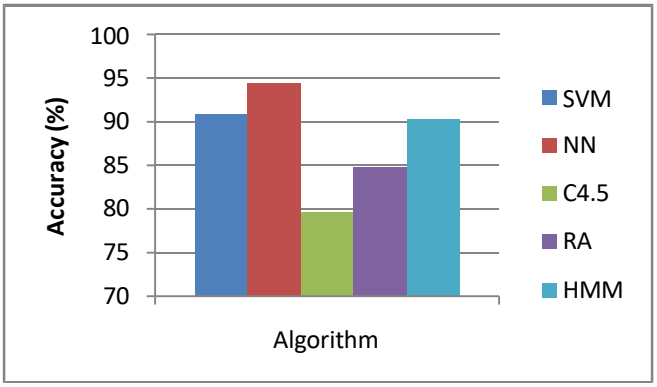

Figure 3.2 mean accuracy

According to the mean accuracy the BPN (back propagation neural network) is accuratly predict the time serise data.

B. Error rate

Error rate of the data mining algorithm told about the incorrectly classified or predicted samples among total samples required to be predicted. Thus this parameter is demonstrating the misrecognition rate of an algorithm. In various applications the following functions are used for measuring error in the model.

$$Error\ rate = 100 - Accuracy(\%)$$

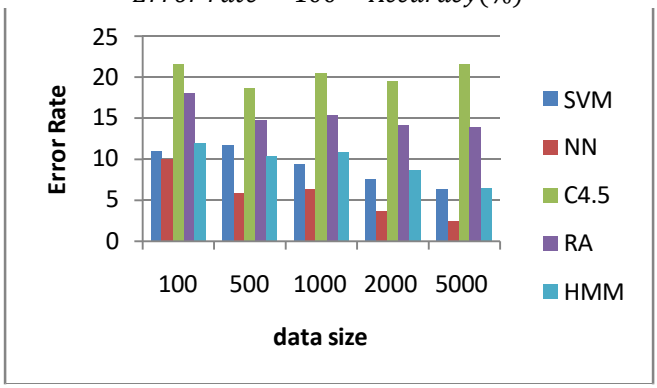

Figure 3.3 error rate

Additionally sometimes the following function is also used.

$$= \frac{100}{}$$

Dataset size	SVM	NN	C4.5	RA	HMM
100	11	10	21.6	18	12
500	11.7	5.8	18.7	14.7	10.4
1000	9.4	6.3	20.5	15.4	10.9
2000	7.6	3.7	19.5	14.2	8.6
5000	6.4	2.5	21.6	13.9	6.5

Table 3.2 error rate

In order to identify the less error producing algorithm the mean error rate of the algorithms are computed. The estimated mean error rate is reported in figure 3.4. According to the algorithm the neural network perform most accurately.

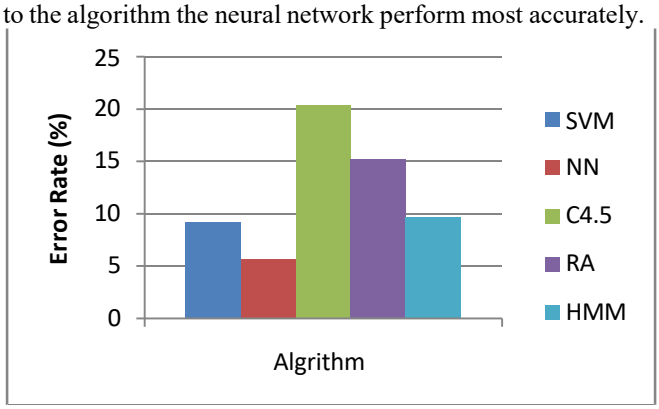

Figure 3.4 mean error rate (%)

C. Memory usages

The memory usages are also known as the memory consumption or space complexity. The performance of the implemented algorithms is given in table 3.3 and their line graph representation is reported in figure 3.5. The experiment with the different dataset size is conduced which is reported in both table as well as in figure. In fixed X-axis provide the dataset size. Additionally the consumed memory is given in Y axis.

Dataset size	SVM	NN	C4.5	RA	HMM
100	17726	17282	17283	17281	16723
500	18273	18573	18659	17726	17281
1000	19969	19287	18927	18086	17889
2000	21071	20028	19672	18452	18361
5000	24937	29371	19836	18766	18762

Table 3.3 memory usages

The memory usage is calculated on the basis of the process execution. Thus to implement this technique in java the following formulation is used.

$$memory\ usage = assigned\ memory - free\ memory$$

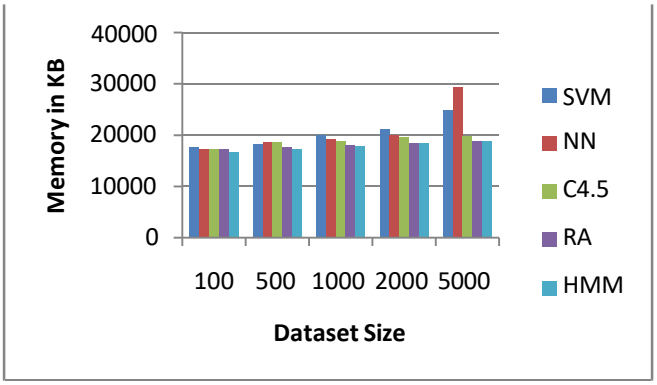

Figure 3.5 memory usages

The memory is measured here in terms of kilobytes (KB). The memory is basically increases with the amount of data size. However, to identify the most efficient algorithm among them the mean memory usages of algorithms computed and reported in figure 3.6.

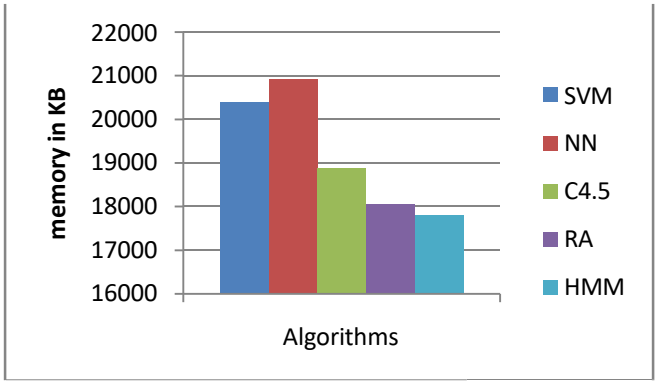

Figure 3.6 mean memory consumption

According to the mean memory usages of the algorithm the HMM usages less amount of memory and BPN algorithm consumes higher amount of memory.

D. Time consumption

The consumption of a data mining technique is given in figure 3.7 and table 3.4.The table contains the observation collected through the experiments with the increasing amount of data. The X axis representsamount of data used with the experiments and the Y axis shows the corresponding time consumed. The time consumed is described here in terms of milliseconds. In order to measure the time for computation the entire scenario is involved means:

= +

To calculate the time for training and testing scenarios the following formula is used.

$$time = Algorithm\ end\ time - start\ time$$

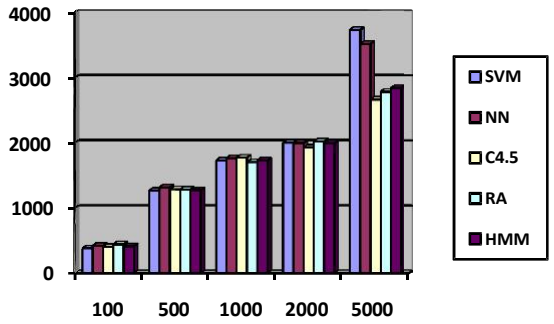

Figure 3.7 time consumption

Dataset size	SVM	NN	C4.5	RA	HMM
100	380	419	403	442	413
500	1266	1315	1287	1282	1269
1000	1727	1761	1771	1702	1729
2000	2001	1996	1928	2018	2018
5000	3728	3516	2661	2781	2781

Table 3.4 time consumption

The table 3.4 shows the performance of the five different algorithmsnamely SVM, NN, C4.5, RA and HMM model. The mean performances of these algorithms are provided in figure 3.8.

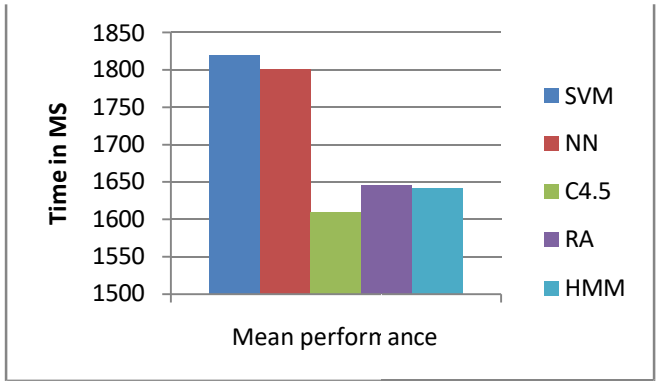

Figure 3.8 mean time consumption of algorithms

According to the mean time consumption the decision tree C4.5 shows the less time consuming algorithm for decision making.

V. CONCLUSION AND FUTURE WORK

The paper is aimed to find a suitable algorithm for predicting the time series data. In this context a real world application is used for collection of data. That data is prepared in five different sizes set and used in this experiment with the five predictive data models. These data models are neural network, support vector machine, regression, hidden Markov model, and decision tree C4.5. The experiments are carried out and their performance is summarized for selection of best among the implemented models. The performance summaries of experiments are given in table 5.1.

Algorithm	Accuracy	Error rate	Memory	Time
SVM	2	4	2	1
BPN	1	5	1	2
RA	4	2	4	4
HMM	3	3	5	3
C4.5	5	1	3	5

Figure 5.1 performance summary

On the basis of the obtained performance summary the higher accuracy in prediction is BPN algorithm is higher, on the other hand the C4.5 algorithm producing higher error rate. Additionally BPN consumes higher amount of main memory during data processing, finally SVM is much expensive in terms of training and classification time.

REFERENCES

[1] Muhammd Jawad Hamid Mughal, "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 2018

[2] Vapnik V, *"the nature of statistical learning theory"*, Springer 1995, New York.

[3] "Chapter 4 Artificial Neural Networks", https://shodhganga.inflibnet.ac.in/bitstream/10603/48/6/chaper%204_c%20b%20bangal.pdf

[4] S. L. Pandhripande and A. Dixit, "Prediction of 2 Scrip Listed in NSE using Artificial Neural Network", International Journal of Computer Applications, Volume 134, No.2, January 2016

[5] "Data Mining - Classification & Prediction", available online: , [accessed 27 April 2016] http://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

[6] S. O. Danso, "An Exploration of Classification Prediction Techniques in Data Mining: The insurance domain", Master's Degree in Advanced Software Engineering, School of Design, Engineering, and Computing. Bournemouth University, September, 2006.

[7] Dr. B. Srinivasan and K. Pavya, "A Study on Data Mining Prediction Techniques in Healthcare Sector", International Research Journal of Engineering and Technology, PP. 552-556, Volume 3, Mar-2016

[8] Kundan Kumar Mishra, Rahul Kaul, "Audit Trail Based on Process Mining and Log", International Journal of Recent Development in Engineering and Technology, Volume 1, Issue 1, Oct 2013

[9] *"Chapter 4. Regression and Prediction"*, O'Reilly, https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch04.html

[10] Shweta Jaiswal, Atish Mishra and Praveen Bhanodia," Grid Host Load Prediction Using GridSim Simulation and Hidden Markov Model", International Journal of Emerging Technology and Advanced Engineering, Volume 4, July 2014.

[11]