Exploratory Data Analysis for Social Big Data Using Regression and Recurrent Neural Networks

R.S. Aswini
Faculty of Computer Science and Engineering, Department of Engineering and Technology.
E-mail: aswini_rs@srmist.edu.in

B. Muruganantham
Faculty of Computer Science and Engineering, Department of Engineering and Technology.
E-mail: murganb@srmist.edu.in

S. Ganesh Kumar
Faculty of Computer Science and Engineering, Department of Engineering and Technology.
E-mail: ganeshk1@srmist.edu.in

A. Murugan
Faculty of Computer Science and Engineering, Department of Engineering and Technology.
E-mail: murugana@srmist.edu.in

Received August 30, 2020; Accepted November 02, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V17I2/WEB17077

Abstract

In general the health care system and hospital takes a major role in the service sector. The clinical treatment successfully increases in the year for the treatment of both the medical and the technical innovations. A sample theoretical challenge exists in the patient flow analysis for real data. An existing method target on the audience and less concentration on the secondary statistical analysis, where the data obtain from the hospital not suitable for the analysis. So this limitation can overcome using the Exploratory Data Analysis (EDA), which helps in analysis of the patients flow in large hospital. The proposed frame work uses a machine learning method for the data classification processes. The feature extraction processes for the patient data applied for the larger hospital dataset and the individual hospital data. Some similar features are allowed to train over the Recurrent Neural Network (RNN) classifier for data modeling using the large hospital dataset. The output of the classifier has the specific details about the patients taken for the EDA method. The linear regression algorithm be the one kind of statistical tool for predicting the relationship between the variables. The proposed frame work is implemented using Mat lab R2014a software and the results were simulated. The relationship between the patient details and hospital information shows the status of the hospital as healthy and un healthy status.

Keywords

EDA, Feature Extraction, RNN, Linear Regression, Medical Record.

Introduction

An Exploratory Data Analysis (EDA) be the user interactive and the best visualization tool for analyzing the Electronic Medical Records (EMRs). The record contains a vast amount of data relate to the medical policy makers, clinical researchers and physicians. The proposed frame work develops a uniform data analysis method for cohort study, which focus on the specific diseases. An interactive feature for divide and conquer way to classify the relatively uniform patients among the individual group. This is repetitive process make an easy way to divide the data into subsets of homogeneous standard. This subset examines the data visually, refined and compared. The final steps derive the data transformation and the feedback from the user to complete the repetitive process.

The method attracts the companies and the professional relates with the health care examines the predictive analysis with the machine learning, which enable to analyze the patients data and determine patients outcomes such as the worse condition or improving the health condition and finding the illness in individual family. So statistics with the EDA analysis the data set and to summarize the main characteristics of the patients, which frames the hypothesis test task and formal modeling. So the EDA forms the formal modeling techniques [1]. The EDA provide benefits for the data set using the statistical and visualization f results in data test accuracy. The classical statistics verify the hypothesis designed for the problem and the fit the specific models, which explains the individual relationships in the data. The hypothesis make the clarity for t he specific problems to understand the data by using the machine learning process. The multidimensional problems can be grouped using the clustering process in two steps. The first step of the clustering process is similar to discriminate analysis. The second step of the clustering process provides the advance level of group membership for the classification process. So the discriminative analysis lags in group membership level[2]. Some level of functions such as the multiple analysis, a free model approach and the iterative data analysis lags in the accuracy of result evaluation techniques. Similar techniques such as the Tukey's methods with the linear model and Poisson distribution and Bayesian methods results with the various plots leads to the complex situations. These problems can be overcome using the best data visualization techniques as EDA [4]. The proposed model designed to evaluate the factors of psychology, medicine and social science related data. The analysis techniques such as the classification and regression of the machine learning clearly predict the feature extraction process but the artificial neural

networks has the training facility for the extracted features with the EDA techniques predicts the result in an accurate way[5]. Considering the factors the proposed uses the Recurrent Neural Networks (RNN) models has the facility to storage the previous data and compare the present during the execution of the process, which results in reduction of the error. The next session 2 discusses about the literature survey and session 3 discusses about Non-graphical EDA. The session 4 discusses about the block diagram representation for the EDA method using RNN. The session 5 discusses about the result analysis and conclusion.

Literature Survey

Plamen P. Angelov et al., [6] proposed model uses the generic or probabilistic approach for the local multi model systems. The data driven approach lags in the non parametric cloud model for storing, extracting and analyzing the data properties. The process involves the meta data analysis in the cloud uses the large memory and complex calculation efficiency. The method lags in the real time data analysis, which involves the time varying information about the patients such as the stress. The author Cornelia Setz et al., [7] uses the wearable sensors for monitoring the real time stress of the individual patients follows the measures the EDA method. The high peaks of the data not fully support the EDA distributions. The instantaneous peaks of the stress signal level for a person needs some classification level. So the peaks form the load for the analysis. The problem leads the way to the sampling method used by the author Sheng-Yao Wang et al.,[8] generates the solution for the peak of the stress signal by sampling the data with the probability model. The author proposed the critical path for the sampling stages with the Distributed Assembly Permutation Flow-Shop Scheduling Problem (DAPFSP). The critical path avoids the invalid path of error data in the process. So the way changes to the local search of the data in the noisy region as the searching phase of the EDA analysis. The critical path for the sampling process needs the internal validation steps for the processes. When analyzing the high peak values for the stress data under goes the various level of the calculation. So the author Yuki Murai et al., [9] predicts the new way to minimize the steps in the procedure using the genetic algorithm (GA), which pre estimate the flow steps in the process and the uses the buffer for the storing the intermediate values of the signal with the help of logical circuits using channel routing algorithm. The routing algorithm takes the long time to process the data and make it difficult to complete the task in timing. The author JIAN-PING FANet al., [10] proposed the best technology to minimize the distance to average distance. The average solution increases the cross – efficiency for the EDA analysis. The positive distance from average solution (PDA) and negative distance from average solution (NDA) calculated to give the variable data length.

The proposed method overcomes the problem of variable length inputs and outputs method using the machine learning method RNN. The variable length of the data samples has undergone the processes of the feature extraction steps to predict the accurate data with the help of EDA techniques.

A. Non-Graphical EDA

The proposed method first analysis raw data in the non- graphical EDA, here the raw data presented in the tabulation format able to visualize, test the independent and dependent categories of variables. The raw data may have some missing component represented by the "NA" (Not Available) category. The missing data produce the disturbance for the following process, which should be corrected as the sensitive values for the processes with the understandable form. The dataset were pre- processed for the multivariable logistic regression. The regression model uses the below calculation for the EDA processes.

B. Regression with EDA

The regression theory describes the dependent variable or concept by an equation, which describes the relevant information for the various influential factors. The metaphor to explain the basic idea for the target variables in EDA uses the equation (1).

$$EDA = VIS+ MA +MF +INT \quad (1)$$

The characteristics of EDA determined using the structural equation, here VIS indicates the visual analysis tool for the graphical representation becomes easy for the people in graphical representation than the mathematical model. The MA indicates the multiple analysis for special features in the EDA approaches for various representations and the different levels of data reduction be the potential inherent structures. The MF be the free model for the analyzing problems which relates the plain data in the graphical representation form. The INT be the interactive approaches, which resembles in data arrangement to show the conceptual knowledge in the next step for processing. The EDA be the one algorithm for the linear regression. The algorithm shows an optimal straight line between the two or more variables. This straight line able to predict unknown variables and the relationship between these variables.

Exploratory Data Analysis Using Regression Model

The EDA is cross classified into the two ways. The first one is the non- graphical and graphical method. The second method either uni-varied or multi- varied normally bi-varied. The non- graphical form involves the statistical detail and the graphical form

works with the pictorial representation of the summarized data. In the proposed frame works extract the features based on the multivariate method choose the more variables at a time, which explores the relationships between the datasets. These features are trained over the recurrent neural network (RNN) to classify the multivariate into groups.

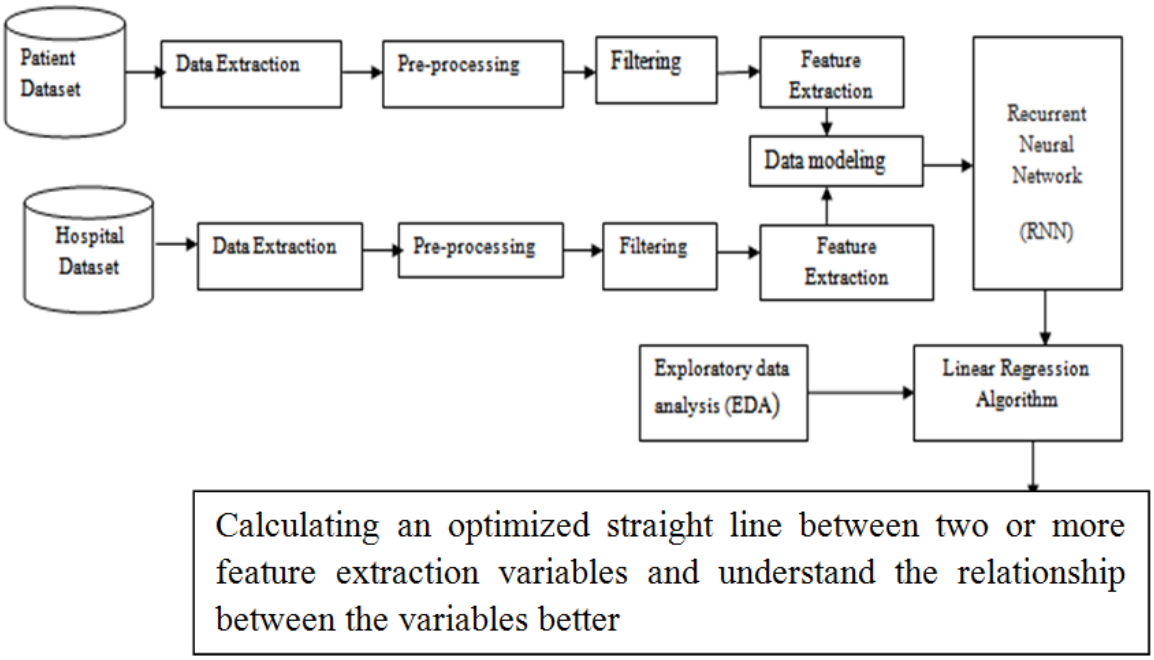

Figure 1.a Block diagram representation EDA using RNN with linear regression

The Figure 1.a shows,

- The storage of medical records containing information of both disciplined and unrestricted nature.
- The handling of data flow around the hospital.
- The gathering of statistics for a management information service.

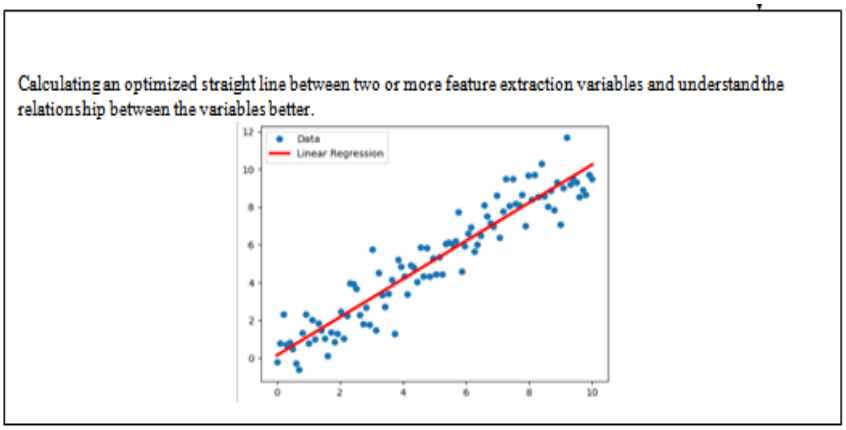

Figure 1.b linear regression output

The figure 1.b shows the linear regression where the straight-line coincide of data with the straight line.

Data Extraction

Table 1.a Data extraction from individual dataset

Input Data	Hospital Dataset	Sex	Age	Weight	Blood Pressure	
	Patients Dataset	Gender	Age	Weight	Systolic	Diastolic

The table 1.a shows the details of the features such as the gender, age, weight and blood pressure in terms of Systolic and Diastolic.

Pre – Processing Stage

In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

A. Filtering

The filtering process removes the unwanted noisy data present in the data set. The data preparation process and the filtering steps consume more processing time. So the data preprocessing steps includes the cleaning, value selection, normalization and data transformation. Additionally feature extraction and the selection process carry the data to next steps for the final training process. The complexity in the data preparing reduces by following the below processes.

B. Data Reduction Algorithm

Step 1: Importing libraries for the dataset.
Step 2: Import the data set in matrix format
Step 3: Adjusting the missing data with the meaning full form
Step 4: Splitting the data into the categorical wise.
Step 5: Separate the data set in to the training set and testing set.
Step 6: Compute the feature extraction process.

C. Feature Extraction Using Statistical Pattern

The concepts of machine learning uses the pattern recognition with the feature extraction starts at the initial level of the measured data and the derived features were the specific

details needed for further processing of RNN classification. These features were trained over all the data present in dataset.

Table 1.b Feature extraction from individual dataset

Parameter condition	Selected Feature	Selection condition
Condition 1	Gender	0- Male, 1-Female
Condition2	Age	(10-20), (20-40), (50-60),(60>)
Condition 3	Weight	(Height-weight): > 65
Condition 4	Blood Pressure	Systolic (<120mmHg) Diastolic (<80mmHg)

The table 1.b shows the Feature extraction from individual dataset with training conditions.

The data present in the dataset were large so these data categorize into the know group of data as the cluster. Finally the data are transformed in to the specific form based on the above condition for the dataset. These results are easy to catch the data during the RNN training process.

D. Recurrent Neural Networks

The Recurrent neural networks, also known as RNNs, are a class of neural networks that allow previous outputs to be used as inputs while having hidden states. In the RNN network artificial neurons are connected as the artificial neural networks, here the nodes are inter connected as the directed graph for the temporal sequence. So the networks behave as the dynamic temporal way in order to use the their internal state as the memory for processing the sequences of inputs.

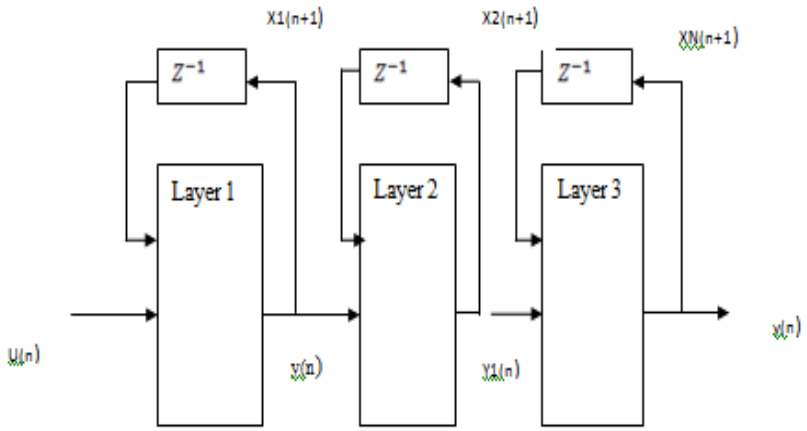

Figure 2 Recurrent multi-layer perceptron (RMLP) architecture with N layers

The figure 2 explains the detailed network structure for the RNN processes. The network investigates the discrete time multi layer perceptron using the local and layer interconnects the feed back connections using the hidden layer. So that the individual nodes needs the training recurrent links to its and to other nodes in the succeeding layer. The connections are delayed using one time unit. The layer details are given below. The layer structure needs the network designed with the fixed sequence for three different types of layers.

One Input Layer

The layer with the one or more recurrent interconnected hidden layer with the recurrent links using the one time delay unit and the forward the interconnect from one hidden layer to the next is completed.

One Output Layer

The nodes of the RMLP be the nets based on the McCul loch and Pitts model. The network details the input net i to j node and j to i node is defined as the total number of subnets is L=2+number of hidden subsets. The layers were numbers from 0 to ends with the infinity.

$$net_{i,j=\sum_{k=1}^{N_{i-1}} w\,{}^{f,i}_{k,j}\; y_{i-1,k(n)}} + \sum_{k=1}^{N_{i-1}} w\,{}^{r,i}_{k,j}\; y_{i,k(n-1)} \quad (1.a)$$

Where $y_{i,j}$ be the activation of node j in the layer i.

$w\,{}^{f,i}_{k,j}$ be the forward weight from node k in the layer (i,1) to node j in the layer (i). The

$w\,{}^{r,i}_{k,j}$ be the recurrent weight for the node k in the layer (i) to node j in the next layer (j).

The N represent the number of nodes in the subnet (j). The time index be (n,1) indicates the feed back in the delay for one time step.

$$y_{i,j}(n) = \boldsymbol{\sigma}(net_{i,j}(n)) \quad (2)$$

or the symmetrical transfer function:

$$\sigma(s) = \tanh s$$

$$\sigma'(s) = \big(1 + \sigma(s)\big)\big(1 - \sigma(s)\big) \quad (3)$$

The feature extraction and the clustering algorithm worked based on the deep noise encoder for the forward [11] and [12]. The algorithm covers the spatial vector for the high dimension data converted in to the low dimension data features using the deep learning network. The experimental results shows the extractive text features for the short text with the clustering methods improves the effective clustering and convert the high dimensional data into the low dimensional data [13] - [14] and [15].

Result and Discussion

The input data from the hospital data sets. The official datasets download using the Medicare.gov Hospital Compare Website provided by the Centers for Medicare & Medicaid Services. These data allow you to compare the quality of care at over 4,000 Medicare-certified hospitals across the country. https://data.medicare.gov/data/hospital-compare.

Dataset 1: Hospital with the fields showed below shown in figure 3

	name	sex	age	wgt	smoke	sys	dia	trial1	trial2	trial3	trial4
YPL-320	'SMITH'	'm'	38	176	1	124	93	18	-99	-99	-99
GLI-532	'JOHNSON'	'm'	43	163	0	109	77	11	13	22	-99
PNI-258	'WILLIAMS'	'f'	38	131	0	125	83	-99	-99	-99	-99
MIJ-579	'JONES'	'f'	40	133	0	117	75	6	12	-99	-99
XLK-030	'BROWN'	'f'	49	119	0	122	80	14	23	-99	-99
TFP-518	'DAVIS'	'f'	46	142	0	121	70	19	-99	-99	-99
LPD-746	'MILLER'	'f'	33	142	1	130	88	0	13	-99	-99
ATA-945	'WILSON'	'm'	40	180	0	115	82	-99	-99	-99	-99
VNL-702	'MOORE'	'm'	28	183	0	115	78	2	-99	-99	-99
LQW-768	'TAYLOR'	'f'	31	132	0	118	86	11	-99	-99	-99
QFY-472	'ANDERSON'	'f'	45	128	0	114	77	8	10	14	-99
UJG-627	'THOMAS'	'f'	42	137	0	115	68	4	9	-99	-99
XUE-826	'JACKSON'	'm'	25	174	0	127	74	-99	-99	-99	-99
TRW-072	'WHITE'	'm'	39	202	1	130	95	8	-99	-99	-99
ELG-976	'HARRIS'	'f'	36	129	0	114	79	-99	-99	-99	-99
KOQ-996	'MARTIN'	'm'	48	181	1	130	92	13	15	21	27
YUZ-646	'THOMPSON'	'm'	32	191	1	124	95	-99	-99	-99	-99
XBR-291	'GARCIA'	'f'	27	131	1	123	79	-99	-99	-99	-99
KPW-846	'MARTINEZ'	'm'	37	179	0	119	77	0	-99	-99	-99

Figure 3 Hospital dataset parameters

Dataset 2: Patient dataset

Machine Learning is exploding into the world of healthcare. When we talk about the ways ML will revolutionize certain fields, healthcare is always one of the top areas seeing huge strides, thanks to the processing and learning power of machines. There's a good chance

you either are or will soon be employed in the healthcare field. A while back, I wrote a list of 25 excellent open datasets for ML and included healthdata.gov and MIMIC Critical Care Database[16],[17] and [18]. Here are 15 more excellent datasets specifically for healthcare as shown in figure 4.

	Gender	Age	Location	Height	Weight	Smoker	Systolic	Diastolic
Smith	'Male'	38	'County General Hospital'	71	176	true	124	93
Johnson	'Male'	43	'VA Hospital'	69	163	false	109	77
Williams	'Female'	38	'St. Mary's Medical Center'	64	131	false	125	83
Jones	'Female'	40	'VA Hospital'	67	133	false	117	75
Brown	'Female'	49	'County General Hospital'	64	119	false	122	80
Davis	'Female'	46	'St. Mary's Medical Center'	68	142	false	121	70
Miller	'Female'	33	'VA Hospital'	64	142	true	130	88
Wilson	'Male'	40	'VA Hospital'	68	180	false	115	82
Moore	'Male'	28	'St. Mary's Medical Center'	68	183	false	115	78
Taylor	'Female'	31	'County General Hospital'	66	132	false	118	86
Anderson	'Female'	45	'County General Hospital'	68	128	false	114	77
Thomas	'Female'	42	'St. Mary's Medical Center'	66	137	false	115	68
Jackson	'Male'	25	'VA Hospital'	71	174	false	127	74
White	'Male'	39	'VA Hospital'	72	202	true	130	95
Harris	'Female'	36	'St. Mary's Medical Center'	65	129	false	114	79
Martin	'Male'	48	'VA Hospital'	71	181	true	130	92
Thompson	'Male'	32	'St. Mary's Medical Center'	69	191	true	124	95
Garcia	'Female'	27	'VA Hospital'	69	131	true	123	79
Martinez	'Male'	37	'County General Hospital'	70	179	false	119	77
Robinson	'Male'	50	'County General Hospital'	68	172	false	125	76

Figure 4 Patient dataset parameters

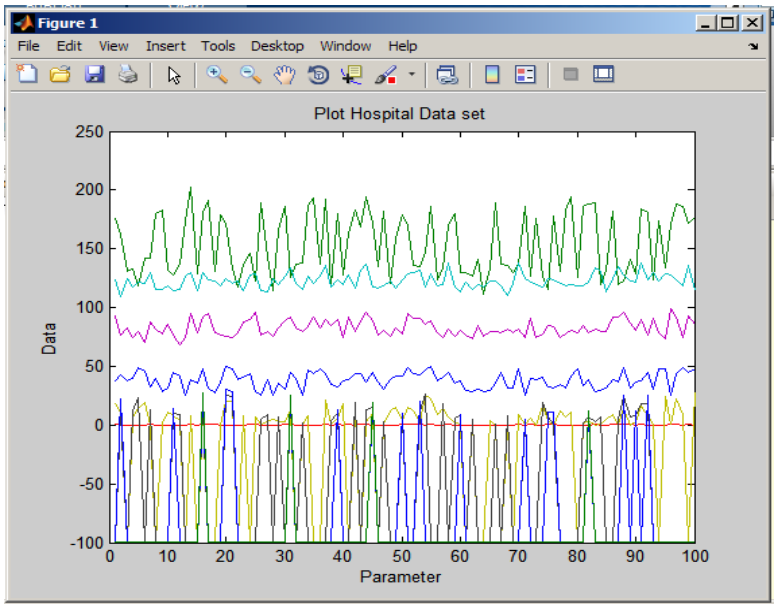

Figure 5 Dataset 1: Hospital

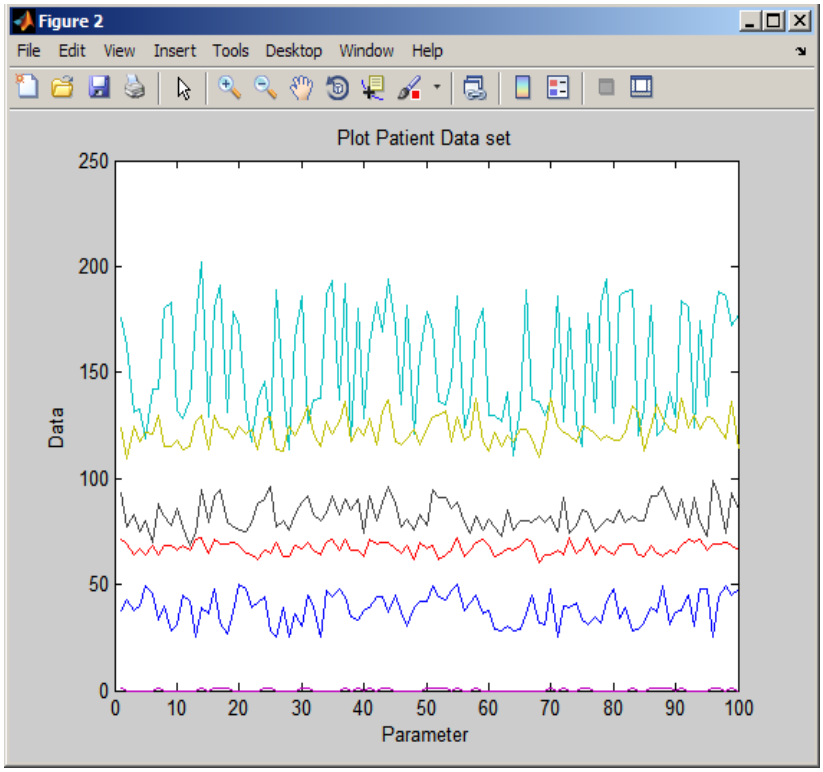

Figure 6 Dataset 2: Patient dataset

The gender, age, location, height, weight, smoker, systolic and diastolic and Self Assessed Health Status details were plotted using the graph for the both datasets.

Linear Regression

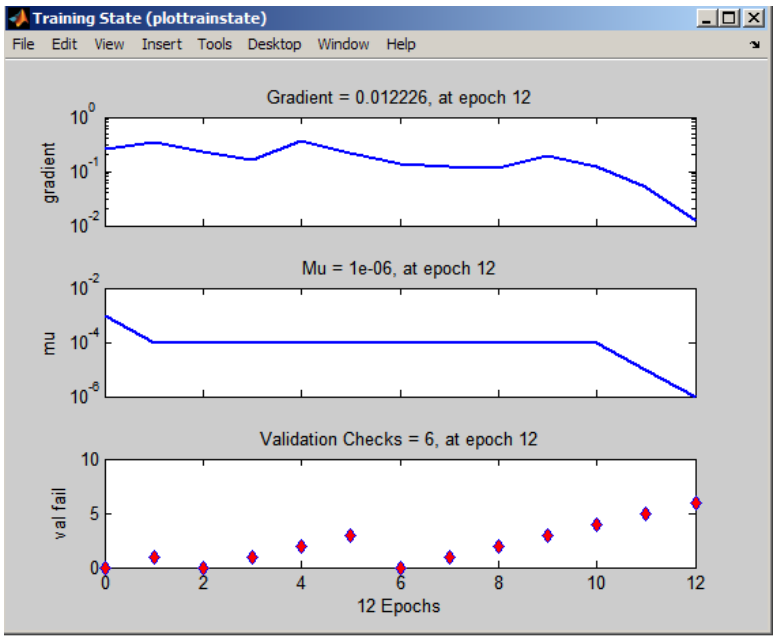

Figure 7 Epochs for linear regression

The figure 7 shows the individual epoch when the dataset, with the data passed form the forward and backward in the neural networks. When the epoch results is big where the data feed to the algorithm divides into several smaller batches.

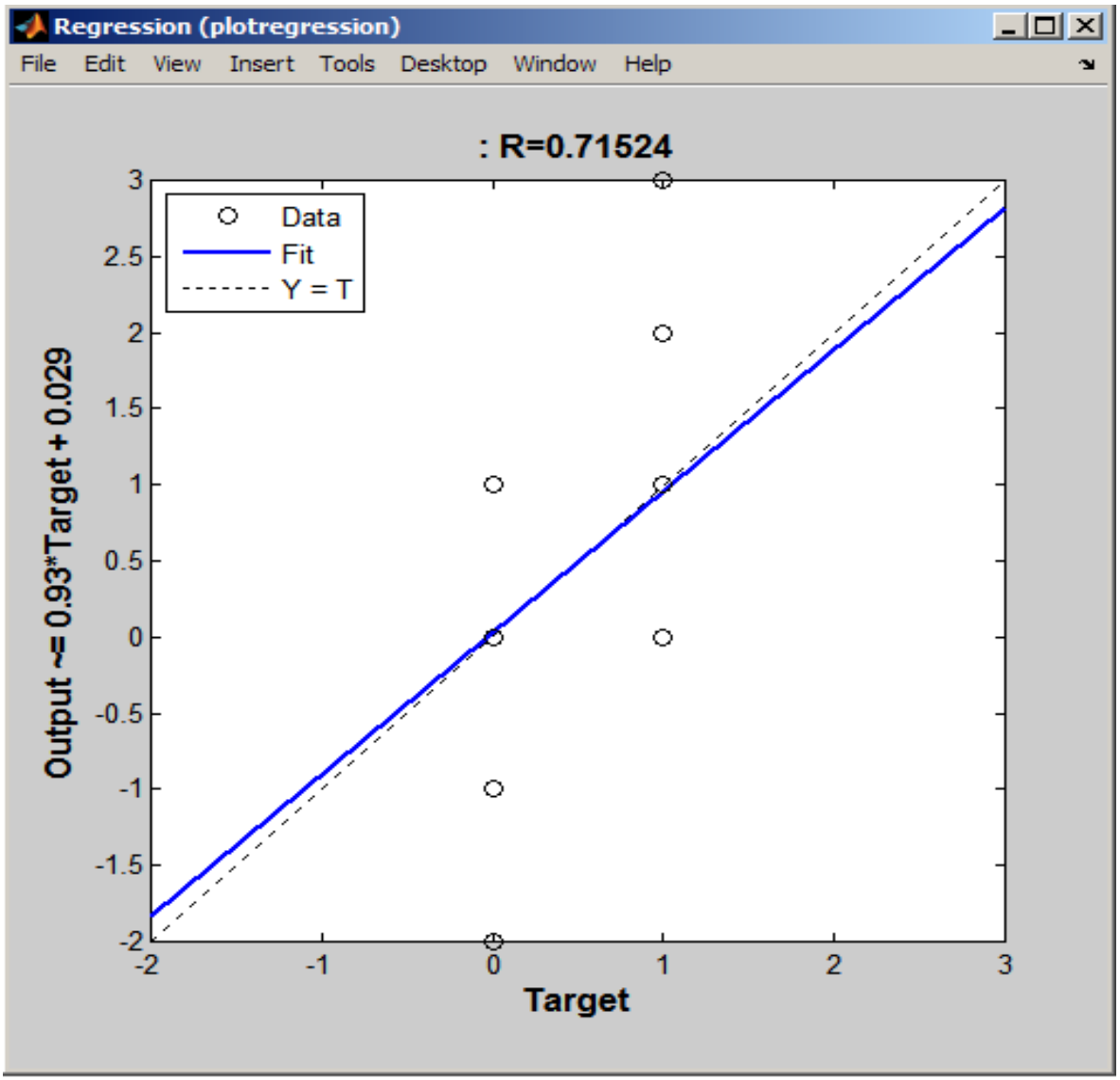

Figure 8 The linear regression of targets **relative to** outputs

The figure 8 plots the linear regression of targets relative to outputs. The data fit with the database were 74.38% of the data features matches with the age, weight, sex and the blood pressure. An independent variable is numerically related to the dependent variable [19].

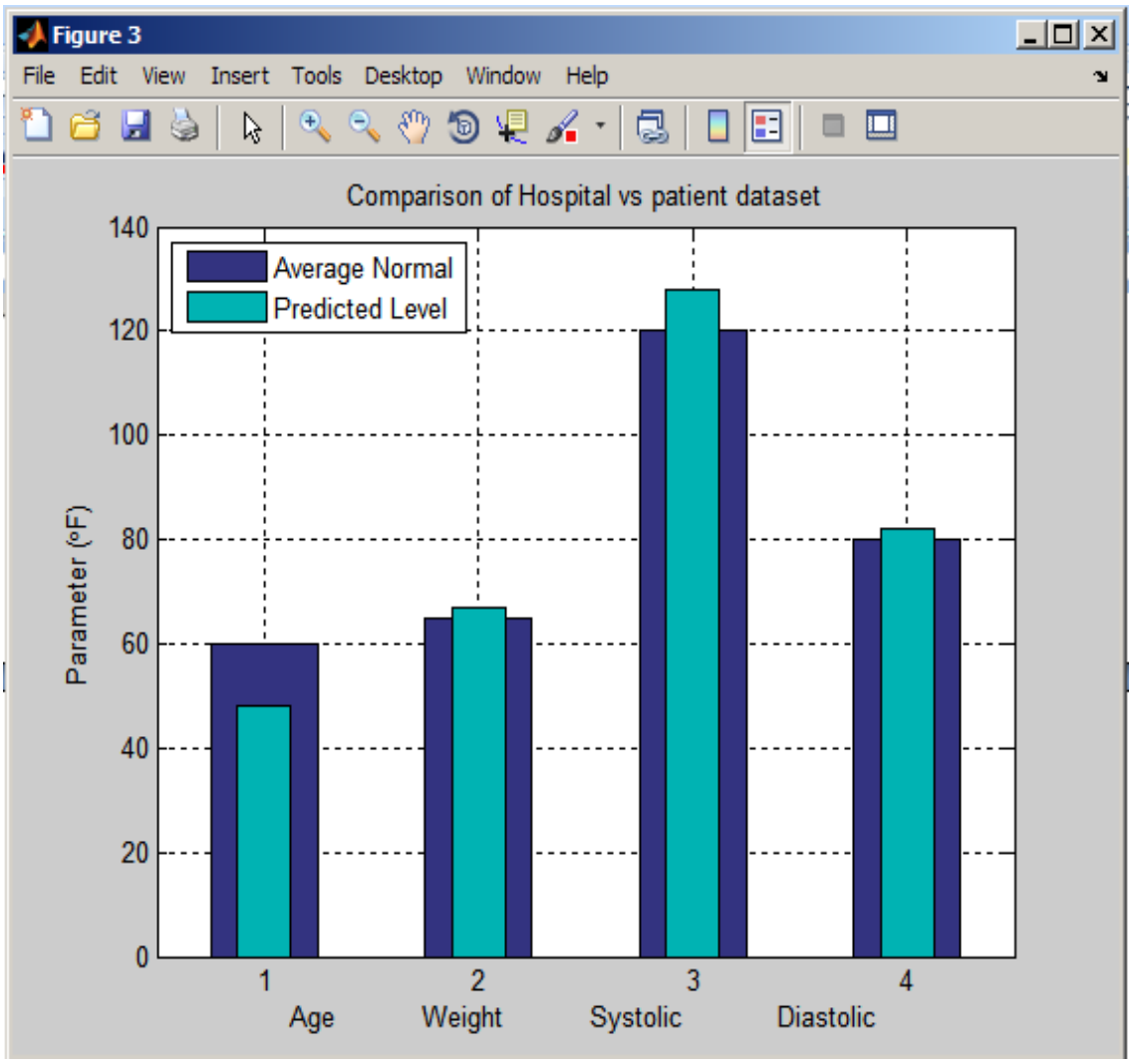

Figure 9 Comparison graph for the hospital and the patient dataset

The figure 9 shows the present data features for the patients dataset is related with the hospital dataset. The age factor is normally lower with the dataset, weight is increased with the normal level, the systolic is higher than the normal level and diastolic also higher. Finally the corresponding dataset features were healthy [19] and [20].

Table 2 Average parameter

Sl.No.	Parameter	Value
1.	Average age of the person	48
2.	Average Weight of the person	67
3.	Average Systolic Pressure of the person	128
4.	Average Dystolic Pressure of the person	82
5.	The given data	Healthy

Table 3 RNN Analysis

Sl.No.	Parameter	Range
1.	Accuracy	86%
2.	Sensitivity	0.98
3.	Specificity	1

The table 2 and 3 shows the average persons for the feature extraction results and the table 3 shows the analysis report for the RNN. The accuracy for the system is 86%, the sensitivity and the specificity for the processes is 98% and the specificity around to 1.

Conclusion and Future Work

The proposed EDA method with the RNN networks gives the performs of the 86% with the patients best with the hospital data set. The two different dataset were analyses through the selective common features such as the age, weight, systolic pressure and diastolic pressure for the person. These information derived to the common factors for training the RNN network. The RNN networks executed based on the time. So the method processed in the time series prediction due to the feature which remember the previous inputs. The patients set details were analyzed to be healthy. In future the work is implemented using Convolution Neural Networks for training more accuracy for the hospital dataset.

References

Huang, C.W., Lu, R., Iqbal, U., Lin, S.H., Nguyen, P.A.A., Yang, H.C., & Jian, W.S. (2015). A richly interactive exploratory data analysis and visualization tool using electronic medical records. *BMC medical informatics and decision making*, *15*(1), 1-14.

Yu, C.H. (2010). Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research*, *3*(1), 9-22.

Velleman, P.F., & Hoaglin, D.C. (1981). *Applications, basics, and computing of exploratory data analysis.* Duxbury Press.

Gelman, A. (2004). Exploratory data analysis for complex models. *Journal of Computational and Graphical Statistics*, *13*(4), 755-779.

Mishra, S., & Palanisamy, P. (2018). Multi-time-horizon solar forecasting using recurrent neural network. *In IEEE Energy Conversion Congress and Exposition (ECCE)*, 18-24.

Angelov, P.P., Gu, X., & Príncipe, J.C. (2017). Autonomous learning multimodel systems from data streams. *IEEE Transactions on Fuzzy Systems*, *26*(4), 2213-2224. http://doi.org/10.1109/TFUZZ.2017.2769039.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2009). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on information technology in biomedicine*, *14*(2), 410-417.

Wang, S.Y., & Wang, L. (2015). An estimation of distribution algorithm-based memetic algorithm for the distributed assembly permutation flow-shop scheduling problem. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *46*(1), 139-149. http://doi.org/10.1109/TSMC.2015.2416127.

Murai, Y., Ayala, C.L., Takeuchi, N., Yamanashi, Y., & Yoshikawa, N. (2017). Development and demonstration of routing and placement EDA tools for large-scale adiabatic quantum-flux-parametron circuits. *IEEE Transactions on Applied Superconductivity*, *27*(6), 1-9. http://doi.org/10.1109/TASC.2017.2721965.

Fan, J.P., Li, Y.J., & Wu, M.Q. (2019). Technology Selection Based on EDAS Cross-Efficiency Evaluation Method. *IEEE Access*, *7*, 58974-58980. http://doi.org/10.1109/ACCESS.2019.2915345.

Al Mashhadany, Y.I. (2012). Recurrent neural network with human simulator based virtual reality. *Recurrent Neural Networks and Soft Computing*, 89-114. http://doi.org/10.5772/35538.

Tutschku, K. (1995). *Recurrent multilayer perceptrons for identification and control: The road to applications.* Univ. Würzburg, Germany, ser. Research Report Series, Report No. 118.

Abraham, A. 129: *Artificial Neural Networks, Handbook of Measuring System Design, edited by Peter H. Sydenham and Richard Thorn.* 2005 John Wiley & Sons, Ltd. ISBN: 0-470-02143-8. Oklahoma State University, Stillwater, OK, USA, 2005.

Baruch, I., Gortcheva, E., Thomas, F., & Ruben, G. (1999). A neuro-fuzzy model for nonlinear plants identification. *Proceedings of the IASTED International Conference Modeling and Simulation (MS '99), Philadelphia, Pennsylvania – USA*, 326-331.

Chéron, G., Duvinage, M., Castermans, T., Leurs, F., Cebolla, A., Bengoetxea, A., & De Saedeleer, C. (2007). Toward an Integrative Dynamic Recurrent Neural Network for Sensorimotor Coordination Dynamics. *Recurrent Neural Networks for Temporal Data Processing*.

Wu, J., Sun, J., & Liang, L. (2012). DEA cross-efficiency aggregation method based upon Shannon entropy. *International Journal of Production Research*, *50*(23), 6726–6736.

Wang, Y.M., Chin, K.S., & Jiang, P. (2011). Weight determination in the cross efficiency evaluation. *Computers & Industrial Engineering*, *61*(3), 497–502.

Stevic, Ž., Vasiljevic, M., Pu ˇÅąka, A., Tanackov, I., Junevicius, R., & Veskovic, S. (2019). Evaluation of suppliers under uncertainty: A multiphase approach based on fuzzy AHP and fuzzy EDAS. *Transport*, *34*(1), 52–66.

Feng, X., Wei, C., & Liu, Q. (2018). EDAS method for extended hesitant fuzzy linguistic multi-criteria decision making. *International Journal of Fuzzy Systems*, *20*(8), 2470–2483.

Khouja, M. (1995). The use of data envelopment analysis for technology selection. *Computers & Industrial Engineering*, *28*(1), 123–132.