International Journal of Scientific & Engineering Research, Volume 11, Issue 3, March-2020
ISSN 2229-5518

Cervical Cancer Risk Factor Prediction Using Behavioral Data

A.B.M. Shihab Uddin
Faculty of Science & Information Technology
American International University-Bangladesh

Shams Ibne Noor
Faculty of Science & Information Technology
American International University-Bangladesh

Sajid Hasan Sifat
Faculty of Science & Information Technology
American International University-Bangladesh

Maisha Masnoon
Faculty of Science & Information Technology
American International University-Bangladesh

Zakia Zerin Haque
Faculty of Science & Information Technology
American International University-Bangladesh

Abstract— Cervical cancer is still a critical global health issue. There are several researches going on cervical cancer prediction using numerous types of data mining techniques. Many of the researches are based on behavioral data while others are based on diagnosis data. Some even have combined the behavioral aspects with the diagnosis findings to build their prediction systems. These researches used data mining techniques to process and manipulate data and come up with outcomes. Machine Learning techniques can improve the cancer prediction. In our work, we will be using machine learning approaches and Neural Network to build an accurate prediction model to predict cervical cancer by using behavioral data sets.
Despite knowing an individual's behavioral risk factors of cervical cancer, it is quite tough to predict the cancer. Different studies disclose that, although having high behavioral risk certain people are never diagnosed with cancer. Alternatively, some people with low behavioral risk are diagnosed with cervical cancer. Nonetheless, behavioral risk factor increases the chance of developing cervical cancer. In this research we used behavioral data to get a mutual pattern so that we can use these patterns to predict the risk factors of a person to have cervical cancer.

Index Terms— Behavioral Data, Cervical Cancer, Data mining, Machine Learning, Pattern Analysis, Risk Factor Prediction.

1 Introduction

According to the studies of World Health Organization, Cervical cancer is the second most common form of cancer for women living in under-developed countries. Approximately 75% of cervical cancer cases in developing countries are diagnosed with the cancer at an advanced stage which is difficult to cure. To detect the pre-cancerous stage, cervical cancer screening is highly recommended by the doctors. However, the screening process of cervical cancer is laboratory oriented which makes it costly and unfeasible for the people of low-incoming countries. As a result, access to the screening program and treatment becomes limited. The limited access to the screening process is one of the main reasons of high mortality rate for cervical cancer of women, in developing and underdeveloped countries.
Cervical Cancer is one of the most preventable types of cancers, which leaves us with an obvious question, why the mortality rate of women around the world because of cervical cancer is increasing day by day? Cancer itself is very unpredictable as many cancers are not genetically determined. In most of the cases cancer is influenced by the lifestyle factors, environmental exposure and sexual behavior. Many researches show that, like most of the cancers, cervical cancer is also governed by behavioral aspects such as smoking, number of sexual partners, age of pregnancy, number of pregnancies and the list goes on. Sexual behaviors, pregnancy, birth control techniques can be either the reason of increasing the risk or can be the reason of reducing the risk of cervical cancer [13] [14] [16]. Habit of using hormonal contraceptives can increase the risk for cervical cancer where usage of intrauterine device (IUD) can reduce the risk [15] [16]. Still there is so much individual variability that, prediction of cancers that are governed by behavioral aspects or the life.

2 Methodology

2.1 Dataset Description

The dataset was collected from UCI machine learning repository. Original source of this dataset is 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values). There are total 32 attributes or rows in the dataset. Among these, we take 11 rows for our task which all are basically behavioral or non-diagnosed data. Details of these selected row will given below:

2.2 Missing Value Replacement

In this research we have used KNN-imputation to replace missing values. KNN is used to find out the closest K neighbors of a missing value based on other variables in a multi-dimensional space. It can handle continuous, discrete, ordinal and categorical type missing data. Using KNN for missing value replacement, we had to consider some parameter of KNN imputation:

1. Number of neighbors
2. The aggregation method
3. Data normalization
4. Numeric attribute distances

2.3 Editing and optimizing non-diagnosed data

Non-diagnosed dataset contains 11 attributes with 858 instances. Those attributes are given below:

1. Age (Numeric)

International Journal of Scientific & Engineering Research, Volume 11, Issue 3, March-2020
ISSN 2229-5518

2. Number of sexual partners (Numeric)
3. First sexual Intercourse (Numeric)
4. Number of pregnancies (Numeric)
5. Smokes (Binary)
6. Smoke years (Numeric)
7. Smoke pack per years (Numeric)
8. Hormonal Contraceptive (Binary)
9. Hormonal Contraceptive years (Numeric)
10. IUD (Binary)
11. IUD years (Numeric)

Now there are two Intuitions among these attributes. Those are described below:

Intuition 1: There are three binary columns which don't have any significance basically. Because every binary column has its supporting next column. For example, IUD holds binary values, zero or one. That means whether a patient use IUD or not. Next column contains how many years a patient uses IUD. Now using these columns, one can clearly take binary decision whether a patient used IUD or not. Therefore, we eliminated three binary columns.

Intuition 2: Smoke pack per year is a unit for measuring the amount a person has smoked over a long period of time. Equation of calculating smoke packs per year is: (number of *cigarettes* smoked per day/20) × number of years smoked. 1 pack year is equivalent to 365.24 packs of cigarettes or 7,305 cigarettes.

$$\begin{aligned}1\ \text{Pack-Year} &= \frac{1Pack}{Day} \times 1year \\ &= \frac{1Pack}{day} \times 365.24\ days \\ &= 365.24\ \text{packs} \\ &= 365.24 \times \frac{20\ Cigarettes}{Pack} \\ &= 7.305\ \text{Cigarettes.}\end{aligned}$$

2.4 Category & priority weight of every attributes

2.4.1 Risk Factor Category

Concept of doing category section came from significance of every attribute. For example, if we take a look at previous section, then we can see that different age value has different kind of impact for causing cervical cancer. We need to distinguish these varieties of significance into the values and that's why we denoted five risk categories of every attributes which are:

(1) Very low (0-10%)
(2) Low (11-30%)
(3) Medium (31-50)
(4) High (51-80)
(5) Very High (81-100)

Boundary range of these categories varied from attributes to attribute given below:

Attribute Name	Very low	Low	Medium	High	Very High	Unit
Age	13-20	21-24	61-70	46-60	25-45	Year(s)
Number of sexual partners	0-1	2-3	4	5-6	7+	Person(s)
First sexual Intercourse	24+	21-23	17-20	15-16	<15	Year(s)
Number of pregnancies	0-1	2-4	5-6	7-8	8+	Year(s)
Smoke years	0-1	1.00 1-3	3.01-5	5.01-7	8+	Year(s)
Smoke quantity per day	0-5	5.01-10	10.01-20	20.01-30	30.01-50	Piece(s)
Hormonal Contracep-tive years	0-2	2.01-4	4.01-5	5.01-8	8+	Year(s)
IUD years	5+	3.01-5	1.01-3	0.01-1	0	Year(s)

2.4.2 Priority weight

We already identified that different values of different attributes have different impact. Just like that if we considereach attribute as a whole then it can be realized that all attributes aren't equally responsible for causing cervical cancer. For example, if we do a comparison then we can notice that hormonal contraceptive and IUD contain most priority weight rather than age for causing cervical cancer. From this concept, we gave every attributes a priority weight value on the scale of 1 to 4. Summation of all priority weight value is 23. We devised every attribute's priority value by 23 to get weighted average.

Attribute Name	Priority weight	Weighted average calculation	Weighted average value	Reference number
Age	1 out of 4	1/23	0.044	08,09
Number of sexual partners	2 out of 4	2/23	0.087	10
First sexual intercourse	2 out of 4	2/23	0.087	11
Number of pregnancies	3 out of 4	3/23	0.131	17
Smoking years	3 out of 4	3/23	0.131	12,13
Smoke quantity per day	4 out of 4	4/23	0.174	12,13
Hormonal contraceptive years	4 out of 4	4/23	0.174	14,15
IUD years	4 out of 4	4/23	0.174	16

International Journal of Scientific & Engineering Research, Volume 11, Issue 3, March-2020
ISSN 2229-5518

2.5 Data scaling using category and priority weight

We couldn't feed raw data into the prediction model directly. Because, currently they had been maintaining an inconsistent scale. That's why we need to do a consistent and proper scaling to scale our data. To do this, we had to use the concept of previous section which is using category and priority weights.

General equation for data scaling is:

$$\frac{(\textbf{Value difference} * \textbf{Percentage per unit}) + \textbf{Previous Percentage}}{100} \times \textbf{Previous Percentage}$$

Where,

Value difference = Input value – lowest value of the range

Per unit percentage

= (difference between percentage range of the category) /(difference between value range of the category)

Here, point to be noted that before multiplying by priority weight value, what we got is actually the individual risk factor percentage for the corresponding value. But how much this individual risk factor will matter for total cause that actually depends on the corresponding attribute weight values.

2.5.1 before Applying Data Scale

Here are the first 10 rows of non-scaled data of the dataset.

id	*Age*	*No of sexual partners*	*First sexual intercourse*	*No of pregnancies*	*Smokes years*	*Smokes packs year*	*Hormonal Contraceptives years*	*IUD Years*
1	17	2	15	2	0	0	0	0
2	19	2	15	2	0	0	1	0
3	18	2	16	1	6	0.25	0	0
4	41	3	17	4	0	0	10	0
5	40	1	18	1	0	0	0.25	0
6	37	2	18	3	0	0	0	3
7	35	3	17	4	0	0	7	0.08
8	35	3	20	2	0	0	0	10
9	35	3	17	6	13	2.6	7	0
10	36	2	15	3	0	0	0	0

2.5.2 after Applying Data Scale

Considering a single row data scenario:

Age = 41, First sexual intercourse Age = 17 years,

Number of sexual partners = 3 persons,

Number of pregnancies = 4,

Smoking years = 0, number of cigarettes per day = 0, hormonal contraceptive years = 10,

Intrauterine Device (IUD) usage years = 0.

Now we need to scale these 8 data values according to their boundary range and priority values.

Age: 25<**41**<45. That means it belongs to the very high category (81-100%). Value difference = 41-25 = 16

$$\textbf{Per unit percentage} = \frac{\textbf{difference between percentage range of the category}}{\textbf{difference between value range of the category}}$$

Per unit percentage = (100-81)/ (45-25) = 0.95

As it falls between 81-100% range that means it contains 80% previous percentage. So,

Previous remaining percentage value = 80.

Priority weight value for age= 0.044

Now putting all these values to the main equation, we get by using

$$\frac{(\textbf{Value difference} * \textbf{Percentage per-unit}) + \textbf{Previous Percentage}}{100} \times \textit{Previous Percentage}$$

Calculation: ((16*0.95) + 80) =

95.2/100 = 0.952 * 0.044 = 0.04189

Here value of age (41) contains 95.2% risk factor according to risk factor boundary but that didn't create any impact because priority weight value of age column is pretty low (0.044).

First sexual intercourse age: Value is 17 and the range is 17-20. So it falls to the medium category (31-50%).

Value difference = 17-17 = 0

per unit percentage = (50-31)/ (20-17) = 6.333

As it falls between 31-50% range that means it contains 30% previous percentage.

So, **previous remaining percentage value** = 30.

Priority weight value for first sexual intercourse age=0.087

Now, putting all these values to the main equation, we get,

Calculation: ((0*6.33) + 30) = 30/100 = 0.3 * 0.087 = 0.0261

Here, 30 or 0.3 is the individual risk factor percentage (30%) of this column. Similarly, we can get the scaled value of the other columns. Here are the first 7 rows of the scaled data from the

ID	Age	No. of sexual partners	1st sexual intercourse	No. of pregnancies	Smokes years	Smoke packs year	Hormonal Contraceptive years	IUD Years	Label value
1	0.00249	0.0087	0.0696	0.01441	0	0	0	0.15566	0.25180
2	0.00374	0.0087	0.0696	0.01441	0	0	0.008698	0.15566	0.26175
3	0.00312	0.0087	0.0435	0.01048	0.078568	0.0029	0	0.156	0.3038
4	0.04224	0.0261	0.02697	0.0393	0	0	0.139337	0.156	0.4305
5	0.0418	0.00522	0.032477	0.01048	0	0	0.002174	0.156	0.2487
6	0.0404	0.0087	0.032477	0.02685	0	0	0	0.052	0.1607
7	0.0396	0.0261	0.02697	0.0393	0	0	0.120742	0.133	0.3865

International Journal of Scientific & Engineering Research, Volume 11, Issue 3, March-2020
ISSN 2229-5518

scaled dataset.

2.6 Building Neural Network Model

2.6.1 Our neural Network structure

Like other traditional neural network model, our model also has three layers which are given below:

(i) Input Layer: As other neural network, our model also contains one input layer. As input has eight dimensions that's why input layer contains eight input neurons. It takes the input from user.

(ii) Hidden Layer: Our neural network has one hidden layer which has six neurons. It processes data that is given by previous input layer by feed forward and backward propagation method.

(iii) Output Layer: We have one output layer with single output neuron in that layer which displays the predicted results.

(iv) Every neuron in every layer is connected with neurons of next and previous layer via synaptic weight. As input and output values are fixed very often, value of these synaptic weights are mainly responsible for how much a model can train.

Input Layer work load

These layers' neuron doesn't have much work to do. It takes the values and send forward to the next layer which is hidden layer via synaptic weight. But before doing this, these input data are also being normalized like similar way otherwise input data format and train data format doesn't match and it may doesn't give expected prediction result.

Hidden Layer work load

This layer is the core part of the neural network because most of the operation is performed in this layer. All neurons are connected with every neuron in input layer. Every connection has a weight value which is set by random number generation. In our model, we use seed () method so that it generates same weighted value.

Data which is forwarded from input neuron are multiplied with weight like that:

If node= n1, weight= W1, input= I1 then

Value, n1=W1*I1

All the multiplied value needs to be added:

Sum= (W1*I1) + (W2*I2) + (W3*I3) + --------- + (Wn*In)

Activation Function

There are many activation function used in neural network. We use sigmoid function in our network model.

Sigmoid Function, (x) = 1/ (1+e^-x)

Sum values are passed through this function and this function returns squished value. Specialty of this function is it always returns the value between 0 to 1 no matter what's the value it takes. That's why it also useful for us to process the data.

Output Layer work load

Like input layer, output layer also doesn't have to do much work. It just shows the final result that is passed from before hidden layer.

Back Propagation

Whole neural network (input, hidden, output layer) is involved in the back propagation method. In the "hidden layer work load" section, we just describe the forward propagation method. After completing forward propagation, there is a difference between expected output and the output that is calculated by forward propagation method. We find the difference between these two as error. We used Mean Square Error (MSE) method to calculate and handle this difference error.

$$\frac{1}{2n}\sum_{x}||y(x)-a||^2$$

Equation to calculate MSE:

Where,

n = Number of training inputs.

x = Set of all inputs.

y(x) = expected output of the network for input x.

a = actual output of the network for input x.

This error is happened because of randomly generated weight values and every weighted value is responsible for this error more or less. Now to find out the adjustment of weight and to minimize the error in each iteration, we used stochastic gradient Descent (SGD) as an optimizer.

Optimization algorithm: It helps to minimize or sometimes maximize the error function which depends on model's internal learnable parameters. Weights and biases are called models internal learnable parameters which are used to compute the output values and learned and updated in the direction of optimal solution.

SGD as an optimization algorithm: As an optimizer, it's used to minimize the objective error function that has the form of a sum:

$$Q(w)=\frac{1}{n}\sum_{i=1}^{n} Qi(w)$$

Here each Summed function, Qi is typically associated with i-th observation to n in the dataset which is used for training. In every iteration, error is passed through the derivative of sigmoid function. This value is multiplied by total error. Then this multiplied value and input value both are gone

International Journal of Scientific & Engineering Research, Volume 11, Issue 3, March-2020
ISSN 2229-5518

through the matrix multiplication. At last, we find the adjustment values which add to the previous weight.

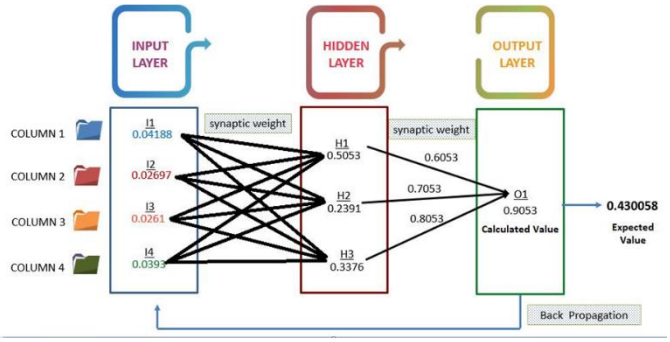

Figure: Neural network model architecture

3 Result

3.1 Individual Risk Percentage of Every Attribute

If running input example set has been continued, then visual representation of these input values is given below.

This is actually the individual risk factor representation which is how much risk factor causes by individual input values. For example, value of age, hormonal contraceptive and IUD contain the most risk factor percentage individually. Though these are already calculated values hence these are displayed using progress bar for better understanding.

3.2 Total risk percentage

Let, neural model gives predicted value 0.43041 for running input example set (which is mentioned in previous sections). Using threshold values, it can easily be said that it falls between 40-50% (0.39494<0.43041<0.4999). Now to get the exact risk percentage there have been some basic calculation needs to be done:

0.43041-0.39494 (predicted value – lower value of the range, 0.39494< x < 0.4999)

0.03547*0.9515 (value difference * per unit percentage)

(0.0337+0.4)*100 (adding 0.4 as it falls between 40-50% ~ 0.4-0.5)

43.37% ~ 43 %

At last, risk factor of having cervical cancer is 43% (Medium category) for this input example set.

4 Conclusion

4.1 Contributions

From the beginning of the research our solely motive was to design a model that can be interactive with the user's perspective. That led us to an interactive model which takes inputs of any female individual's behavioral experiences or routines into the model. The model then calculates the data which given and shows the results to that individual of the percentage of risk of having cervical cancer without any medical diagnosis.

4.2 Possibilities for Future Works

4.2.1 Improvements and Modifications

For our research we took medical references to identify risk category and priority weight. As we are not medical students. it is clearly possible to see that there can be significant amount of more research to gather more knowledge so these two factors can be improved in future by further medical research. The neural network model that we created can be modified if data quantity is substantial in future.

4.2.2 Future and Diagnostic Approaches

We worked with only non-diagnosis data, which gave us the insight for the risk factor perspective towards the behavioral experiences of any individual's. In future diagnosis data also can be included in the existing infrastructure to get more knowledge and insights for more meaningful results.

5 References

[1] Cervical cancer and sexual lifestyle: a systematic review of health education interventions targeted at women. By Jonathan Shepherd Greet Peersman Ros Weston Ibrahim Napuli

[2] Sexual behaviour of women with human papillomavirus (HPV) lesions of the uterine cervix. Syrjänen K, Väyrynen M, Castrén O.

[3] Cervical cancer in younger women. By Andrews FJ, Linehan JJ, Melcher DH.

[4] Cervical cancer in pregnancy: reporting on planned delay in therapy. By Duggan B, Muderspach LI, Roman LD, Curtin JP, d'Ablaing G, Morrow CP.

[5] Cervical Cancer and Screening in Great Britain', Report of British Medical Association.

[6] Cervical Cancer: A Global Health Issue, Women 's Health Research Institute.

[7] Applications of machine learning in cancer prediction and prognosis Cancer Informant. By J.A. Cruz, D.S. Wishart.

[8] https://www.cancer.net/cancer -types/cervical -cancer/statistics

[9] https://www.cancerresearchuk.org/health -professional/cancer -statistics/statistics -by -cancer -type/cervical -cancer/incidence#heading -One

[10] Multiple Sexual Partners as a Potential Independent Risk Factor for Cervical Cancer: a Meta -analysis of Epidemiological Studies: Zhi -Chang Liu, Wei - Dong

International Journal of Scientific & Engineering Research, Volume 11, Issue 3, March-2020
ISSN 2229-5518

Liu , Yan -Hui Liu , Xiao -Hua Ye, Si -Dong Chen

[11] Early age at first sexual intercourse and early pregnancy are risk factors for cervical cancer in developing countries: K S Louie, S de Sanjose, M Diaz, X Castellsagué, R Herrero, C J Meijer, K Shah, S Franceschi, N Muñoz and F X Bosch.

[12] How does tobacco smoke contribute to cervical carcinogenesis?: Philip E. Castle

[13] Smoking and cervical cancer: Pooled analysis of the IARC multi -centric case - control study: Martyn Plummer, Rolando Herrero, Silvia Franceschi, Chris J.L.M. Meijer, Peter Snijders, F. Xavier Bosch, Silvia De Sanjose, Nubia Munoz

[14] Effect of oral contraceptives on risk of cervical cancer in women with human papillomavirus infection: the IARC multicentric case -control study

[15] Cervical cancer and use of hormonal contraceptives: a systematic review: Jennifer S Smith PhDa, Jane Green DPhilb, Amy Berringtonde Gonzalez DPhilb, Paul Apple by MScb, Prof Julian Peto DScc, Martyn Plummer MSca, Silvia Franceschi MDa, Prof ValerieBeral MD.

[16] Invasive Cervical Cancer and Intrauterine Device Use: Deborah Lassise, David A Savitz, Richard F Hamman, Anna e Baron, Louise A Brinton, Robert S Levines.

[17] https://www.uptodate.com/contents/cervical -cancer -in -pregnancy