Infant Mortality Prediction Using Artificial Neural Network

Hasan Tarek(0818)
Saara Sheneen(0833)
Sefat-E-Mehedi(0839)
Institute of Information Technology
University of Dhaka

Abstract— **Infant mortality is the death of young children under the age of 2. Many factors have impact on infant mortality. If this can be predicted before a child's birth, the rate can be decreased by taking proper steps. In this paper we have used artificial neural network to predict the rate of infant mortality based on those attributes. This project might help decision makers of the government to impact on attributes that cause infant mortality. Doctors also can predict if a newborn requires beyond regular care if predicted probability is high.**

I. Introduction

An artificial neuron network (ANN) is a computational model based on the structure and functions of biological neural networks. Information that flows through the network affects the structure of the ANN because a neural network changes - or learns, in a sense - based on that input and output.

ANNs are considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found. ANN is also known as a neural network.

Infant mortality is the death of young children under the age of 2. This death toll is measured by the infant mortality rate, which is the number of deaths of children under two year of age per 1000 live births. Infant mortality is an important measure of human development, related to the level of welfare of a society.

Infant Death is still a burning issue in Bangladesh. Due to lack of facilities and many factors, infant death rate is 0.038% in this under-developed country. Patterns and trends in causes of infant mortality help decision-makers assess needs, prioritize interventions, and monitor progress. However, data on causes of death tend to be limited in developing countries, including Bangladesh.

In this project, we have tried to find out which factors cause infant death using neural networks. For this purpose, we have taken lots of features which can also be treated as factors or attributes for predicting the rate of infant mortality. The objective of this paper is to build a model based using artificial neural network which will be trained using the past data records of both successful and unsuccessful births.

This paper is organized by following sections. Section 2 contains related works, section 3 contains methodology, section 4 contains findings and result, section 5 contains conclusion.

II. Related Works

A study about infant mortality with children under one year old was performed in using DM techniques over the integrated SIM and SINASC databases of the municipality of Rio de Janeiro, RJ, Brazil, between the years 2008 and 2012 [3].

A paper [1] has shown that the infant mortality rate (IMR) defines as the risk for a live born child to die before its first birthday is known to be one of the most sensitive and commonly used indicators of the social and economic development of a population (MasuyStroobant Gourbin, 1995). The association between deprivation and poor survival in infancy was already documented with survey data as early as 1824 (Villerm, 1830 quoted by Lesage-Dugied, 1972). The association between socio-economic factors and infant mortality was further reinforced when improvements in overall infant mortality levels over time ran parallel with general social and economic development in most industrialised countries during the twentieth century. Furthermore, since the Second World War, corroboration of the strong inverse relationship between socio-economic development and mortality rates has been found repeatedly among countries and areas within countries. At the individual level, significant social inequalities are repeatedly recorded, even when the overall IMR reaches very low levels (Haglund et al., 1993). Links between individual-level social inequalities and regional (aggregate-level) differences are partly explained by relatively high spatial concentration of the deprived

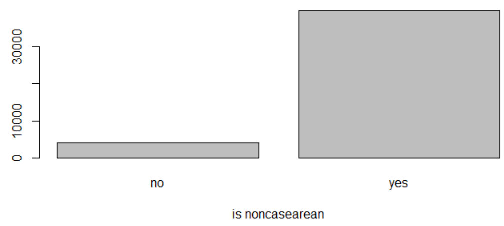

Fig. 1. noncesarean frequency diagram

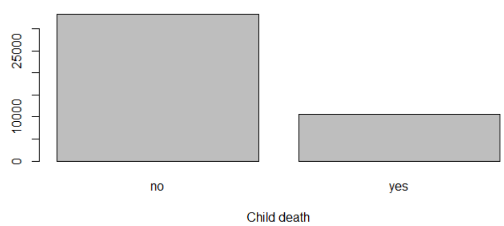

Fig. 2. child death frequency diagram

and of populations of lower social class (United Nations, 1953; Masuy-Stroobant, 1983).

Trends in the causes of child mortality serve as important global health information to guide efforts to improve child survival. With child mortality declining in Bangladesh, the distribution of causes of death also changes. Pneumonia remained the top killer of children under 5 in Bangladesh between 1993 and 2004. The increasing importance of neonatal survival is highlighted by the growing contribution of neonatal deaths and several neonatal causes [4].

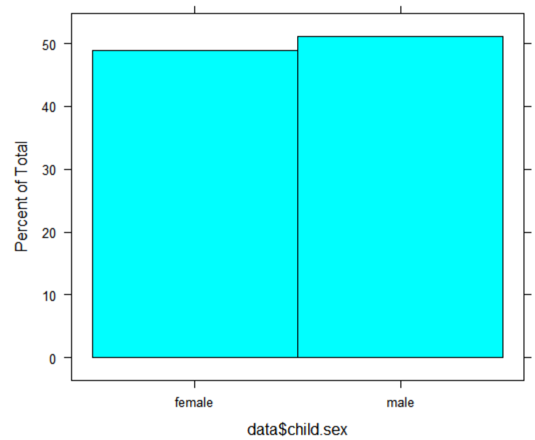

Fig. 3. Gender distribution bar diagram

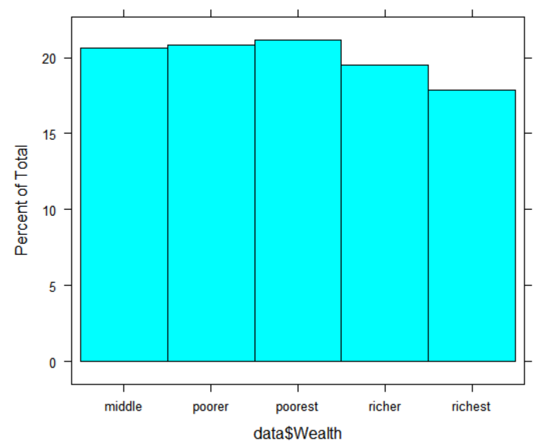

Fig. 4. Bar diagram of wealth status

III. Data Description

We have taken our dataset from The DHS Program(demographic and health serveys)[5]. We can use the following graphs to visualise the overall dataset. From our dataset we have extracted 17 attributes and graphically plotted four of them as demo.

In our dataset (fig 1) we have found out above nearly 80% new born babies are given normal birth and 20% of newborn babies are caesarean.

Death attribute(fig 2) can take two values yes/no. In our dataset we have found out that 60% of children are given birth as dead and 40% are given birth alive.

We have considered 50% male babies and 48% female babies for our prediction so that we can get an unbiased result (fig 3). We have considered economical condition of families(fig 4). This wealth variable can take five values such as middle class, poorer, poorest, richer and richest class. Here,21% families are from middle class, 22% families are poorer, 23% families are poorest, 18% are richer and 17% are richest. In fig 5 and 6, we have represented a summary of our considered attributes below where we have calculated the mean median outliers value for each quantitative attribute and for qualitative attributes we have showed the total number of considered data. Both fig 5 and fig 6 depict overall statistical analysis and distribution of the entire dataset.

IV. Methodology

A. Attribute Selection

The Digital Health Service program(DHS) of USAID dataset contains 12 sub divisions. They are characteristics of respondents, household characteristics of respondents, Infant and child mortality,maternal and newborn health ,marriage and sexual activities, Fertility, Fertility preferences, Fertility regulation, child health, nutrition and of child and women, HIV related attributes, community characteristics. Each division contains a variable number

```
Education.in.single.years household.members
Min.   : 0.00             Min.   : 1.000
1st Qu.: 0.00             1st Qu.: 4.000
Median : 4.00             Median : 5.000
Mean   : 4.05             Mean   : 5.631
3rd Qu.: 7.00             3rd Qu.: 7.000
Max.   :17.00             Max.   :25.000
Number.of.eligible.women.in.the.household     Wealth
Min.   :0.0000                              middle :9041
1st Qu.:0.0000                              poorer :9121
Median :0.0000                              poorest:9269
Mean   :0.6319                              richer :8535
3rd Qu.:1.0000                              richest:7806
Max.   :7.0000
total.children   total.sons.died   total.daughters.died
Min.   : 1.000   Min.   : 0.0000   Min.   :0.0000
1st Qu.: 2.000   1st Qu.: 0.0000   1st Qu.:0.0000
Median : 3.000   Median : 0.0000   Median :0.0000
Mean   : 3.701   Mean   : 0.2412   Mean   :0.1969
3rd Qu.: 5.000   3rd Qu.: 0.0000   3rd Qu.:0.0000
Max.   :15.000   Max.   :10.0000   Max.   :5.0000
```

Fig. 5. Summary of the overall dataset(a)

```
last.5.years.total.birth first.birth.age last.child.in.caesarean
Min.   :0.0000           Min.   :10.00   no :33205
1st Qu.:0.0000           1st Qu.:15.00   yes:10567
Median :0.0000           Median :17.00
Mean   :0.4252           Mean   :17.47
3rd Qu.:1.0000           3rd Qu.:19.00
Max.   :4.0000           Max.   :46.00
             twin         month.born      child.sex
3rd of multiple:    4   Min.   : 1.000   female:21376
no             :43114   1st Qu.: 3.000   male  :22396
yes            :  654   Median : 7.000
                        Mean   : 6.554
                        3rd Qu.:10.000
                        Max.   :12.000
previous.birth.interval.month breastfeeding.month casearean
Min.   :  9                   Min.   : 0.00       no :33205
1st Qu.: 25                   1st Qu.: 7.00       yes:10567
Median : 37                   Median :15.00
Mean   : 45                   Mean   :15.49
3rd Qu.: 58                   3rd Qu.:24.00
Max.   :267                   Max.   :35.00
          size.of.child.on.birth death
average              :29660      no : 4002
larger than average : 4516       yes:39770
smaller than average: 5577
very large          : 1055
very small          : 2964
```

Fig. 6. Summary of the overall dataset(b)

of attributes such as characteristics of the respondents part contains age, residence, wealth status, marital status, education, literacy, access to the media, employment. Infant and child mortality division contains child age, mothers age at birth, birth order, previous birth interval. Maternal and newborn health contains place of antenatal care, is baby caesarean, assistance during delivery, breastfeeding month, size of child, weight of a child. Among these attributes we selected 17 attributes which are corresponded to infant mortality manually. They are:

1) Workable household members
2) Wealth status of a newborns family
3) Total children of a mother
4) Number of births in the last five years
5) Age of mother at first birth
6) Education level of mother
7) Average BMI of mother during pregnancy
8) Child sex
9) Previous birth interval in month
10) Is newborn caesarean
11) Size of a child
12) Weight of a child
13) Is newborn vaccinated
14) Is newborn twin
15) Level of child malnutrition
16) Initiation time of breastfeeding
17) Duration of breastfeeding in months

B. Missing Value Handling

Missing value is an obstruction to proceeding for building machine learning models. This dataset contains total 43274 observations and 4.3% of them contain missing values of one or several attributes. Attribute values are missing completely at random, Which means missing values are distributed randomly across all over observations. To be sure that property of missing values is completely at random, we compared into two parts of the dataset, one with missing observations and other without missing observations. On a t-test of each part we found that there is no difference between the means of these two samples for each attribute with missing values. So we concluded that missing values follow completely at random property. Now we can apply techniques for completely at random distributed missing observations.

There are two types of attributes with missing values:

1) Numerical missing value
2) Categorical missing value

C. Imputation of Numeric Missing Value with Mean Substitution

As this dataset does not contain time series observations, mean substitution technique is good enough. We took ten observations of upper row and ten observations of lower row, averaged them and put the average value into the middle empty cell. This technique might affect overall variance of the dataset but does not alter other true population parameters.

D. Imputation of Categorical Missing Value with KNN

The k nearest neighbours is an algorithm that is used for simple classification. The algorithm uses feature similarity to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. This can be very useful in making predictions about the missing values by finding the ks closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the neighbourhood. In our case we chose k=10 to trade between bias and variance of the KNN model.

E. One Hot Encoding

Categorical variables are not eligible to train a learning model. To convert this kind of categorical text data into model-understandable numerical data, we used one hot encoding technique. We took a column which has categorical data, which has been label encoded (each class was replaced by an arbitrary numeric value), and then split the column into multiple columns. The numbers are replaced by 1s and 0s, depending on which column has what value. Python scikit-learn library was used to perform one hot encoding.

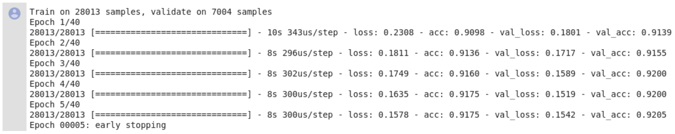
```
Train on 28013 samples, validate on 7004 samples
Epoch 1/40
28013/28013 [==============================] - 10s 343us/step - loss: 0.2308 - acc: 0.9098 - val_loss: 0.1801 - val_acc: 0.9139
Epoch 2/40
28013/28013 [==============================] - 8s 296us/step - loss: 0.1811 - acc: 0.9136 - val_loss: 0.1717 - val_acc: 0.9155
Epoch 3/40
28013/28013 [==============================] - 8s 302us/step - loss: 0.1749 - acc: 0.9160 - val_loss: 0.1589 - val_acc: 0.9200
Epoch 4/40
28013/28013 [==============================] - 8s 300us/step - loss: 0.1635 - acc: 0.9175 - val_loss: 0.1519 - val_acc: 0.9200
Epoch 5/40
28013/28013 [==============================] - 8s 300us/step - loss: 0.1578 - acc: 0.9175 - val_loss: 0.1542 - val_acc: 0.9205
Epoch 00005: early stopping
```

Fig. 7. Training accuracy from different fold

F. Feature Scaling

Scaling refers to change of range of values without altering their distribution. Neural model works better when features are on a relatively similar scale and close to normally distributed. We applied standard scaling technique to normalize feature values. Standard scaling equation:

$$x' = (\mu - x)/\sigma \tag{1}$$

Python scikit-learn library was used to perform standard scaling.

G. Model Generation

A feedforward neural model with 2 hidden layers was built with keras Sequential library. Categorical cross entropy was as a loss function with adam optimizer and accuracy metrics.

H. Model Evaluation

Both 80-20 split and 10 fold cross validation approach was used to evaluate and validate the model.

V. Findings & Analysis

From 10 fold cross validation and 80-20 split, we have found accuracy given in the following table:

Fold	Accuracy
1st fold	0.9198264
2nd fold	0.92804934
3rd fold	0.92392049
4th fold	0.92346356
5th fold	0.92209276
6th fold	0.9266621
7th fold	0.91957962
8th fold	0.91935115
9th fold	0.93237377
10th fold	0.92277816
80-20 split	0.9246145060101896

TABLE I
Accuracy from different fold

Overall accuracy of the model is 92.38%

The following graph(fig 7) represents accuracy of the model from different fold. From the plot of loss(fig 8), we can see that the model has comparable performance on both train and validation dataset (labeled test). As the epoch number increases the validation loss decreases

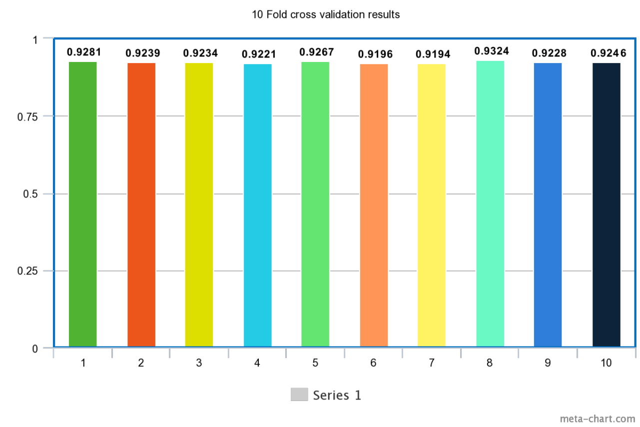

Fig. 8. Bardiagram of accuracy from different fold

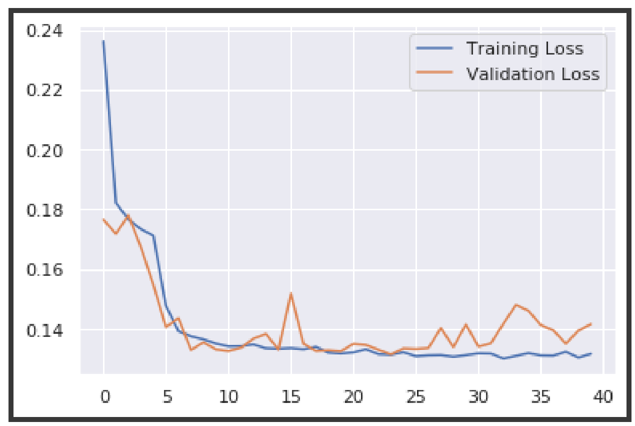

Fig. 9. Loss graph

If these parallel plots start to depart consistently, it might be a sign to stop training at an earlier epoch.

From the below loss graph we can see that the validation loss decreases at each training epochs

we also uses Early stopping method, so the model can not be over fitted.

VI. CONCLUSION

In this paper we analysed a dataset from USAID to classify infant mortality. This classification can be helpful to prevent infant death. Infant mortality rate largely depends on these factors mentioned in this report. We have already seen that; these factors give a sound accuracy in our prediction. So, by taking necessary steps regarding on these factors may reduce child death.

References

[1] **Masuy-Stroobant, Godelieve. (2002). The Determinants of Infant Mortality: How Far are Conceptual Frameworks Really Modelled?. 10.1007/978-1-4020-4676-6_2. .**

[2] **Sanjay Kumar, Ari Viinikaineny and Timo Hamalainenz, 2016. Machine Learning Classification Model For Network Based Intrusion Detection System**

[3] Zhou D, Frize M., Predicting probability of mortality in the neonatal intensive care unit
[4] Li Liu, Qingfeng Li, Rose A Lee, Ingrid K Friberg, Jamie Perin, Neff Walker, and Robert E Black, Trends in causes of death among children under 5 in Bangladesh, 1993-2004: an exercise applying a standardized computer algorithm to assign causes of death using verbal autopsy data. 2011.
[5] dataset link-https://dhsprogram.com/data/Model-Datasets.cfm 15:54, 5 November,2019