EVALUATING THE PERFORMANCE OF NEURAL NETWORK USING FEATURE SELECTION METHODS ON PIMA INDIAN DIABETES DATASET

[1] Dr. Raghavendra S., [2] Santosh Kumar J., and [3] Dr. Raghavendra B. K.
[1] Associate Professor, Department of Computer Science and Engineering, CHRIST (DEEMED TO BE UNIVERSITY), Bengaluru, India
[2] Associate Professor Department of Computer Science and Engineering, KSSEM, Bengaluru, India
[3] Professors, Department of Computer Science and Engineering, KSSEM, Bengaluru, India

ABSTRACT: Diagnosis of Diabetes disease at beginning stage is important for healthier treatment. In today's scenario equipment's like sensors are used for discovery of infections. Accurate classification techniques are necessary for automatic detection of disease samples. This study utilizes data mining techniques for classification of Diabetes patients. Neural network algorithm was implemented for classification using R platform. Classification and prediction of medical datasets poses real challenges in Data Mining. To deal with these challenges and Artificial Neural Network (ANN) is commonly used. LR enables us to examine the relationship between a categorical outcome and a set of descriptive variables. LR explains that there can be one or more self-governing variables that can establish the problem outcome. ANN resembles the human brain and here the information is processed by simple elements called neurons and signals are transmitted between the neuron from the experimental results it is identified that for Diabetes dataset NN with 10 fold using percentage split prediction correctness of 84.52% is achieved. Artificial neural networks are considered as a field of artificial intelligence. It has also been applied in many disciplines, including biology, psychology, statistics, mathematics, medical science, and computer science. Feature selection is the process of choosing a small subset of features that is sufficient to describe the target model. In this research work an attempt has been made to evaluate neural network model with feature selection methods such as forward selection and backward elimination using cross validation sampler and percentage split on publicly available medical datasets. The classification accuracy is used to measure the performance of the model. From the experimental results it is confirmed that the neural network model with backward elimination feature selection method using percentage split gives more efficient result.

Keywords – Artificial Neural Network, Feature Selection Method, Percentage Split, Cross validation sampler

I. INTRODUCTION

In recent years, electronic health records in modern hospitals and medical institutions larger to get better the value of patient care and increase the output and efficiency of health care Diabetes. So methods for efficient processor based analysis are needed due to the inadequacy of traditional manual data analysis. Machine learning methods have been a wonderful support for making prediction of a particular system by training. In recent year's machine learning has been the developing, reliable and supporting tool in medical area. Due to recent advances in machine learning, medical analysis improves diagnostic accuracy, reduces cost and reduces human resource.

Medical analysis is a tough and complete task, and it must be passed out well and exactly. ANN is believed as a arena of Artificial Intelligence. The progress of the representations was enthused by the neural construction of humanoid brain. ANN have stayed applied in many disciplines and recently, ANN have developed as a very widespread model and used in the diagnosis of many diseases [1].

ANN is measured as an vital arena of Artificial Intelligence. The ANN model growth was driven by the neural design of humanoid brain. ANNs is successfully applied in various fields such as environmental science, study of numbers, study of medicine, study of computers etc. ANNs are also being used in many business areas like accounts and audits, funding, managing and decision making, promotion and manufacture etc. ANNs have turned out to be a well-liked model and recently they are used to identify diseases and to forecast the patients' survival proportion [2].

Feature selection is the procedure of recognizing and eradicating as much of the immaterial and redundant data as possible. Feature selection is frequently considered as a essential pre-process phase to analyze these data, as this technique can decrease the dimensions of the data and frequently leads to better analysis [2].

The planned research work is mainly paying attention to obtain better classification accuracy with less number of attributes by which we reduce the amount of time required for prediction and also improve the classification accuracy. This will result in reducing the number of tests that is to be done while predicting the presence or absence of disease.

II. LITERATURE SURVEY

Multilayer Feed-forward Network: This network is made of input, hidden and output layers. Supervised learning is an approach to find the input-output association from the training using a set of data. Learning system is fed with the input data and generates output, which is then comparing with the target to calculate the error signal. The error is sent to the learning system for more training until the least value of error is generated [3][4]. The learning procedure takes place by inputting the data to be train by the network. The information from the input layer is spread to the hidden layer for information process. Then the output layer will extra process the information to obtain the results. The outputs are then compared with the preferred values for error computations.

The multi-layer perceptron neural networks (NN) is compared with logistic regression (LR), to recognize important covariates and their connections and to associate the designated variables with those regularly used in medical harshness of disease guides for breast cancer. The LR is selected as an acknowledged standard for forecast by biostatisticians in order to assess the NN. From the outcome it is found that the correctness of NN model is documented as influential means, when related with LR.

The authors identified the most and least important extrapolative issues for breast cancer survival analysis by means of feature evaluation indices derived from multilayer feedforward back propagation neural networks (MLJFBPNN), fizzy k-nearest neighbor classifier (FK-"V) and a logistic regression-based backward stepwise method (LR). The data used for the survival analysis were collected from 100 women who had been clinically diagnosed with breast disease in the form of carcinoma or benign conditions [6]. The data set consists of seven different histological and cytological prognostic factors and two corresponding outputs to be predicted. The result shows that each method identified a different set of factors as being the most important and therefore it is suggested that it could be inappropriate to rely on one method's outcome for assessment of the factors.

For prediction and diagnosis of various diseases with good accuracy Data Mining techniques are widely used. The two most successful data mining tools, Neural Networks and Genetic Algorithms are used for prediction of Diabetes disease. To initialize the Neural Network weights global optimization advantage of Genetic Algorithms is used. Using this technique the learning is faster, more stable and accurate While diagnosing a disease the patient has to undergo various tests which are costly and sometimes all the tests are not required. For automated detection of Diabetes is an intelligent and effective methodology is designed based on Neural Network. There exists many methods to diagnose Diabetes but the main drawback is that the patient has to undergo various tests. Using this method user can check whether he/she is suffering from Diabetes or not. Artificial Neural Networks is used to construct distributions to carry out plausible reasoning in the field of medicine. It describes a comparison between Multivariate Logistic Regression (MLR) and the Entropy Maximization Network (EMN) in terms of explicit assessment of their predictive capabilities. The EMN and MLR have been used to determine the probability of harboring lymph node metastases at the time of initial surgery by assessment of tumor based parameters. Both predictors were trained on breast cancer patient records. From the result it shows that the Maximum Entropy Estimation is an alternative to Multivariate Logistic Regression for analyzing small data sets of binary outcome data.

III. METHODOLOGY

The methodology followed in the proposed work is in figure 1. It consists of the following steps:

3.1 Data Collection

The Diabetes was selected from UCI Machine learning repository for this study. It is a trial of the entire Indian population gathered. The dataset comprised of 345 rows and seven different Columns. The class value was reported based on these parameters as either 1 or 0 to represent the Diabetes.

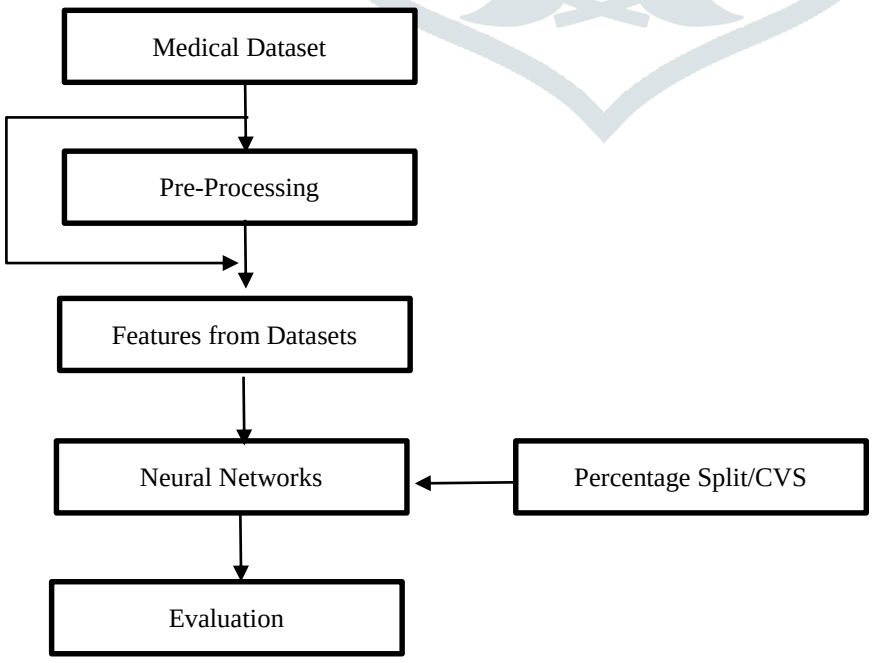

Figure 1: Proposed Frame work using neural network

3.2 **Pre-processing:**

Pre-processing is applied to standardize the missing values. The missing parameter along with their instances was replaced by 0 values.

3.3 **Randomization and splitting of dataset**

The features chosen in the earlier step were approved to develop classification model. The Dataset was randomized to obtain an random permutated sample. It was followed by dividing of the dataset into training and test sets.

3.4 **Classification algorithms**:

Different data mining algorithms like of NN were implemented in R platform for classification. R is a admired statistical computing structure for performing data mining experiments. A 10-fold cross validation is applied. The algorithms are briefly discussed below:

3.4.1 Neural Network**:** Neural Networks are a machine learning structure that attempts to mimic the learning pattern of natural neural networks. Biological neural structure have interconnected neurons with dendrites that get inputs, and then based on these inputs they create an output signal through an axon to another neuron. To create a neural network, we simply start to add layers of perceptions together, creating a multi-layer perceptron model of a neural network. We have an input layer which directly take in your feature inputs and an output layer which will create the outputs. Any layers in between are known as hidden layers because they don't directly "see" the feature inputs or outputs.

IV. RESULTS AND DISCUSSION

To compare the Accuracy of LR SVM RF and ANN model for diabetes data set using cross validation sample and percentage as test options. The stipulation of the datasets is as shown in table 1.

Sl. No	Dataset	Instances	Total attributes	classes
1	Diabetes	345	7	2

Table 1. Specification of medical datasets

The subsets of the feature obtained after applying the forward selection is shown in table 2.

For full attribute set of Diabetes dataset the classification accuracy attained is 84.52% by NN with K fold as shown in table 2.

Technique Used for Finding Classification Accuracy	**Percentage Split**				
	50%	**66%**	**70%**	**75%**	**80%**
NN	83.29	83.11	82.88	82.8	83.4
NN with 10 Fold	83.70	**84.52**	83.35	83.55	83.8

Table 2. Accuracy attained for full set of attributes.

The different subsets of features resulted after applying the forward selection method is shown in table 3.

Subset No.	**Subset of Attributes**	**No. of Attributes**
1	pres, outcome	2
2	pres, pedi, outcome	3
3	preg, pres, pedi, outcome	4
4	preg, pres, skin, pedi, outcome	5
5	preg, pres, skin, insu, pedi, outcome	6
6	preg, pres, skin, insu, pedi, age, outcome	7
7	preg, pres, skin, insu, mass, pedi, age, outcome	8

Table 3: Different subsets obtained after applying forward selection

The classification accuracy achieved for the different combinations of the feature is shown through table 4 through table 10.

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	77.6	77.61	77.7	78.13	**78.17**
NN with 10 Fold	77.13	77.62	77.23	77.7	77.56

Table 4. Accuracy attained for attributes pres

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	78.51	77.89	**77.99**	79.0	79.24
NN with 10 Fold	77.49	78.10	77.97	77.60	78.73

Table 5. Accuracy attained for attributes pres and pedi

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	78.76	78.14	77.55	78.3	78.44
NN with 10 Fold	78.44	79.47	78.58	79.05	78.60

Table 6. Accuracy attained for attributes pres, preg and pedi

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	79.05	76.98	77.57	78.68	78.84
NN with 10 Fold	78.20	78.82	78.9	78.64	78.94

Table 7. Accuracy attained for attributes pres, preg, pedi and skin

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	78.8	78.36	**76.68**	79.24	79.25
NN with 10 Fold	78.3	78.67	79.19	78.22	78.19

Table 8. Accuracy attained for attributes pres, preg, pedi, skin and insu

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	78.95	79.06	79.63	79.38	79.11
NN with 10 Fold	79.50	79.15	79.76	79.75	79.93

Table 9. Accuracy attained for attributes pres, preg, pedi, skin, insu and age

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	80.62	80.29	80.87	80.92	81.88

NN with 10 Fold	80.90	80.82	80.93	80.67	80.80

Table 10. Accuracy attained for attributes pres, preg, pedi, skin, insu, age and mass

The different subsets of features resulted after applying the forward selection method is shown in table 11.

Subset No.	Subset of Attributes	No. of Attributes
1	preg, plas, skin, insu, mass, pedi, age, outcome	8
2	preg, plas, skin, insu, mass, age, outcome	7
3	plas, skin, insu, mass, age, outcome	6
4	plas, insu, mass, age, outcome	5
5	plas, mass, age, outcome	4
6	plas, mass, outcome	3
7	plas, outcome	2

Table 11: Different subsets obtained after applying backward elimination

The classification accuracy achieved for the different combinations of the feature is shown through table 12 through table 18.

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	83.11	83.08	83.02	83.12	83.96
NN with 10 Fold	83.66	84.09	83.21	84.14	84.41

Table 12. Accuracy attained for attributes preg, plas, skin, insu, mass, pedi and age

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	82.70	82.48	82.30	82.29	82.70
NN with 10 Fold	83.73	83.78	84.06	83.99	**85.24**

Table 13. Accuracy attained for attributes preg, plas, skin, insu, mass and age

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	82.69	83.10	82.87	83.26	83.58
NN with 10 Fold	83.01	82.92	84.13	83.55	83.02

Table 14. Accuracy attained for attributes plas, skin, insu, mass and age

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	82.80	83.11	82.87	83.25	83.63
NN with 10 Fold	82.92	83.48	83.77	82.57	83.69

Table 15. Accuracy attained for attributes plas, insu, mass and age

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	82.76	83.23	82.96	83.43	81.88

NN with 10 Fold	83.50	83.63	83.52	84.15	82.87

Table 16. Accuracy attained for attributes plas, mass and age

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	82.75	82.96	82.84	83.22	83.66
NN with 10 Fold	83.30	83.14	82.70	83.56	83.69

Table 17. Accuracy attained for attributes plas, mass

Technique Used for Finding Classification Accuracy	Percentage Split				
	50%	66%	70%	75%	80%
NN	81.94	82.18	81.99	82.67	82.88
NN with 10 Fold	82.65	82.61	82.06	82.41	82.88

Table 18. Accuracy attained for attribute plas

V. CONCLUSION

The research work compares the different machine learning methods like NN and NN with 10 Fold for Diabetes dataset. From the experimental results it is recognized that for Diabetes dataset with NN along 10 Fold using split ratio of 66%, prediction accuracy **84.52%** is achieved.

Entropy features selection using backward elimination results with full set attribute accuracy is **84.41%** as show in table 12 and as show in table 13 for reduces set of attributes (Pregnancies, Glucose, Skin Thickness, Insulin, BMI, Age & Outcome) we getting accuracy of **85.24 % with NN 10 fold.**

VI. REFERENCES

[1]. Raghavendra B.K., Jay B. Simha, "Performance Evaluation of Logistic Regression and Neural Network Model with Feature Selection Methods and Sensitivity Analysis on Medical Data Mining", International Journal of Advanced Engineering Technology (Vol. II, Issue: I, January-March 2011), pp. 288-298.

[2]. Raghavendra B.K., S.K. Srivatsa, Raghavendra S, Shivashankar S.K., "Comparison of Logistic Regression and Neural Network Model with and without hidden Layers", Universal Journal of Applied Computer Science and Technology, Vol.1, 2011, pp. 49-53.

[3]. Raghavendra S, and Indiramma M., "Performance Evaluation of Logistic Regression and Artificial Neural Network Model with Feature Selection Methods using Cross Validation Sample and Percentage Split on Medical Datasets", Second International Conference on Emerging Research in Computing, Information, Communication and Applications (ERCICA-2014), August-2014.

[4]. Ankita Dewan and Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 2015, page(s): 704-706.

[5]. Sunita Soni, Ujma Ansari, Dipesh Sharma and JyotiSoni, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer application (0975-8887), vol. 17, no.8, March (2011).

[6]. Seker H., Odetayo M., Petrovic D., Naguib R.N.G., Bartoli C., Alasio L., Lakshmi M.S., and Sherbet G.V., "An Artificial Neural Network Based Feature Evaluation Index for the Assessment of Clinical Factors in Breast Cancer Survival Analysis", Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, pp. 1211-1215.