Thesis Report on

Spinal Cord Injured (SCI) patients Length of Stay (LOS) prediction based on admission data

Authored By:

Tabib Ibne Mazhar
ID: 17141041

Nusrat Jahan Suha
ID: 17341006

Department of Computer Science and Engineering,
BRAC University

Supervised by
Professor Dr. MD. Haider Ali
BRAC University

Co-Supervised by
Dipankar Chaki
BRAC University

Declaration

We declare that, this thesis report is our own work and has not been submitted for any other degree or professional qualifications. All sections of the paper that use quotes or describe an argument or concept developed by another author, have been referenced in the reference section.

Professor Dr. MD. Haider Ali
Supervisor

Tabib Ibne Mazhar
Author

Nusrat Jahan Suha
Author

Abstract:

In order to take better care and to ensure better facilities to the inpatients, predicting length of stay serves great importance. Since, the resources and the doctors are limited in the hospital especially in a developing countries like Bangladesh, it is quite difficult to provide proper healthcare to the inpatients. Not only because of limited hospital resources but also, it is difficult for the inpatients to bear the expense for a long period as well. In addition to that, if doctors can predict length of stay at the early stage of preadmission, they can map a well instructed way for example, which treatment, which instrument will treat patient best. As a result the patient can start his treatment with a slight assumption of the expenses. If we can predict accurate length of stay, patients do not have to leave in between the treatment without medical advice. Keeping all this point in mind, we decided to developed a study using machine learning algorithm and artificial neural network (ANN) to predict length of stay for Spinal Cord Injured (SCI) patients. For this purpose we chose Centre for the Rehabilitation of the Paralysed (CRP). They provided us around 500 inpatients data who has been released from the hospital after completing their treatment.

Acknowledgement

We are grateful to our supervisor, Professor Dr. MD. Haider Ali , for his immense support and valuable ideas throughout the work. He helped us greatly in his busy schedule and guided us. Without his guidance, it would have been impossible to work in the field of data mining and big data analysis.

We are also grateful to our co-supervisor Dipankar Chaki , who have helped us for the past 12 months and made time for us in his busy schedule. He supported us no matter what the situation was.

We are thankful to Nahid Quader from CSE dept, who has helped us with his expertise in python and AAN.

Nevertheless, we would like to express our gratitude to Department of Computer Science and Engineering, BRAC University and our teachers for helping us with all the necessary support.

Index

1. Introduction

Have you ever thought of a life that is confined to wheelchair for the whole life? Have you ever thought of that person, who has to depend on others for every movement? Well, that's the life of most of the people who is suffering from spinal cord injury. The spine consists of the spinal cord, which is a group of nerves that are protected by the individual vertebrae of the spine. The main function of the spinal cord is to send signals from the brain to other regions of the body. It is the main messenger throughout the body. A spinal cord injury (SCI) is damage to the spinal cord that causes changes in its function, either temporary or permanent.

These changes translate into loss of muscle function, sensation, or autonomic function in parts of the body served by the spinal cord below the level of the lesion. Injuries can occur at any level of the spinal cord and can be classified as complete injury, a total loss of sensation and muscle function, or incomplete, meaning some nervous signals are able to travel past the injured area of the cord. Depending on the location and severity of damage along the spinal cord, the symptoms can vary widely, from pain or numbness to paralysis to incontinence. The prognosis also ranges widely, from full recovery in rare cases to permanent tetraplegia (also called quadriplegia) in injuries at the level of the neck, and paraplegia in lower injuries. Complications that can occur in the short and long term after injury include muscle atrophy, pressure sores, infections, and respiratory problems.

The effect of spinal cord injury depend on can be divided into two section, complete and incomplete. A complete injury means there is no function below the level of the injury – no sensation and no voluntary movement. Both sides of the body are equally affected. In terms of incomplete injury, people can move parts of their body. They can feel their paralyzed body parts which cannot be moved or people can move one parts of their body than others.

In bangladesh in 4 people in every 5 In August, 2009, randomly selected 154 people who died because of SCI from 1979 to 1999, were analyzed retrospectively and showed that most of them around 91% was men and mean age of sustaining injury was 33.85 years. Moreover 86.1% injuries was traumatic in organ. Only 16% survived beyond 10 years and 54.4% died within 5 years[14].

Why this study is important? Mainly this study is important in order to allocate the resources efficiently. Since there is a close relation between hospital resources and patient's length of stay at the hospital [13,14]. In japan, a comparison study has made which says that, inpatients capacity and the ratio of involuntary admissions positively correlated with the greater number of LOS [13]. Again, guardian association such as problem management, communication, cooperation among nurse and physicians are greatly correlated with low LOS [14].

We proposed to develop a study to predict the Length of Stay for Spinal Cord Injured patients Using Artificial Neural Network. We used both multiple linear regression and ANN for this study and both of them was able to predict (LOS). For ANN the accuracy rate was between 70-80% for cross validation binary output. For 3 category target class, it was near 70%.

1.1.Overview

Why the study was done?

- Spinal cord injury is one of the most dangerous diseases which leads to cause various other diseases.
- Prediction of length of stay results in efficient resources allocation
- In Bangladesh, this is the first study which has been developed to predict length of study

What it will do?

- It will predict the length of stay of patients.
- It will predict whether the length of stay of an inpatient will be less than 100 days or not.

What are the findings?

- We have found both MLR, ANN and SVM are able to predict the LOS but accuracy level fluctuated a lot.

2. Literature review

[4] This study has been developed for predicting future crime trend in Bangladesh. They have developed linear regression model for this purpose. To find out all the linear regression parameters, gradient descent algorithm has been used. In this study, linear regression performance is outstanding. The dataset has been divided into two portion, one is metropolitan region and other is divisional region. For both portions, they have analyzed the prediction for individual crime. In the conclusion it has been proved in their study that, with the growing population, the number of crime is also on increase.

[3] They proposed to predict the Length of Stay for cardiology patients analysing the data before admitting him/her at the hospital. The purpose was to develop an Artificial Neural Network(ANN) for prediction using three primary diagnosis, coronary atherosclerosis (CAS), heart failure (HF), and acute myocardial infarction (AMI). Pearson correlation coefficient has been used to visualize the relation between data. But the outcome suggested to use another model. After that, ANN has been used. With compare to logistic regression, the result using ANN either. It has been observed for CAS patients, the accuracy for this models is ranged from 88.07% to 91.53%. Again Linear regression's performance is better than ANN with LOS of 2 days. On the contrary, with the LOS of more than 5 days it was unable to forecast the LOS accurately for both Linear regression and ANN . For AMI and HF patients. The models was not as effective as predicting LOS for CAS patients. With the LOS of 3,5,6,7 days, ANN was able to predict more accurately than LR. On the other hand who has a LOS of more than 18 days, both LR and ANN was not able to produce the prediction accurately.

[6] They proposed to predict length of stay for acute psychiatric patients and they also proposed to find, whether negligence is reason for longer length of stay. They divided the data into two parts equally. From the first one they derived the function for LOS prediction and with the second group, they worked for searching the disorganizations. To estimating the effect of each variable individually on Length of stay, Multiple linear regression has been applied and also for predicted LOS. an assurance coefficient war acquired which indicates the qualification for the next step. After that this function has used to predict the LOS of the second group. The difference of LOS-PLOS was secure as well. They developed a Receiver

Characteristic (ROC) curve to examine new difference of LOS-PLOS for predicting the inefficiency issues.

[9] They aimed to develop a study to forecast gold price based on available economic factors. They used linear regression to develop a model. They made 4 model, first one is developed by gathering all related factor in account. After that stepwise regression is applied for reducing the number of variable and solving the multicollinearity problem as well.then model 3 was developed consisting the lagged variable and removed the variables having coefficients which are indifferent from zero. Lastly Prais-Winsten procedure was followed to re calculate the coefficient and estimate model 4. Lasty the lagged variable model which has smaller MSPR was selected. For comparing the models and to derive the output, "forecast-1" was used which is a naive process.

We have also study some other papers related to LOS prediction [7,10,14], most of them they have used Multiple linear regression. We have encouraged from paper [3] and [9]. We have processed our data according to [3]. In addition, we have also followed their working methodology.

3. Data set

3.1 Data collection:

Collecting data was one of the most difficult work. In a country like bangladesh it's too difficult to collect. For our model, we targeted the patients who have suffering spinal cord injury. In order to collecting patient's details we have selected rehabilitation centre named Centre for the Rehabilitation of the Paralysed (CRP). CRP is kind of a association who provides care, assessment and help based on their mental, financial and physical condition. In our country their is no other organization like CRP. For that purpose we collected about data from october 2015 to july 2017, containing around 500 patients details. In order to collect data we had to provide several application to CRP. It took us about 2 months for the approval. Finally e are provided with their admission and discharged details. We got around 500 patient's details, though we have cut them in numbers.

3.2 Data description:

Our data set contains around 500 patient's data. Currently we are working with the discharged people's data. In order to build a model which will be suitable for our work. We have considered 13 attributes which are described below them in short,

1.Our first two parameters are age and sex.

2.Afterwards our next attribute is the time difference between patients injury date and admission date. Because it's really important to know how many days he had delayed to start his treatment.

3.Our next predicted value is P.sore. P . sore— also called pressure ulcers and decubitus ulcers — are injuries to skin and underlying tissue resulting from prolonged pressure on the skin. P. sores most often develop on skin that covers bony areas of the body, such as the heels, ankles, hips and tailbone... If any SCI patient has P.sore then it usually takes longer time to recover. Consequently, their length of the stay gets higher.

4. Diagnosis. 4 types of Diagnosis. Tt, Tp, Ntt and Ntp.

- Tp is paralysis of the legs and lower body, typically caused by spinal injury or disease.
- Tt or Tetraplegia, also known as quadriplegia, is paralysis caused by illness or injury that results in the partial or total loss of use of all four limbs and torso.

5.Then comes ASIA. ASIA means, American Spinal Injury Association by whom spinal cord has been classified and the levels are determined. Its classification system is divided into 5 points on a scale of A to E. It explains its one patient's ASIA is E, meaning that his bones are in good condition and he can function on his own. Again, if someone is rated with ASIA A, indicates he is almost disable and paralysed. ASIA can be categorized into two types,

- Admission ASIA
- Discharge ASIA

If a patient's ASIA is denoted by A at the time of admission and if he is found with ASIA B, C or D, it explained that the patient has improved. In the contrary if his/her ASIA if found in descending order, for example when a patient admits with ASIA E and discharged with ASIA A, B or C, it indicates that the his/her condition has been deteriorated.

6.SCIM

7. N/L or neurological level. We got both admission N/L and discharge N/L.

8. S/L or skeletal level.

9. LOS or length of stay.

We have used Linear regression to determine the relation between the data and predicting the LOS for a random patient. Currently we are using different number of classifier and trying to figure out their accuracy level and which classifier is better for the dataset. American Spinal Injury Association by whom spinal cord has been classified and the levels are determined [2].

3.3 Data modification:

In order to modify this data set to our model we had to remove numbers of patient's data. For example, the inpatients who has been referred to somewhere else or inpatients who has died or who has re admitted. So, patients are excluded in this study who are discharged willingly against medical advice, in order to additional rehabilitation service, inpatients who has been referred to another medical center since their details do not lead to any result as their treatment has n stopped in the middle [8]. In this case we have found 75 number of referred inpatient, people has stopped the treatment willingly, people has been died. So we had to remove all those data. After that, we have normalized our dataset by dividing all values with the highest value to create a workable environment with our model.

4. Working Methodology

Predictors	V13	Sig
v1	0.007	0.887
v2	-.137**	0.008
v3	-0.009	0.867
v4	0.014	0.793
v5	-.389**	0
v6	-.435**	0
v7	-.335**	0
v8	-.112*	0
v9	0.01	0.854
v10	-0.033	0.53
v11	-.301**	0
v12	-.417**	0

Table number 1: Pearson correlation between predictor and LOS

Pearson correlation was used to learn about the relations between the length of stay and each predictors. And the correlation has been summarized in the table. In table-1, we can observe that, most of the predictors are highly correlated with the target class. Before applying linear regression we plotted histogram graph as well to check whether the data are normally distributed or not. From figure-1, we can observe that our graph has positive skewness and It is not normally distributed. Since linear regression does work better for data which are not normally distributed, here it can be applied.

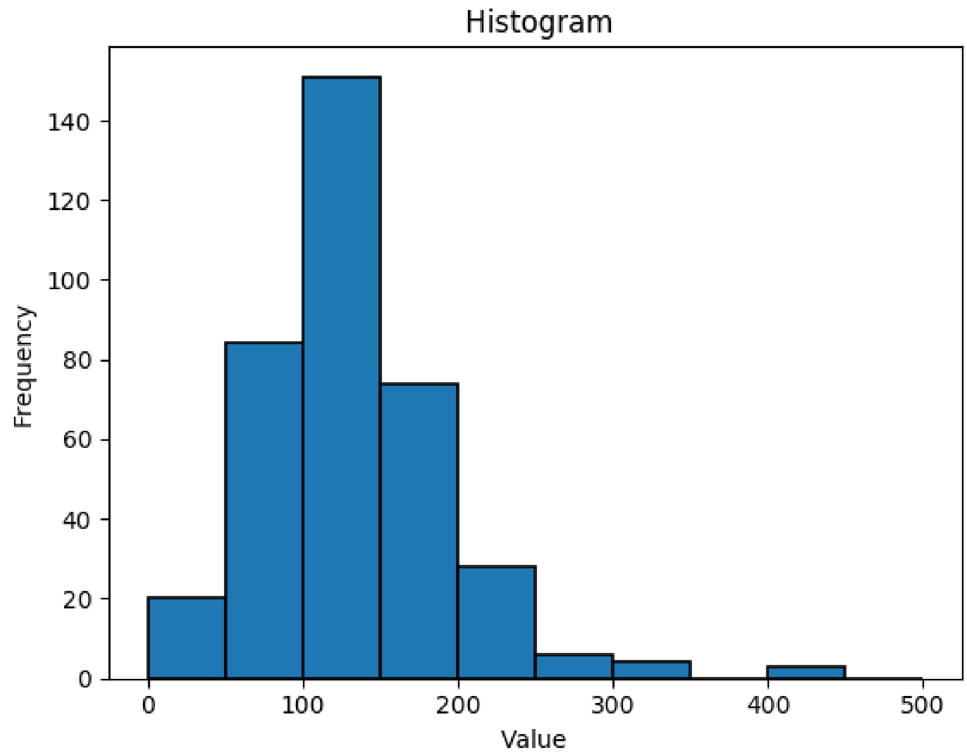


Figure1 : Histogram of the data

4.1 Multiple Linear regression (MLR):

Linear regression predicts the relationship between the predictors who have cause and effect relation. Usually there is one dependent variable, which depends on other independent variable. If one or more conditions causes a certain event, linear regression is a suitable for that study. In other words it can be said, linear regression is preferable for prediction study. The main concentration of linear regression study is to research the relation between dependent and independent variable and to produce a linear equation which defines the relation between them[11]. When one dependent predictor depends on more than a one independent predictor, the study is called Multiple Linear regression(MLR). Multiple linear Regression hypothesis can be expressed as,

$$y = \beta_0 + \beta_{1x_1} + \beta_{2x_2} + \ldots \beta_{nx_n}$$

Where the $\beta_0, \beta_1, \beta_2 \ldots\ldots \beta_n$ is regression parameter, x is predictor and y is predicted value. Since we wanted to develop a prediction model, MLR is chosen first as MLR most preferable for prediction model.

For this study we developed 2 types of model one is developed after classifying data and another one is developed with the unclassified data. Once we trained the linear regression with classified data the our regression model is,

y=2.75+ (-.026)* x1 + (-0.311)* x2 + (-0.036)* x3 + 0.149 * x4 + (-.407)* x5 + (- 1.032) * x6 + (0.155) * x7+ (-.088)*x8+ 0.061*x9+ (-0.159)*x10+ (-.904)*x11+ (-.762)*x12.

Here, the regression parameters are β_0=2.75, β_1= -.026, β_2= -0.311, β_3= -0.036, β_4=0.149, β_5=-.407, β_6 =- 1.032, β_7 = 0.155, β_8 = -.088, β_9=0.061, β_{10} = -0.159, β_{11} = -.904, β_{12} = -.762.

For this model the value of adjusted R squared is .34. Which defines that, 34% of the predictors explain this model. After that, another model was developed with the unclassified data and our regression model is,

y= 276.127+ (.087)*x1+ (-25.618)* x2 + (-.004)* x3 +4.179 * x4 + (-31.895)* x5 + (-17.196) * x6 + (3.365) * x7+ (-.677)*x8+ 0.533*x9+ (-.495)*x10+ (-.590)*x11+ (-.518)*x12.

Where the regression parameters are, β_0=276.127, β_1= .087, β_2= -25.618, β_3= -.004, β_4= 4.179, β_5= -31.895, β_6 = -17.196, β_7 =3.365, β_8 = -.677, β_9=0.533, β_{10} = -.495, β_{11} = -.590, β_{12} =-.518.

The adjusted R squared value for this model is .33 which refers that 33% of the variables, explain the model. These parameters are used to predict the Length of Stay.

4.2 Artificial Neural Network(ANN):

An Artificial Neural Network (ANN) is a computational model that is inspired by the way biological neural networks in the human brain process information. Artificial Neural Networks have generated a lot of excitement in Machine Learning research and industry. The basic unit of computation in a neural network is the neuron, often called a node or unit. It receives input from some other nodes, or from an external source and computes an output.

Each input has an associated weight (w), which is assigned on the basis of its relative importance to other inputs. The node applies a function *f* (defined below) to the weighted sum of its inputs as shown in in Figure 1 below:

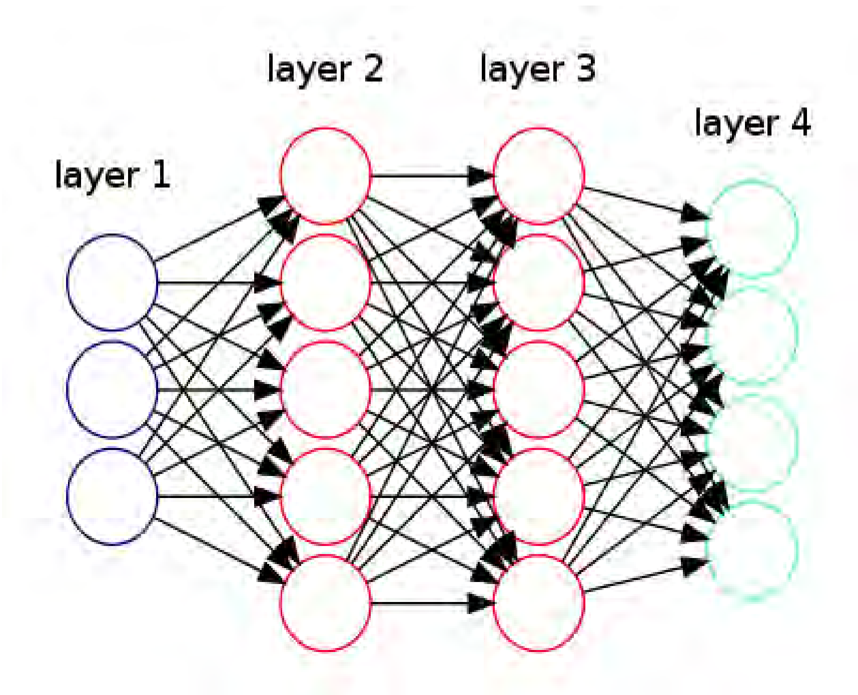


Figure 2: Layers of Neural Network

The above network takes numerical inputs X1 and X2 and has weights w1 and w2 associated with those inputs. Additionally, there is another input 1 with weight b (called the Bias) associated with it. t Y from the neuron is computed as shown in the Figure 1. The function f is nonlinear and is called the Activation Function. The purpose of the activation function is to introduce nonlinearity into the output of a neuron. This is important because most real world data is nonlinear and we want neurons to learn these nonlinear representations. The feedforward neural network was the first and simplest type of artificial neural network devised [3]. It contains multiple neurons (nodes) arranged in layers. Nodes from adjacent layers have connections or edges between them. All these connections have weights associated with them. An example of a feedforward neural network is shown in Figure,

A feedforward neural network can consist of three types of nodes:

Input Nodes: The Input nodes provide information from the outside world to the network and are together referred to as the “Input Layer”. No computation is performed in any of the Input nodes – they just pass on the information to the hidden nodes.

Hidden Nodes : The Hidden nodes have no direct connection with the outside world (hence the name “hidden”). They perform computations and transfer information from the input nodes to the output nodes. A collection of hidden nodes forms a “Hidden Layer”. While a feedforward network will only have a single input layer and a single output layer, it can have zero or multiple Hidden Layers.

Output Nodes : The Output nodes are collectively referred to as the “Output Layer” and are responsible for computations and transferring information from the network to the outside world.

TensorFlow:Tensor Flow is an open-source software library for machine learning across a range of tasks, and developed by Google to meet their needs for systems capable of building and training neural networks to detect and decipher patterns and correlations, analogous to the learning and reasoning which humans use. It is currently used for both research and production at Google, often replacing the role of its closed-source predecessor, DistBelief. TensorFlow was originally developed by the Google Brain team for internal Google use before being released under the Apache 2.0 open source license on November 9, 2015.

Keras: Keras is an open source neural network library written in Python. It is capable of running on top of MXNet, Deep Learning, Tensorflow, CNTK or Theano. Theano: Theano is a numerical computation library for Python. In Theano, computations are expressed using a NumPy-esque syntax and compiled to run efficiently on either CPU or GPU architectures.

Activation function:

Softmax: In mathematics, the softmax function, or normalized exponential function is a generalization of the logistic function that "squashes" a K-dimensional vector z of arbitrary real values to a K-dimensional vector σ(z) of real values in the range [0, 1] that add up to 1. The function is given by

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, \ldots, K.$$

for j = 1, …, K.

In probability theory, the output of the softmax function can be used to represent a categorical distribution – that is, a probability distribution over K different possible outcomes. In fact, it is the gradient-log-normalizer of the categorical probability distribution.

The softmax function is used in various multiclass classification methods, such as multinomial logistic regression, multi class linear discriminant analysis, naive Bayes classifiers, and artificial neural networks. Specifically, in multinomial logistic regression and linear discriminant analysis, the input to the function is the result of K distinct linear functions, and the predicted probability for the j th class given a sample vector x and a weighting vector w is:

$$P(y = j \mid \mathbf{x}) = \frac{e^{\mathbf{x}^\mathsf{T}\mathbf{w}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}^\mathsf{T}\mathbf{w}_k}}$$

This can be seen as the composition of K linear functions

$\mathbf{x} \mapsto \mathbf{x}^\mathsf{T}\mathbf{w}_1, \ldots, \mathbf{x} \mapsto \mathbf{x}^\mathsf{T}\mathbf{w}_K$ and the softmax function (where $\mathbf{x}^\mathsf{T}\mathbf{w}$ denotes the inner product of x and w). The operation is equivalent to applying a linear operator defined by w to vectors x, thus transforming the original, probably highly-dimensional, input to vectors in a K-dimensional space R^K..

The softmax function is often used in the final layer of a neural network-based classifier. Such networks are commonly trained under a log loss (or cross-entropy) regime, giving a nonlinear variant of multinomial logistic regression. Since the function maps a vector and a specific index i to a real value, the derivative needs to take the index into account:

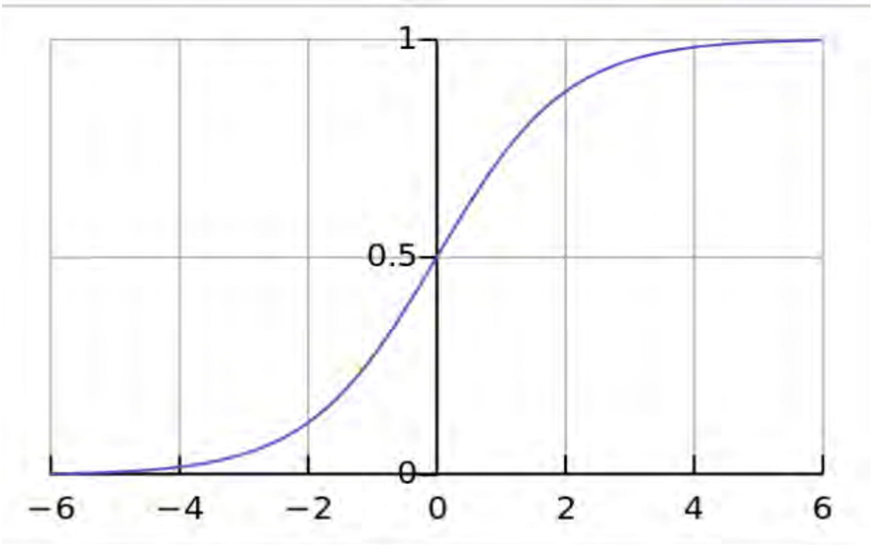


Figure 3: softmax function

$$\frac{\partial}{\partial q_k}\sigma(\mathbf{q}, i) = \cdots = \sigma(\mathbf{q}, i)(\delta_{ik} - \sigma(\mathbf{q}, k))$$

Here, the Kronecker delta is used for simplicity (cf. the derivative of a sigmoid function, being expressed via the function itself).

Sigmoid function:

A sigmoid function is a bounded differentiable real function that is defined for all real input values and has a non-negative derivative at each point.

A sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid curve. Often, sigmoid function refers to the special case of the logistic function shown in the first figure and defined by the formula

$$S(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{e^x + 1}.$$

Other examples of similar shapes include the Gompertz curve (used in modeling systems that saturate at large values of x) and the ogee curve (used in the spillway of some dams). Sigmoid functions have domain of all real numbers, with return value monotonically increasing most often from 0 to 1 or alternatively from −1 to 1, depending on convention.

A wide variety of sigmoid functions have been used as the activation function of artificial neurons, including the logistic and hyperbolic tangent functions. Sigmoid curves are also common in statistics as cumulative distribution functions (which go from 0 to 1), such as the integrals of the logistic distribution, the normal distribution, and Student's t probability density functions.

Relu:

In the context of artificial neural networks, the rectifier is an activation function defined as:

$$f(x) = x^{+} = \max(0, x),$$

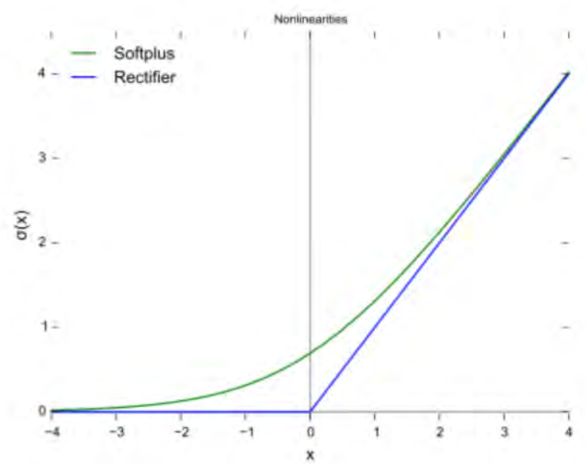


Figure 4: Relu function

where x is the input to a neuron. This is also known as a ramp function and is analogous to half-wave rectification in electrical engineering. This activation function was first introduced to a dynamical network by Hahnloser et al. in a 2000 paper in Nature with strong biological motivations and mathematical justifications.[2] It has been used in convolutional networks[3] more effectively than the widely used logistic sigmoid(which is inspired by probability theory; see logistic regression) and its more practical[4] counterpart, the hyperbolic tangent. The rectifier is, as of 2015, the most popular activation function for deep neural networks.[5]

A unit employing the rectifier is also called a rectified linear unit (ReLU).

A smooth approximation to the rectifier is the analytic function

$f(x)=\ln[1+\exp(x)]$,

which is called the softplus function.[7] The derivative of softplus is $f'(x)=\exp(x)/[\exp(x)+1]=1/[1+\exp(-x)]$, i.e. the logistic function.

Rectified linear units find applications in computer vision[3] and speech recognition[8][9]using deep neural nets.

Elu:

Exponential linear units try to make the mean activations closer to zero which speeds up learning. It has been shown that ELUs can obtain higher classification accuracy than ReLUs.

$$f(x) = \begin{cases} x & \text{if } x >= 0 \\ a[\exp(x) - 1] & \text{otherwise} \end{cases}$$

a is a hyper-parameter to be tuned and a>=0 is a constraint.

Optimizer:

Adagrad: It simply allows the learning Rate -η to adapt based on the parameters. So it makes big updates for infrequent parameters and small updates for frequent parameters. For this reason, it is well-suited for dealing with sparse data.

It uses a different learning Rate for every parameter θ at a time step based on the past gradients which were computed for that parameter.

Previously, we performed an update for all parameters θ at once as every parameter θ(i) used the same learning rate η. As Adagrad uses a different learning rate for every parameter θ(i) at every time step t, we first show Adagrad's per-parameter update, which we then vectorize. For brevity, we set g(t,i) to be the gradient of the loss function w.r.t. to the parameter θ(i) at time step t .

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i}.$$

Adagrad modifies the general learning rate η at each time step t for every parameter θ(i) based on the past gradients that have been computed for θ(i). The main benefit of Adagrad is that we don't need to manually tune the learning Rate.Most implementations use a default value of 0.01 and leave it at that. AdaDelta:

It is an extension of AdaGrad which tends to remove the decaying learning Rate problem of it. Instead of accumulating all previous squared gradients, Adadelta limits the window of accumulated past gradients to some fixed size w.

Instead of inefficiently storing w previous squared gradients, the sum of gradients is recursively defined as a decaying mean of all past squared gradients. The running average $E[g^2](t)$ at time step t then depends (as a fraction γ similarly to the Momentum term) only on the previous average and the current gradient.

$E[g^2](t)=\gamma.E[g^2](t-1)+(1-\gamma).g^2(t)$, We set γ to a similar value as the momentum term, around 0.9.

$\Delta\theta(t)=-\eta\cdot g(t,i).$

$\theta(t+1)=\theta(t)+\Delta\theta(t).$

$$\Delta\theta_t = -\frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t.$$

As the denominator is just the root mean squared (RMS) error criterion of the gradient, we can replace it with the criterion short-hand:

$$\Delta\theta_t = -\frac{\eta}{RMS[g]_t} g_t.$$

The final formula for Parameter Updates.

Another thing with AdaDelta is that we don’t even need to set a default learning Rate .

Adam:

Adam stands for Adaptive Moment Estimation. Adaptive Moment Estimation (Adam) is another method that computes adaptive learning rates for each parameter. In addition to storing an exponentially decaying average of past squared gradients like AdaDelta ,Adam also keeps an exponentially decaying average of past gradients M(t), similar to momentum:

M(t) and V(t) are values of the first moment which is the Mean and the second moment which is the uncentered variance of the gradients respectively.

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}.$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}.$$

The formulas for the first Moment(mean) and the second moment (the variance) of the Gradients

Then the final formula for the Parameter update is —

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t.$$

The values for β1 is 0.9 , 0.999 for β2, and (10 x exp(-8)) for ϵ.

Adam works well in practice and compares favorably to other adaptive learning-method algorithms as it converges very fast and the learning speed of the Model is quiet Fast and efficient and also it rectifies every problem that is faced in other optimization techniques such as vanishing Learning rate , slow convergence or High variance in the parameter updates which leads to fluctuating Loss function

4.3 Support Vector Machine:

In machine learning, support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

5.Results

Results of both linear regression and Artificial Neural Network has been discussed in this section. These methods were run to forecast the length of stay of spinal cord injured patient. Initially it was planned to apply Artificial neural network only, but later on we decided to use Multiple Linear regression(MLR), Artificial Neural Network(ANN) and Support Vector Machine(SVM).

5.1 Multiple linear regression:

For MLR, same data set was used for both training model and testing. The Mean Absolute

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\boldsymbol{x}_i - \boldsymbol{x}|$$

Error(MAE) was used to find out the error. The formula for MAE is,

where Xi and X are the predicted LOS and actual LOS for the ith test data, i= 1,2……….n and n is the number of testing instances. The calculated value of MAE for this model is 43.97. We have used the regression equation to predict some of patient's length of stay. And some

y1=276.127+ (.087)* 55 + (-25.618)* 1+ (-.004)* 15+4.179 * 2 + (-31.895)* 1+ (-17.196) * 1+ (3.365) * 1+ (-.677)*9+ 0.533*9+ (-.495)*9+ (-.590)*11+ (-.518)*28= 191.221~191

y2=276.127+ (.087)* 35 + (-25.618)* 1+ (-.004)* 28+4.179 * 1 + (-31.895)* 1+ (-17.196) * 1+ (3.365) * 1+ (-.677)*24+ 0.533*24+ (-.495)*24+ (-.590)*15+ (-.518)*68= 152.48~153

y3=276.127+ (.087)* 35 + (-25.618)* 2+ (-.004)* 33+4.179 * 2 + (-31.895)* 2+ (-17.196) * 2+ (3.365) * 4+ (-.677)*25+ 0.533*8+ (-.495)*7+ (-.590)*10+ (-.518)*85= 85.38~85

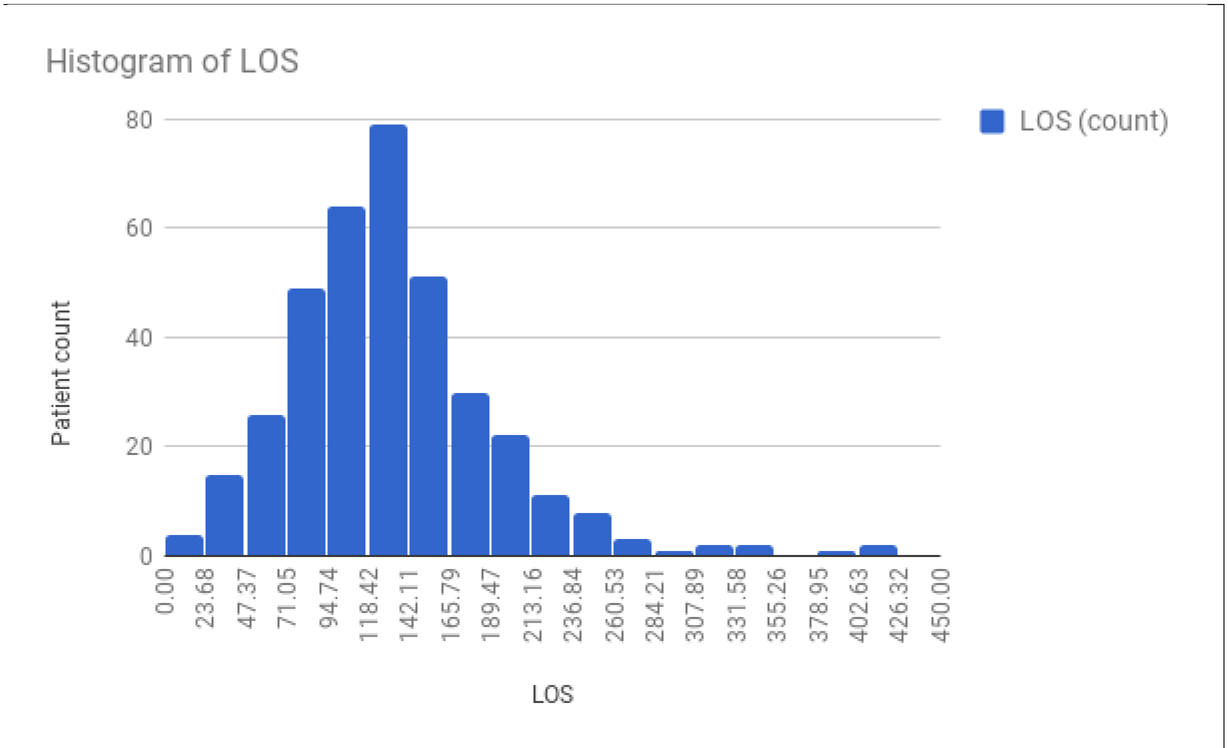
y4=276.127+ (.087)* 32 + (-25.618)* 1+ (-.004)* 104+4.179 * 1+ (-31.895)* 1+ (-17.196) * 1+ (3.365) *1+ (-.677)*21+ 0.533*22+ (-.495)*18+ (-.590)*14+ (-.518)*5= 161.621~163

Now we can compare these values with the actual length of In table-****, we have compare these values with their actual values. In addition we have added some more predicted LOS corresponded to -their actual LOS,

Table 2:Comparison between actual LOS and predicted LOS.

Now, if we observe the table-2, we can see the model is able to predict more accurately when it is around the concentrated area. To be more clear about the concentrated area see in the the figure:5,

Number of instance	Actual LOS	Predicted LOS
y1	19	63
y2	55	87
y3	77	85
y4	109	66
y5	121	107
y6	118	173
y7	134	111
y8	154	161
y9	136	131
y10	177	154
y11	158	153
y12	169	191
y13	144	132
y14	244	163



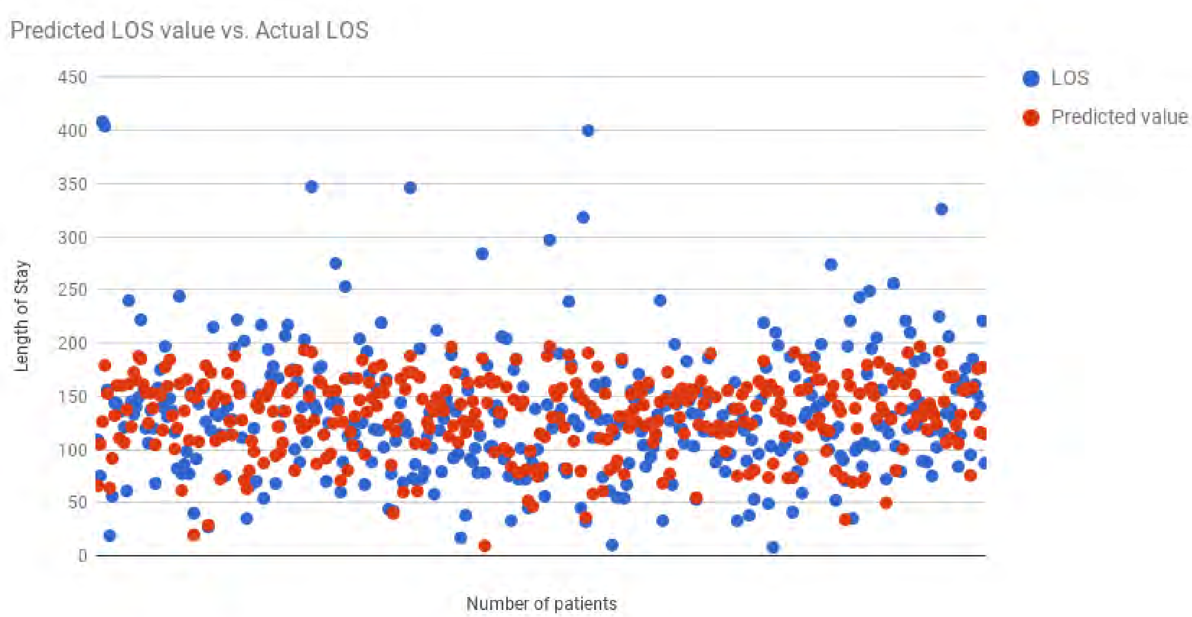


Figure 6: Scatter plotting of actual LOS and predicted LOS against the number of patients.

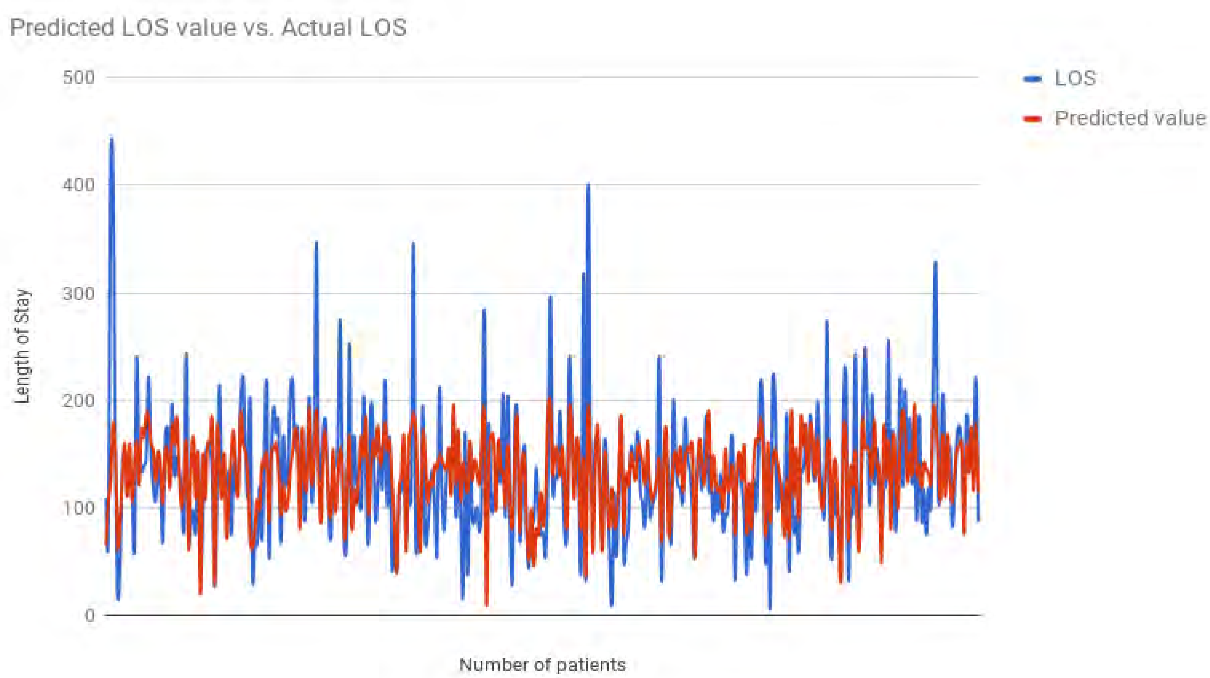


Figure 7: Scatter plotting of actual LOS and predicted LOS against the number of patients.

5.2 Artificial Neural Network:

For all the cases in neural network, we have used the label encoder to encode the output class. We have used the keras classifier to classify the model. We have divided the whole dataset into training set and testing set. The testing set was standard 33% of the data and training set was 66%. The testing set was chosen randomly. We have used random function to do that. We used fit function to fit the data for the better output. After that we have used function named estimator.predict() to predict the output.

When the target class was divided into 4 categories the following results were found,

input activation function	Input node number	output activation function	optimizer	epoch	batch_size	Accuracy(%)	Standard Deviation (%)
Relu	21	sigmoid	adam	200	8	48	2.90
Relu	21	softmax	ada grad	200	8	48.64	3.31
Relu	31	softmax	rmsprop	200	8	47.30	1.15
Relu	13	softsign	rmsprop	200	8	50.53	3.83
Relu	21	softmax	nadam	200	8	46.49	0.20

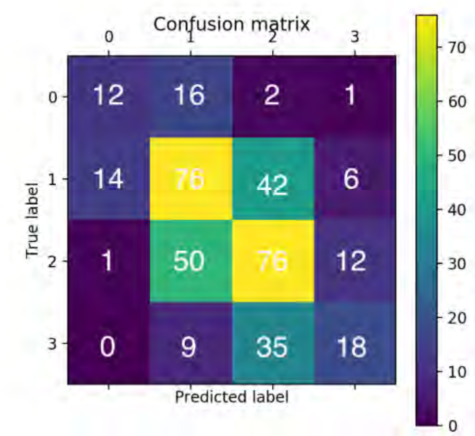


Table 3: 4 catagories target class output.

Figure 8: Plotting of the confusion matrix with accuracy of 50.53%

Here we have used the rmsprop as the optimizer with increased the accuracy level but was not more than 50%.

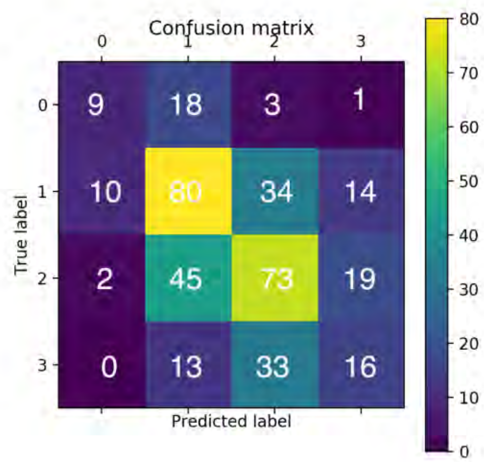


Figure 9: Plotting of the confusion matrix with accuracy of 48%

Binary Output:

When we tried to predict the binary output, if the patient's LOS is greater than 100 or not, the following things we found

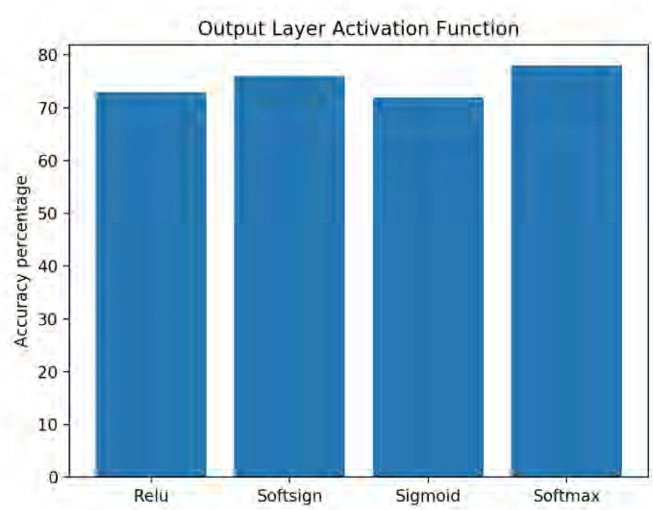


Figure 10: Output layer activation function accuracy.

From the above figure we can see that the softmax activation function works the best in the binary classifier.

Softmax=78%

Softsign=77%

Relu=74%

Sigmoid=72%

input activation function	Input layer Node	output activation function	optimizer	epoch	batch_ size	Accuracy(%)	Standard Deviation
Relu	21	sigmoid	nadam	200	8	78.53	4.55
Relu	31	softmax	nadam	200	8	80.28	2.91
Relu	21	sigmoid	nadam	200	8	79.47	4.35
Relu	21	softmax	adagrad	200	8	78.52	3.09
Relu	21	softmax	adamax	200	8	77.98	2.11

Table 4: binary target class output

In the above table, we can see that softmax activation function and nadam optimizer works best and gives the best accuracy.

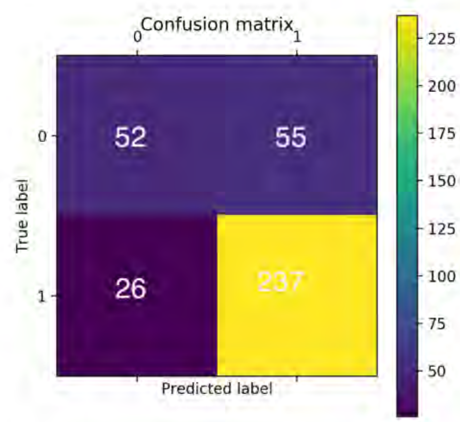


Figure 11: Plotting of the confusion matrix with accuracy of 79%

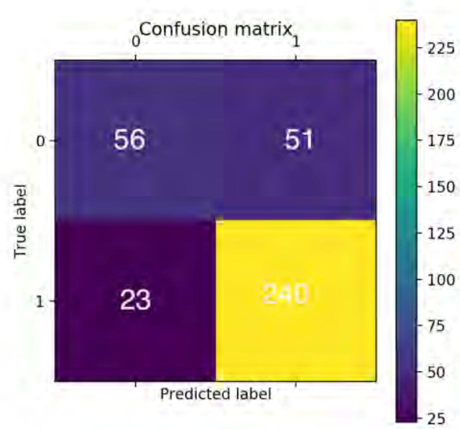


Figure 13: Plotting of the confusion matrix with accuracy of 80%

3 category target class: We have divided the target class into three categories. The categories are from 0- 100, 101-200 and 201 to the rest.

input activation function	Input layer Node	output activation function	optimizer	epoch	batch size	Accuracy(%)	Standard Deviation
Relu	21	softmax	rmsprop	200	8	67.30%	(2.58%)
Relu	31	softmax	nadam	200	8	67.30%	(7.09%)
Relu	21	softmax	adadelta	200	8	69.20%	(4.69%)
Relu	21	softmax	adagrad	200	8	66.77%	(5.32%)
Relu	21	softmax	adamax	200	8	68.67%	6.99%

Table 5:binary target class output

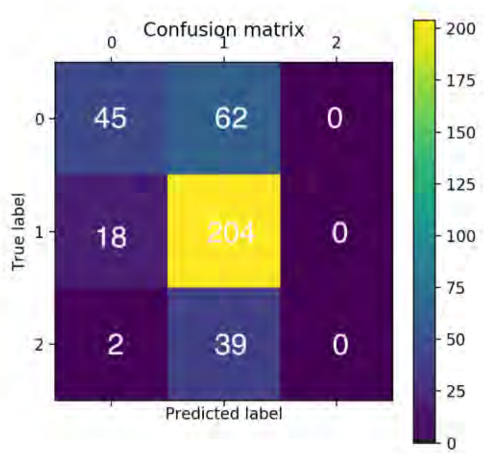


Figure 14: Plotting of the confusion matrix with accuracy of 67%

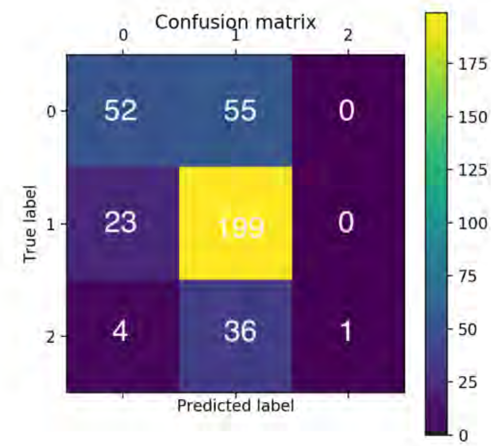


Figure 15: Plotting of the confusion matrix with accuracy of 67%

For the above two matrices the accuracy level is almost same but the the later matrix predicted one value from the third category which is better than the previous matrix. The only difference was the optimizer. RMSprop worked better than the nadam optimizer in the above case.

5.2.2 Binary target class(126 days):

We again divided out target class into two categories but this time we divided the length of stay less than 126 days and greater than 126 days. This made equal distribution of target class and helped to create better model.

input activation function	Input layer Input Node	output activation function	optimizer	epoch	batch _size	Accuracy (%)	Standard Deviation
Relu	21	softmax	rmsprop	200	8	69.30%	2.91
Relu	31	softmax	nadam	200	8	65.30%	2.7
Relu	21	softmax	adadelta	200	8	70.00%	1.42

Table 6: binary target class (126) output table.

In the above table we can clearly see that adadelta works better than any other optimizer though rmsprop came very close. Standard deviation is almost the same in the above cases.

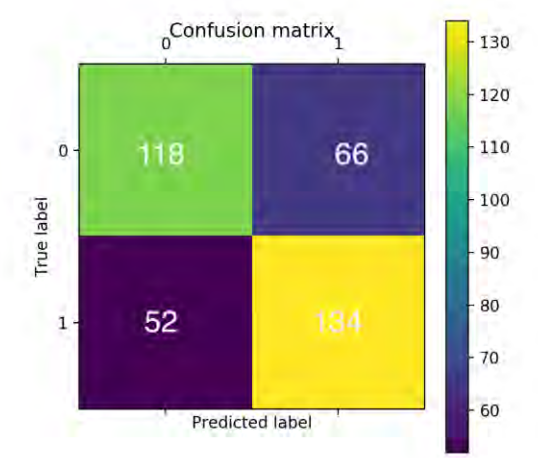


Figure 16: Plotting of the confusion matrix with accuracy of 68%

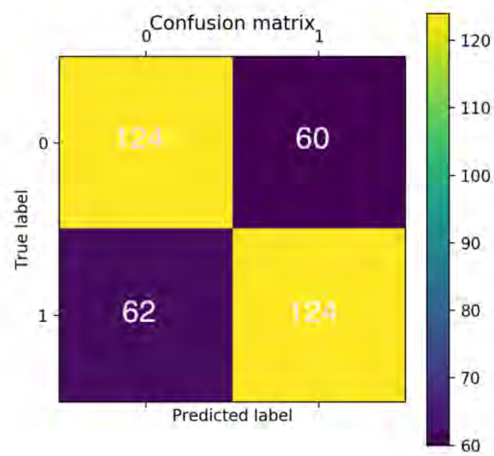


Figure 17: Plotting of the confusion matrix with accuracy of 67%

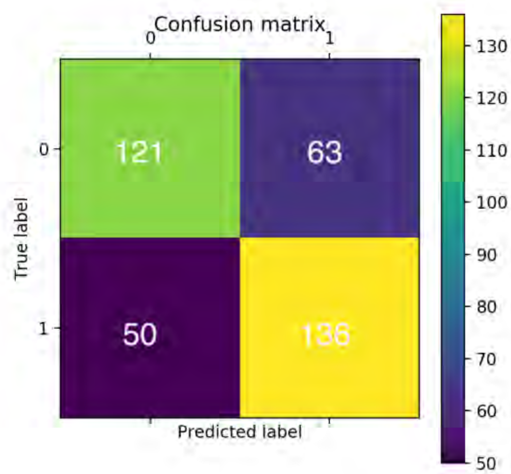


Figure 18: Plotting of the confusion matrix with accuracy of 70%

For the above three matrices the accuracy rate is almost the same but the optimizer adadelta worked best in the above scenario.

5.3 SVM:

For evaluating SVM results , we have used the python library. SVM shows almost similar results of artificial neural network. Binary output shows the highest accuracy , 80% whereas 4 category target class shows accuracy of 44%.

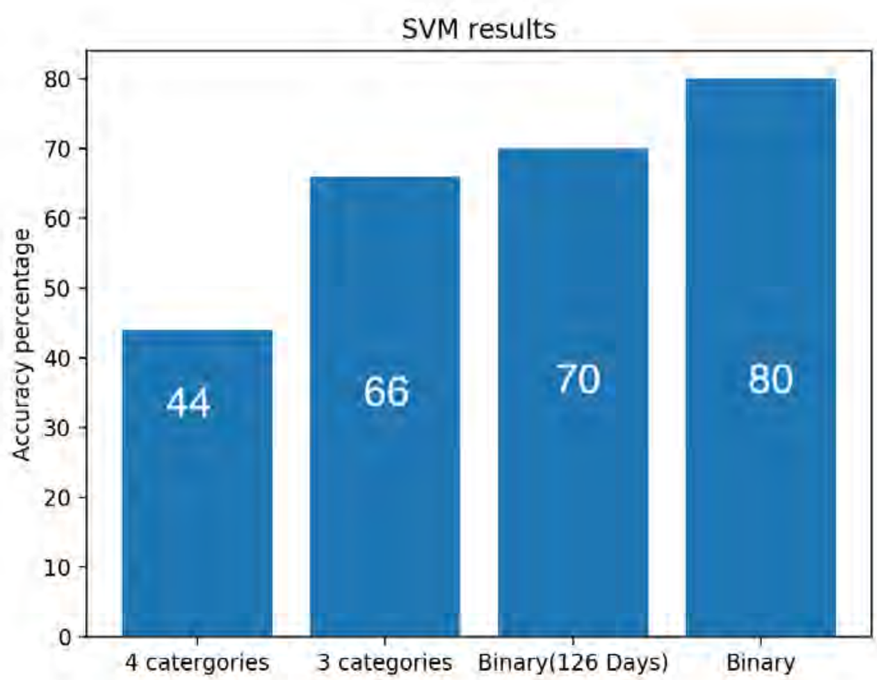


Figure 19: Accuracy level of SVM

6. Discussion

In our study, we tried to predict the LOS of SCI patients. Our plan was to predict the LOS of CRP patients. In addition, we tried to find out whether the patients will stay longer than 100 days or not, as it a one of the most important information for CRP. for this purpose we first chose MLR model for our study. But this model is not able to predict all kind of LOS. Moreover the adjusted R squared value is also very low. The main reason behind this is for MLR, limited data, the number data we collected was quite low. Since it was medical data, very less number of pattern has been noticed. For all these reason MLR was not able to redict all of the LOS.

In ANN, As we have seen above , when we have divided the target class in 4 categories each with 60 days, the result was not so impressive. But when we divided the target class into 3 categories or two categories, the result started to improve. When we divided the target class into three categories, we divided it in 0-100 days, 101-200 days and the 201 to the rest. We found the accuracy level to be around 70 percent, but the problem was most the prediction indicates towards 100-200 days as this category has the most patients. Again when we divided the target class into two classes, one less than 100 and one greater than 100, the accuracy level was near 80% percent but the problem was most of the prediction indicated towards more than 100 days class. So in order to avoid this, we divided the target class into two categories, more than 126 days and less than 126 days. This made the target class into even distribution . The accuracy level of this category prediction was near 70%. For our data set, the artificial neural networks result was not satisfactory due to several reasons

1. In order the predict future using artificial neural network, there have to be some kind of patterns. But as we have collected real life data from a hospital and as the data was medical data, there was very little pattern in the data. For example, we saw that the admission SCIM had a greater impact on the LOS but other parameters did not have that kind of impact.

2. For better prediction , the neural network need a lot of data. But after careful consideration we could only use 370 data of the patients. If we consider all the previous works that had done on prediction of LOS , they had almost 5 to 10 times the data we had. So if we could had that amount of data, our prediction would have improved.
3. We had only 12 parameters to work with in this data set. In order to better predict with the neural network we needed lot more parameters that had influence on the LOS.

7.Conclusion and Future Work

In our thesis work, we tried to Length of Stay of the spinal cord injured patients in Bangladesh.

We have worked on the data of only two years of 370 patients. For our dataset prediction with linear regression worked better than the neural network . To predict the length of stay with neural network, we needed huge sets of data but in Bangladesh , data collection is very tough. We had to wait two months to get the data from CRP. We have tried to collect data from different sources but it was not successful.

In the beginning we got data of 500 patients but we had to exclude 130 patients data as there were some error in the dataset. So in the future we intend to collect more data from CRP to improve our model.

We planned to use different classifier like random forest or decision tree but due to time constraints we could not use them. We wanted to modify our neural network and svm model but we could not do it . So that will also be our goal to improve the model.

Our initial goal was to make a model that can be used for the better treatment of the SCI patients but our model is in prototype condition. Right now it is not usable for the general public but we are planning to build a software, which will help the doctors to predict length of stay of the SCI patients at a early stage.

8.References

[1].http://www.icdri.org/Mobility/centre_for_the_rehabilitation_of.htm.

[2].C. Pavese, M. Schneider, M. Schubert et a, “Prediction of Bladder Outcomes after Traumatic Spinal Cord Injury: A Longitudinal Cohort Study”, PLOS medicine journal, June 21, 2016.

[3].Pei-Fang (Jennifer) Tsai, Po-Chia Chen, Yen-You Chen, et al., “Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network,” Journal of Healthcare Engineering, vol. 2016, Article ID 7035463, 11 pages, 2016.

[4]. Abdul Awal, Md & Rabbi, Jakaria & Hossain, Sk & Hashem, M.M.A.. (2016). Using Linear Regression to Forecast Future Trends in Crime of Bangladesh. . 10.1109/ICIEV.2016.7760021.

[5]. R. Jiménez, R. Lam, M. Maro and M.Marot, “Observed-predicted length of stay for an acute psychiatric department, as an indicator of inpatient care inefficiencies. Retrospective case-series study”, BMC Health Services Research, 17 February 2004.

[6].L. McLay, M. USNR el at, “Predictors of Length of Stay in a Psychiatric Ward Serving Active Duty Military and Civilian Patients”, MILITARY MEDICINE, 170, 3:219, 2005

[7].R Fine, P & L Stover, S & DeVivo, Michael. (1987). “A methodology for predicting lengths of stay for spinal cord injury patients. Inquiry : a journal of medical care organization”, provision and financing. 24. 147-56.

[8].Z. Ismail, A. Yahya and A. Shabri, “Forecasting Gold Prices Using Multiple Linear Regression Method ”, American Journal of Applied Sciences 6 (8): 1509-1514, 2009.

[9]G. Uyanik, N. Guler, “A study on multiple linear regression analysis”, Procedia social and behavioral sciences, 2013.

[10]. A. MINAI and R. WILLIAMS, “On the Derivatives of the Sigmoid”, Neural Networks. Vol. 6, pp. 845-853, 1993.

[11].H. Imai, J. Hosomi, H. Nakao et al., "Characteristics of psychiatric hospitals associated with length of stay in Japan," Health Policy, vol. 74, no. 2, pp. 115–121, 2005.
[12]. A. Razzak, S. Helal, R. Nuri, "Life Expectancy of Persons with Spinal Cord Injury (SCI) Treated in a Rehabilitation Centre at Dhaka, Bangladesh", Disability, CBR & Inclusive Development, Vol 22, No.1, 2011.
[13].S. M. Shortell, J. E. Zimmerman, D. M. Rousseau et al., "The performance of intensive care units: does good management make a difference?" Medical Care, vol. 32, no. 5, pp. 508–525, 1994.