

INTERPRETATION OF NEURODEGENERATIVE DISEASE AND ASSOCIATION WITH CERVICAL AND OVARIAN CANCER

By
MD. RAIHAN AHMED
ID: 161-35-1416

Under the supervision of
SAYED ASADUZZAMAN
LECTURER (SENIOR SCALE)
DEPT. OF SWE, DIU

This paperwork is submitted to the Software Engineering Department to
qualify the prerequisites for the authorization of graduating B.Sc. in
Software Engineering.



DAFFODIL INTERNATIONAL UNIVERSITY

SOFTWARE ENGINEERING DEPARTMENT

Dhanmondi, Mirpur Road, Dhaka, Bangladesh
December 2019

APPROVAL

This thesis titled as “**Interpretation of Neurodegenerative Disease and Association with Cervical and Ovarian Cancer**”, submitted by **Md. Raihan Ahmed, ID: 161-35-1416** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc in Software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan

Professor and Head

Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Chairman

Dr. Md. Asraf Ali

Associate Professor

Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

Asif Khan Shakir

Lecturer

Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Prof Dr. Mohammad Abul Kashem

Professor

Department of Computer Science and Engineering
Faculty of Electrical and Electronic Engineering
Dhaka University of Engineering & Technology, Gazipur

External Examiner

DECLARATION

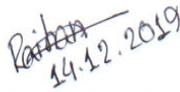
I, **Md. Raihan Ahmed** do hereby declare that this report has been done by me under the supervision of **Sayed Asaduzzaman**, Senior Lecturer, Dept. of Software Engineering, Daffodil International University. We also declare that this report nor any part of this report has been submitted elsewhere for award of any degree.

Supervised By:



.....
Sayed Asaduzzaman
Senior Lecturer
Department of Software Engineering
Daffodil International University

Submitted by:



.....
Md. Raihan Ahmed
ID: 161-35-1416
Software Engineering Department
Daffodil International University

ACKNOWLEDGEMENT

First of all, I might want to to express my sincere appreciation to the Omnipotent for providing me the potential to complete this thesis work. Again, I would express my gratitude to my advisor Sayed Assaduzzaman, Senior Lecturer, Software Engineering Department for his continuous support on my B.Sc study and research for his motivation, support, exuberance, forbearance and huge cognition. His coherent guidance encouraged me to the whole time of performing my research and writing this thesis. I could not imagine of having a better mentor and supervisor for my undergrad study. Beside my advisor, I am also grateful to one of my favorite teacher Kawsar Ahmed, Professor(Assistant), Dept. of ICT, MBSTU. I followed his guidelines that helped me to complete my thesis work without any difficulty.

My sincere thanks to the entire Software Engineering Department at Daffodil International University for providing good education and knowledge.

I also express my gratitude to all my teachers Dr. Yousuf Mahbulul Islam Professor and Vice Chancellor, Daffodil International University; Dr. Touhid Bhuyian, Professor and Head, Dept. of Software Engineering. I would like to convey my heartiest feelings to Nazia Nishat, Senior Lecturer, dept. of Software Engineering and Dr. Md. Asraf Ali, Professor(Associate), dept. of SWE, Daffodil international University for enlightening me the first glance of research. The knowledge that I have learned from the classes for completing my degree of bachelor in Software Engineering level were indispensable for this thesis. In course of conducting the study vital information were collected through books, journals, electronic media and other secondary sources.

Table of Contents

APPROVAL	ii
DECLARATION.....	Error! Bookmark not defined.
ACKNOWLEDGEMENT	iv
List of Tables	vi
List of Figures.....	vii
ABSTRACT	viii
1. INTRODUCTION.....	1
1.1. Background.....	1
1.2. Motivation of Research	3
1.3. Research Gap.....	5
1.4. Research Question	5
1.5. Research Objectives	6
1.6. Research Scope	6
1.7. Thesis Organization.....	6
2. LITERATURE REVIEW.....	7
3. RESEARCH METHODOLOGY	9
3.1. Data Collection.....	9
3.2. Data Preprocessing.....	9
3.3. Evaluation of the performance of Machine Learning Models	10
3.4. Comparison of Feature Selection with individual Evaluators	11
3.5. Deriving Key Features.....	12
3.6. Equation	13
4. RESULT AND DISCUSSION.....	13
4.1. Knowledge Discovery	27
4.2. Calculation of Risk Level.....	32
4.3. Research Output.....	41
5. CONCLUSION	44
6. FUTURE WORK	45
7. REFERENCES.....	45

List of Tables

Table 1. Accuracy of machine learning classifiers of stress.....	14
Table 2. Accuracy of machine learning classifiers of ovarian cancer.	15
Table 3. Accuracy of machine learning classifiers of Cervical Cancer.....	16
Table 4: Info gain, gain ratio, gini index and chi-square test of stress.	17
Table 5: Importance of factors in data table based on info gain, gain ratio and gini index of cervical cancer.	18
Table 6. Importance of factors in data table based on info gain, gain ratio and gini index of ovarian cancer.	19
Table 7. Comparison between the factors using different algorithms of stress.....	20
Table 8. Comparison between the factors using different algorithms of cervical cancer.	22
Table 9. Comparison between the factors using different algorithms of Ovarian Cancer.	24
Table 10. Weightage value of Stress.....	28
Table 11. Weightage value of Cervical Cancer.....	29
Table 12. Weightage value of Ovarian Cancer.....	30
Table 13. Weightage value of Stress, Cervical and Ovarian Cancer.....	31

List of Figures

Figure 1. Box plot Age.....	32
Figure 2. Box plot Age first intercourse	33
Figure 3 Number of children.....	33
Figure 4. Box plot of Age of husband.....	35
Figure 5. Box plot of Marital Status	36
Figure 6. Menopause box plot	36
Figure 7. Probabilities of 10 valuable attributes	37
Figure 8. Decision tree between the factors of cervical cancer	38
Figure 9. Age box plot	38
Figure 10. Age of husband box plot.....	39
Figure 11. BMI box plot	39
Figure 12. Probabilities 10 highly influenced factors.....	40
Figure 13. Decision tree between the factors of ovarian cancer	41

ABSTRACT

Stress can affect on health if on certain time vital moves were not made to control pressure than it might create cancer in human body. Cervical and ovarian cancer is one of these days the most frightening illness amongst females in the midst approaching nation alike Bangladesh. The Bangladesh community are lacking behind in educational activity and awareness about these two cancers. There is no prediction for Bangladeshi female in this modern age. Purpose: To find out the association between factors and the most significant factors of stress, cervical, ovarian cancer. Contribute a prediction on befalling cervical and ovarian cancer based on their worthy factors as well as stress parameters. Methods: A study has been made with case control on 298 patients having cervical and 522 patients of ovarian cancer. Cases of 197 and control of 100 were considered for cervical cancer. In case of ovarian cases of 267 and control of 254 beheld for data mining analysis. Data of 161 persons were taken on stress from NINH. Performance were analyzed with model e.g. Logistics Regression, Random Forest, AdaBoost, Naïve Bayes, Neural Network, kNN, CN2 rule Inducer, Decision Tree, Quadratic Classifier were compared with their standard metrics. For certainty info gain, gain ratio, gini index were revealed of both cervical and ovarian cancer. Attributes were ranked using different feature selection evaluators. Then the most significant analysis was made with the significant factors. Factors like children, age of first intercourse, age of husband, pap test, age are significantly higher factors of cervical cancer. Contradictorily, genital area infection, pregnancy problem, use of drugs, abortion, number of children important factors of ovarian cancer. The analysis was made with significant factors of stress, cervical and ovarian cancer that will help us to forecast the risk of occurring cervical or ovarian cancer and might help to abate the cancer not just from Bangladesh but also all over the world.

1. INTRODUCTION

Business related stress and post awful stress issue is considered as a genuine factor for some illnesses. It is seen that individuals suffering from stress disorders have 55% chance to die with cervical cancer. Ladies who experienced at least 6 side effects of post-traumatic stress disorder had more serious danger of developing ovarian cancer. According to WHO, the second leading disease is cancer, it causes 9.6 million death in 2018[27]. Uncontrolled increase of irregular cells exceeding their regular territory with the ability to attack or potentially spread to different organs is cancer. Among different types of cancer cervical and ovarian cancer is the most prominent hazard to female's wellbeing. Due to cervical and ovarian cancer every year, over and above 300 000 women dies further half a million were diagnosed.

A data mining and machine learning approach has taken place in this paper. As it is seen that stress is a significant factor of cervical and ovarian cancer. A relation has been depicted of stress with cervical and ovarian cancer using data mining techniques and machine learning algorithms. A prediction has been made of the risk level of those deadly diseases. A comparison with the machine learning models was done. We further examined association between significant factors.

1.1. Background

Stress is linked with diverse neurodegenerative issue, specifically, the pathophysiological importance of pressure in Alzheimer's infection and several diseases. Some previous studies also shown that stress spur on cervical and ovarian cancer [20,1]. National Cancer Institute summarizes that, cervical tumor constitutes in the cervix, a body part associating with uterus and vagina [4, 13]. Human Papillomavirus(HPV) is the main reason behind occurring cervical cancer [4]. Sayed Asaduzzaman conducted a study where he finds cervical slowly develops without

showing any indication in the beginning seemingly hard to discover but can be noticed with several pap test [13]. Cervical malignancy grew mostly in the middle of cervix advancement namely squamous metaplasia [13]. It includes changing of the columned epithelium within the cervix into squamous epithelium at some stage in puberty, where transitional cells support HPV replication [4, 13]. HPV contamination persistency in this case prompt cervical intra-epithelial neoplasia (CIN) 2 or CIN3 injuries and, inevitably, improvement of intrusive cervical disease[4, 14]. A few components like long haul utilization of oral contraceptives, smoking, number of children, previously it was a sexually transmitted diseases (STIs) with chlamydia trachomatis, few herpes infections and HIV increments the danger of cervical malignancy among Human Papiloma Virus DNA confirming females [13]. Estrogen and its receptors are unequivocally connected with HPV diseases and enhances cervical malignant growth in aggregation with human papilloma virus oncogenes [13]. A recent study was made across Sweden incorporating 4245 patients with recently diagnosed cervical cancer [24]. The experiments explicit that psychiatric disorders and stressful life occasions around malignancy diagnosis are related with expanded cancer explicit mortality among patients with cervical cancer, autonomous tumor attributes and treatment methodology [24]. Another study was made with 70 patients in china with insomnia provoked by cervical cancer [1]. This irregular checked preliminary selected patients with sleep deprivation that arises or exacerbated by cervical disease [1].

Through psychoneuroimmunology relation between the psychological and physiological characteristics of cancer risk and progress has been considered [27]. The enduring activity of the hypothalamic-pituitary-adrenal (HPA) axis in the continuing stress reaction and in depression most likely weakens the resistance of the system and accelerates to the advancement of certain kinds of cancer growth [27].

According to American Cancer Institute, ovarian cancer is supposed to start in the ovaries but recent knowledge exhibits that numerous ovarian tumors may actually begin in the fallopian tubes, which holds two ovaries after the body of uterus. Likely cervical, ovarian cancer is hard to recognize [17]. The ovaries lie profound inside the abdominopelvic pit, making them hard to view or feel [17]. Epithelial ovarian disease stays an exceptionally dangerous [18]. According to American Cancer Institute, ovarian cancer are supposed to start in the ovaries but recent knowledge exhibits that numerous ovarian tumors may actually begin in the fallopian tubes, which holds two ovaries after the body of uterus. Regardless of serious research endeavors over the previous decade directed toward improved location and treatment of ovarian malignant growth, most of the ladies determined to have ovarian malignant growth surrender to the deadly disease [18]. It was at first accepted that ovarian malignant growth needed admonition signs, in spite of the fact that we currently realize that there are unpretentious side effects that may recommend sickness [17]. Another obstacle to early recognition is the absence of approved screening instruments to recognize malady [17]. The Papanicolaou ("Pap") test is utilized to screen for cervical cancer and the mammogram to screen for breast cancer, however there is no approved and strong test that can distinguish ovarian disease [13, 17].

A study was made using the data of Nurses' Health Study found a substantial relationship between treatment for PTSD and growth of ovarian cancer [20]. It has shown that in a experiment, use of stress hormones activates inflammation in ovarian epithelial or cancer cells [20].

1.2. Motivation of Research

Cancer is stimulated by stress which is a monstrous issue for the scientists due to lack of loyal treatment to fix the overwhelming sickness in people since antiquated

time. In spite of the concept that the improvement of new era in cancer growth has been multiplying step by step, however there is still prerequisite of disease to overcome the treatment [9]. There is no solution for disease after totally influenced [9]. Death is inescapable [9]. Lung cancer growth the mostly well-known reason for cancer death around the world [8]. The event of lung cancer has expanded quickly and turn the most widely recognized disease in men in many nations [8]. Lung cancer growth represents around 1,095,000 new disease cases and 951,000 deaths every year in men, and 514,000 cases and 427,000 deaths in women [8]. A full research was made by Kawser Ahmed to build up a framework that can be utilized by an individual to test his risk level of Lung Cancer [8]. And utilizing the acquired knowledge an experiment was able to predict the risk level of lung cancer [9]. Skin protects your body from the many viruses and bacteria that we are exposed to daily. A research was carried out to build up a system that can be utilized by an individual for knowing his risk level of skin cancer [10]. In South Asia like India, Bangladesh, Nepal, Myanmar, Pakistan, and Tibet and so on 76,000 female died for breast cancer in a year [5]. A single study using data of 160 patients from find out the impact of different factors [5]. That work shows the correlation among the factors of breast cancer [5].

Presently Type-1 Diabetes is also a shocking sickness in Bangladesh. Type 1 diabetes, which is known as adolescent diabetes or insulin-subordinate diabetes, is an interminable condition where pancreas delivers mostly zero insulin. With 306-man information (Case bunch 152 and Control Group-154) has been gathered from Dhaka dependent on a particular questionnaire to show the association and criticalness among the degree of elements [6]. When abnormal chemical reactions in the body modify the normal metabolic process causes metabolic disorder. A survey was carried out on obesity, type-2 diabetics, hypertension, cardiovascular disease, it is portrayed that they

might be precisely interconnected with metabolic disorder [15]. Obesity is a perplexing disease including an extreme amount of body fat. In Bangladesh, obesity has become an enormous issue [23]. Obesity is a term applied to overabundance body weight with a strangely high extent of body fat [23]. This ailment additionally produces type 2 diabetes, high blood cholesterol, hypertension, heart issues [23]. In an analysis, a dataset of 259 people was undertaken to find out risk factors of obesity using data mining and statistical analysis was made to find out more concerned factors of obesity [23].

1.3. Research Gap

While conducting a research many question or problem arises on a topic or area, which has not been addressed suitably or at all given in a field of study, known as research gap.

- a. There is no risk factor developed for neurodegenerative disease along with cervical and ovarian cancer.
- b. No application (mobile app or website) for this disease has been created for prediction.
- c. No analysis ranking is done for neurodegenerative disease along with cervical and ovarian cancer.

1.4. Research Question

Initial step of research is an answerable investigation to a particular issue namely research question.

- a. How can I find risk factors for this disease along with cervical and ovarian cancer?
- b. How to make an application based on the findings?
- c. How to make prediction for this disease along with cervical and ovarian cancer?

- d. How to analysis and rank through analysis?

1.5. Research Objectives

The expectation to achieve from a research is research objective. Aim is to find the information and understanding that is needed so as to respond to my research question.

- a. To detect significant risk factors of neurodegenerative disease.
- b. To make an application based on machine learning.
- c. To detect risk level of neurodegenerative disease.
- d. To analyze the risk factors and find out the co-relation between them.

1.6. Research Scope

Scope refers to the territory of how far the research has been analyzed i.e. each one of those things that will be covered in this research. My research scopes are,

- a. Data Mining
- b. Bio-informatics
- c. Machine Learning
- d. Bio-statistics
- e. Data Science

1.7. Thesis Organization: In the following sections, at first the previous work were reviewed in section 2. In section 3, the risk prediction models and their techniques behind prediction were discussed elaborately. We conduct experiments on three datasets in section 4 and it was conducted with the help of knowledge discovery. Their efficiency on prediction is shown with a set of figures and tables. This section also contains the output of our research which is the mobile application that I prepared for risk prediction. For preparing this application at first I made an equation to differentiate

risk level and the prepared an algorithm. The algorithm is provided in section 4. Finally, this work is concluded on section 5 and future work is proposed on section 6.

2. LITERATURE REVIEW

A search or inquiry of the accessible literature in the particular issue or topic or area of research or theory. By doing so, a research problem is explored and depiction, summary and critical evaluation is obtained of the matter.

A significant concern is air contamination (inside and outside) for public wellbeing. Numerous studies emphasizes its dangerous effects on health conditions [2]. In recent years it is also viewed as a vigorous risk factor of neuropathology [2]. Continuing studies have also characterizes that environmental or behavioral factors enhances the risk for Parkinson's disease [28]. Again, neurotoxic metals, for example, lead, mercury, aluminum, cadmium and arsenic, certain pesticides and metal-based nanoparticles have been associated with AD[11]. The danger to lead, manganese, solvents and a few pestiferous has been identified as risk factors of PD[11]. A review illustrates dangerous impact of environmental factors in neurodegenerative disease and its long term potential [2]. Also examined use air pollutant exposures and unexpected brain effects of urban people [2]. Ecological etiologies are connected with two of the most well-known neurodegenerative diseases, AD and PD have been verified from a research[11]. An study provides a report on the of Parkinson's disease, afterwards they also focuses on epidemiological advances of the most recent 10 years and they suggested for Parkinson's disease avoidance and treatment [28]. Progressively mild traumatic brain injury (TBI) is a well-settled risk factor for some neurodegenerative diseases including Alzheimer's infection, Parkinson's ailment, and amyotrophic parallel sclerosis (ALS)[12]. Recent survey summarizes the prevailing literature with respect to the epidemiology of chronic traumatic encephalopathy, post-TBI dementias and

Parkinson's disease, and the epidemiology of MTBI and neurodegenerative consequences after traumatic brain injury [12, 29]. A study estimates the risks of neurodegenerative disease from idiopathic rapid eye movement (REM) sleep behavior disorder (RBD) [25]. Genes are also associated with Alzheimer disease. A survey found out some genes linked with Alzheimer disease. apoE 4 allele of apoprotein E is a critical risk factor of Alzheimer disease [16]. Investigation of whole genome uncovered the genes increments the risk of Alzheimer disease: PLD3 and TREM2 [3]. Also audits the connection between AD risk genes and cellular and neuropathologic aspects of AD [3]. Decoding the administrative code of gene expression and comprehend the transcriptional impacts of genome variety has become one of the major challenges of the past decades [7]. A system was developed using deep learning-based framework namely ExPecto, that can precisely predict, ab initio from a DNA arrangement, the tissue-explicit transcriptional impacts of mutations, including those that are uncommon or that have not been watched [7].

These disease are mostly dangerous. Most of these have no remedy after a certain period of time. Predicting the condition plays a significant role in the diagnosis process. A paper describes about the usage of machine learning techniques for predicting disease advancement and severe diseases like age-related macular degeneration [22]. In a undertaken survey, comparison between different machine learning techniques e.g. decision tree, K nearest neighbors, random forest [14]. Modern technology has facilitated the prediction of harmful diseases. For this different prediction models were introduced. To integrate these models with medical rules a system was developed called PRIME [26]. Vikas Chaurasia presented a report where they used accessible technology to establish prediction models for breast cancer using Naïve Bayes, RBF Network, J48 algorithms [21].

3. RESEARCH METHODOLOGY

This paper uses well-known data mining and machine learning model were compared with metrics such as, accuracy, precision, recall, F1, support [14]. This was find using sklearn library of python and orange machine learning and data mining toolkit. We further propose an equation based on the difference between the result metrics of these two toolkits. Using apriori algorithm correlation among the significant factors which describes the dependency among the factors. Feature selection was performed using ranker algorithm. Key factors on the data analysis was derived for all the evaluators of ranker algorithm [6]. Afterwards it was compared among them and the most worthy attribute was obtained. For prediction it is important to find out the significant factors. Here, importance of factors has been gathered according to info gain, gain ratio, gini index [6].

3.1. Data Collection

In total 866 data were collected from various diagnosis center of patients suffering from diseases like, ovarian, cervical and stress disorder. Data of 161 female patients were collected those who were experiencing cervical cancer. Data was collected from a set of questionnaire which includes 26 attributes. 522 patients of ovarian cancer were interviewed with a set questions which contains 47 risk factors. Again those people who had stress disorder were inquired with a survey form of 15 questions and 161 patients of NINH provided their valuable information.

3.2. Data Preprocessing

Data cleaning, data integration, data selection, data transformation are four leading tasks of data pre-processing to convert the dataset from noisy, inconsistent data to a format suitable for mining and learning predictability. In data cleaning phase conflicting and inconsistent data were removed. Valuable data were joined in data

integration phase. Data suited for the analysis were retrieved from dataset in data selection phase. Finally in data transformation, data were converted to proper structures fits for data mining and machine learning analysis. To ignore collision of the data a small amount of data were altered [4].

3.3. Evaluation of the performance of Machine Learning Models

In this lesson, eight classifiers known as SVM, Random Forest, Logistic Regression, AdaBoost, Naïve Bayes, Neural Network, kNN, CN2 rule Inducer were used for evaluation with orange and almost 10 classifiers namely SVM, Random Forest, Logistic Regression, AdaBoost, Naïve Bayes, Neural Network, kNN, Gaussian Process, Decision Tree, Quadratic Classifier were used for assessment of machine learning models. In this context the performance were measured using standard metrics like area under ROC curve (AUC), precision, classification accuracy, recall, specificity, F measure, support. Additionally to compare the performance Receiver Operation Characteristics (ROC) had been engaged. A decision tree were constructed with the important factors of cervical, ovarian and stress datasets.

a. Performance measures: Classification accuracy rates for the datasets were analyzed.

For each dataset two classes were identified namely positive and negative. There are four possibilities for a single prediction e.g. true positive, true negative, false positive, false negative. True positive and true negative described as how many correct predictions were made. False positive and false negative provides how many incorrect predictions were made of positive and negative classes when they actually belong to positive and negative classes.

i. Accuracy: It defines the number of correct predictions that were correctly classified from total number of predictions in ratio.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

ii. Precision: Explains the number of positive predictions that were correctly classified by the classifier from total number of positive predictions.

$$\mathbf{precision} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}}$$

iii. Recall: Characterizes the fraction of correct positive predictions of the whole equation.

$$\mathbf{recall} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$$

iv. F-measure: It predicts on the average value of precision and recall.

$$\mathbf{F - measure} = \frac{\mathbf{2} \times (\mathbf{precision} \times \mathbf{recall})}{\mathbf{precision} + \mathbf{recall}}$$

v. Specificity: Measures the number of whole negative prediction those are accurately identified by the classifier.

$$\mathbf{Specificity} = \frac{\mathbf{TN}}{\mathbf{TN} + \mathbf{FP}}$$

vi. Support: Defines the number of datasets that were analyzed after training and splitting the whole dataset. From my analysis it is seen that sklearn uses only 24% of the whole dataset that were used.

$$\mathbf{Support} = \mathbf{0.24} \times \mathbf{no. of data in the dataset}$$

vii. Receiver Operating Characteristics curve (ROC): Drawing true positive rate vs false positive rate in a graph portrays performance of classification model can be which is known as ROC curve. True positive rate is a synonym of recall and false positive rate is also known as specificity.

viii. Area under the ROC curve (AUC): Measurement of two-dimensional area below the whole ROC curve from (0,0) to (1,1). A combining measure of performance is provided by AUC of the models.

3.4. Comparison of Feature Selection with individual Evaluators

In this research, seven attribute inspectors namely classifier attribute eval, correlation attribute eval, gain ratio attribute eval, info gain attribute eval, oneR attribute eval, relief attribute eval, symmetrical uncert attribute eval of ranking algorithm were used for documenting the strongly associated risk factors with the diseases. These evaluator ranks between the parameters based on the impact of causing the disease. The results were compared with each other and the most significant and important factors were taken into account.

3.5. Deriving Key Features

The first and foremost concept to judge the probability is to find the significant factors through various analysis. The study was undertaken with lots of derivatives and algorithms to find out the significant factors. The level of important factors were acquired using information gain, gain ratio, gini index.

3.5.1. Information Gain: It is a measurement of the decrease of uncertainty. It is estimated from entropy. Entropy is the measurement of probability of changeability of the processed information. The higher the entropy, the harder it is to make any determinations from that data.

$$\text{entropy}(p_1, p_2 \dots p_n) = -p_1 \log p_1 - p_2 \log p_2 \dots \dots - p_n \log p_n$$

$$E(x) = \sum_{i=1}^c -p_i \log p_i$$

$$E(T, X) = \sum_{c \in X} P(c)E(c)$$

Summation of the feature probability of values times the log probability of same label. By deducting the value of label and features from entropy of label information gain is obtained.

$$\text{Gain}(T, X) = E(T) - E(T, X)$$

3.5.2. Gain Ratio: Alters information gain by taking ignored information into account including the number and sizing of the branches that reduces the bias of information gain.

$$\text{SplitEntropy}(T, X) = \sum_{c \in X} -\frac{T_x}{T} \log \frac{T_x}{T}$$

$$\text{Gain Ratio} = \frac{\text{Gain}(S, A)}{\text{SplitEntropy}(S, A)}$$

3.5.3. Gini Index: It measures the impurity of a single feature. It is obtained by subtracting the sum of squared probabilities from one.

$$\text{Gini}(X) = 1 - \sum_{i=1}^c (p_i)^2$$

From information gain we acquired the certainty of individual features for a specific label. Gain ratio provides us the same including the intrinsic information of the dataset. Gini index provides how much filthy an individual factor is. All of these values are gathered in terms of 0 and 1.

3.6. Equation

Analyzing chi-square test and results of feature selection evaluators I found out the most significant factors which is working behind cervical and ovarian cancer with in connection with stress. Then these factors were given different scores based on their significance level. Afterwards the following equation was defined to separate the risk levels of an individual.

$$\text{Prediction difference} = (\sum \text{highest} - \sum \text{lowest}) \div 4$$

4. RESULT AND DISCUSSION

We have got 160 people provided data for stress questionnaire where 112 patients experienced to diseases for stress. For cervical cancer data 298 women were

collected where 198 patients were experienced cervical cancer. Among 522 data's 267 women were suffering from ovarian cancer. Here total analysis of the datasets were analyzed with different models of machine learning and then their performance were measured with various metrics. Table 1, Table 2, table 3 depicts the conducted study on machine learning models of stress, cervical and ovarian cancer. A rule based AI strategy were used for finding attractive relations between factors with regards to huge databases. Stress data is not yet analyzed. But it has some co-factors related to both cervical and ovarian cancer. The important risk factors were leveled according to info gain, gain ratio and gini index shown in table 4, table 5 and table 6. So far we have got some important variables of both cervical and ovarian cancer. Here, number of children, age, age of husband, age of first intercourse and a non significant parameter marital status for cervical cancer are also shown in box plot.

Table 1. Accuracy of machine learning classifiers of stress

Model	Tool	AUC	CA	F1	Precision	Recall	Specificity	Support
SVM	Orange	0.746	0.744	0.698	0.736	0.744	0.462	-
	Sklearn	0.615	-	0.600	0.590	0.620	-	39
Random Forest	Orange	0.746	0.725	0.707	0.705	0.725	0.525	-
	Sklearn	0.692	-	0.700	0.730	0.690	-	39
Logistic Regression	Orange	0.796	0.769	0.752	0.757	0.769	0.579	-
	Sklearn	0.769	-	0.770	0.760	0.770	-	39
AdaBoost	Orange	0.744	0.788	0.785	0.783	0.788	0.683	-
	Sklearn	0.790	-	0.790	0.790	0.790	-	39
Naïve Bayes	Orange	0.753	0.763	0.763	0.763	0.763	0.672	-
	Sklearn	0.785	-	0.780	0.790	0.780	-	39
Neural Network	Orange	0.767	0.750	0.747	0.745	0.750	0.631	-
	Sklearn	0.692	-	0.570	0.480	0.690	-	39
kNN	Orange	0.646	0.669	0.627	0.618	0.669	0.394	-
	Sklearn	0.615	-	0.600	0.590	0.620	-	39
CN2 rule Inducer	Orange	0.721	0.694	0.696	0.699	0.694	0.595	-
	Sklearn	-	-	-	-	-	-	-
Decision Tree	Orange	0.700	0.713	0.717	0.723	0.713	0.639	-
	Sklearn	0.692	-	0.700	0.710	0.690	-	39
Quadratic Classifier	Orange	-	-	-	-	-	-	-
	Sklearn	0.744	-	0.700	0.750	0.740	-	39

Again for convenience certain factors were taken to draw box plots. Figure 1-13 shows elaborate analysis of the significant factors from the findings. A decision tree along with significant factor probabilities were shown as figure. Figure 7, 12 are the predicted probabilities and figure 8, 13 were trees of cervical cancer and ovarian cancer.

Table 2. Accuracy of machine learning classifiers of ovarian cancer.

Model	Tool	AUC	CA	F1	Precision	Recall	Specificity	Support
SVM	Orange	0.883	0.742	0.704	0.741	0.742	0.8	-
	Sklearn	0.861	-	0.85	0.87	0.85	-	72
Random Forest	Orange	0.868	0.755	0.745	0.744	0.755	0.841	-
	Sklearn	0.972	-	0.97	0.97	0.98	-	72
Logistic Regression	Orange	0.863	0.735	0.721	0.72	0.735	0.828	-
	Sklearn	1	-	1	1	1	-	72
AdaBoost	Orange	0.86	0.742	0.737	0.737	0.742	0.828	-
	Sklearn	1	-	1	1	1	-	72
Naïve Bayes	Orange	0.851	0.621	0.627	0.642	0.621	0.836	-
	Sklearn	0.958	-	0.96	0.96	0.96	-	72
Neural Network	Orange	0.847	0.718	0.719	0.721	0.718	0.838	-
	Sklearn	0.986	-	0.99	0.99	0.98	-	72
kNN	Orange	0.845	0.735	0.725	0.723	0.735	0.833	-
	Sklearn	0.861	-	0.85	0.87	0.85	-	72
CN2 rule Inducer	Orange	0.821	0.674	0.675	0.676	0.674	0.815	-
	Sklearn	-	-	-	-	-	-	72
Decision Tree	Orange	-	-	-	-	-	-	-
	Sklearn	0.986	-	0.99	0.98	0.99	-	72
Quadratic Classifier	Orange	-	-	-	-	-	-	-
	Sklearn	0.431	-	0.3	0.22	0.5	-	72

Different attribute evaluator were used for ranking the parameters based on the significance of enhancing the diseases. Seven attribute inspector were used and result were analyzed to find out important factors. Table 5 and Table 6 shows the ranking between the factors. From these ranking of attributes by those evaluators it is possible

to find the required vital factors that we need in doing our prediction. Children, age, age of first intercourse, husband's age, use of oral contraceptives were valuable factors for cervical cancer.

Table 3. Accuracy of machine learning classifiers of Cervical Cancer.

Model	Tool	AUC	CA	F1	Precision	Recall	Specificity	Support
SVM	Orange	0.921	0.835	0.835	0.841	0.835	0.838	-
	Sklearn	0.761	-	0.76	0.77	0.76	-	126
Random Forest	Orange	0.926	0.843	0.843	0.843	0.843	0.843	-
	Sklearn	0.778	-	0.78	0.79	0.78	-	126
Logistic Regression	Orange	0.933	0.848	0.848	0.848	0.848	0.848	-
	Sklearn	0.793	-	0.79	0.79	0.79	-	126
AdaBoost	Orange	0.900	0.827	0.827	0.827	0.827	0.827	-
	Sklearn	0.762	-	0.76	0.76	0.76	-	126
Naïve Bayes	Orange	0.921	0.750	0.740	0.790	0.750	0.741	-
	Sklearn	0.785	-	0.78	0.79	0.78	-	126
Neural Network	Orange	0.912	0.814	0.814	0.814	0.814	0.813	-
	Sklearn	0.770	-	0.77	0.77	0.77	-	126
kNN	Orange	0.917	0.839	0.839	0.839	0.839	0.839	-
	Sklearn	0.761	-	0.76	0.77	0.76	-	126
CN2 rule Inducer	Orange	0.912	0.816	0.816	0.816	0.816	0.815	-
	Sklearn	-	-	-	-	-	-	-
Decision Tree	Orange	0.773	0.835	0.835	0.835	0.835	0.835	-
	Sklearn	0.754	-	0.75	0.76	0.75	-	126
Quadratic Classifier	Orange	-	-	-	-	-	-	-
	Sklearn	0.762	-	0.76	0.77	0.76	-	126

Again, in case of ovarian cancer the most essential factors we get from analyzing info gain, gain ratio, gini index is those women who had problem during pregnancy, taken

abortion, having infection in genital area, lately menopause, affected by cervical cancer and breast cancer already and related variables are absorption of tentative drug etc. The following table 4 shows the detailed analysis of information gain, gain ratio and gini index on stress.

Table 4: Info gain, gain ratio, gini index and chi-square test of stress.

Attributes	#	Info. gain	Gain ratio	Gini	χ^2
stress treat diseases?	2	0.124	0.289	0.076	2.526
occupation	16	0.095	0.037	0.045	2.913
impact on disorder?	2	0.087	0.109	0.054	5.025
impact on depression/scizophrenia?	2	0.069	0.071	0.038	8.416
religion	4	0.041	0.047	0.022	0.222
stress impact on memory?	2	0.039	0.042	0.023	2.920
impact on health?	2	0.032	0.045	0.020	1.524
Age		0.024	0.012	0.013	0.009
stress idea?	2	0.023	0.032	0.014	1.085
family member	4	0.016	0.012	0.009	1.323
event cause stress?	2	0.010	0.040	0.007	0.112
monthly income	4	0.004	0.003	0.002	0.220
impact on anxiety?	2	0.003	0.059	0.001	0.003
sex	2	0.000	0.000	0.000	0.004

The above table shows the model accuracy of various machine learning algorithms. Here, logistics regression shows the highest accuracy with highest area under Roc curve. Fig 4, Fig 5, Fig 6, Fig 10 displays the flowcharts that I derived from my analysis. For the both machine learning and data mining algorithms were used for the derivation. From the results obtained from feature selection of ranker algorithm a data mining algorithm, the results were considered with the chi- square test result. Afterwards both of the significances were taken into account which is used to derived scoring patterns. Then the scores I given to the most significant factors were used for risk prediction. Following the algorithm, I have drawn flowcharts that is required to use in the application for predicting the risks of diseases.

Table 5: Importance of factors in data table based on info gain, gain ratio and gini index of cervical cancer.

Attributes	#	Info. Gain	Gain ratio	Gini	χ^2
Children	3	0.634	0.481	0.321	100.508
Age_of_husband	5	0.579	0.372	0.314	119.297
Age	4	0.545	0.312	0.302	19.214
Age_first_sex	2	0.453	0.453	0.219	74.112
Condom_or_diaphragm	2	0.296	0.385	0.182	93.685
Oral_contraceptics	2	0.253	0.316	0.158	25.445
Education_level	4	0.192	0.102	0.113	84.508
eat_fruits	2	0.146	0.183	0.067	10.750
Had_hysterectomy	2	0.143	0.194	0.060	31.472
Knowledge_of_cervical_cancer	2	0.139	0.257	0.088	51.066
family_had_cervical	2	0.117	0.175	0.048	26.396
Occupation	6	0.114	0.075	0.058	3.996
Area_of_residence	3	0.097	0.067	0.063	7.789
Family_members	3	0.087	0.067	0.054	6.773
vaccine_carvical_cancer	2	0.072	0.276	0.040	25.610
diethylstilbestrol	2	0.068	0.139	0.027	16.244
Sex_partner	2	0.065	0.115	0.029	16.897
Expenses	3	0.055	0.081	0.024	2.267
Sexual_infections	2	0.046	0.121	0.018	11.168
Marital_status	4	0.041	0.073	0.017	0.636
Smoking	2	0.031	0.107	0.012	7.614
Family_income	4	0.028	0.015	0.018	0.462
Social_status	4	0.023	0.013	0.014	0.728
Pap_test	2	0.016	0.091	0.006	4.061
Any_type_of_cancer	2	0.008	0.078	0.003	2.030

After analyzing the significances of the factor of cervical, ovarian and stress I derived an algorithm for predicting the risk levels of the diseases which is,

- 1: Start
- 2: read weights
- 3: $total_weights \leftarrow \sum weights$
- 4: $pd \leftarrow (\sum highest - \sum lowest) \div 4$
- 5: if $total_weights \leq pd + \sum lowest$ then print LOW RISK
- 6: else if $total_weights \leq (pd*2) + \sum lowest$ then print MEDIUM RISK
- 7: else if $total_weights \leq (pd*3) + \sum lowest$ then print HIGH RISK
- 8: else print VERY HIGH RISK
- 9: Stop

Table 6. Importance of factors in data table based on info gain, gain ratio and gini index of ovarian cancer.

Attributes	#	Info. Gain	Gain ratio	Gini	χ^2
Problem During Pregnancy?	2	0.408	0.414	0.250	148.636
Abortion?	2	0.355	0.355	0.225	118.676
Have Infection in Genital Area	2	0.342	0.356	0.213	84.280
Menopause	3	0.270	0.181	0.153	47.347
Affected By cervical Cancer?	2	0.264	0.308	0.160	47.090
Affected By Breast Cancer	2	0.251	0.278	0.157	52.243
Knowledge about ovarian cancer?	2	0.247	0.296	0.149	114.100
Take tentative drug?	2	0.246	0.281	0.152	111.921
Close Relative affected by Breast Cancer	2	0.241	0.287	0.146	40.581
Had Breast Cancer Prior to Age 40	2	0.240	0.284	0.146	41.612
Estrogen Pill taking after Menopause	2	0.237	0.291	0.142	110.485
Menopause after 50?	2	0.231	0.276	0.141	39.288
Ever Had a Hysterectomy	2	0.228	0.268	0.141	40.543
Condom/ Diaphragm	2	0.221	0.274	0.134	104.769
Any Birth Control Pill?	2	0.221	0.232	0.143	93.780
Use Napkin	2	0.220	0.263	0.135	103.446
Any Replacement Therapy Taken	2	0.214	0.249	0.134	100.124
Pregnancy after 35	2	0.212	0.245	0.133	39.975
pap test?	2	0.211	0.246	0.132	99.114
Children after 35	2	0.208	0.237	0.131	40.476
Regular Exercise?	2	0.206	0.213	0.135	85.197
Takes Hormone After Menopause	2	0.206	0.251	0.127	98.322
Diet Maintain	2	0.200	0.238	0.125	95.336
Cancer History In family?	2	0.197	0.215	0.127	43.440
Take adequate fruit	2	0.185	0.187	0.122	72.132
STI?	2	0.171	0.200	0.110	32.078
Food contains high fat?	2	0.113	0.151	0.071	58.026
Age	3	0.107	0.148	0.060	46.190
Menopause End age?	3	0.100	0.063	0.067	7.758
Oral Contraception	2	0.099	0.163	0.059	52.634
Education	3	0.070	0.058	0.047	3.586
Age of Husband	4	0.069	0.081	0.042	5.796
BMI?	3	0.061	0.085	0.039	6.488
Social Class	3	0.055	0.045	0.037	1.028
Height?	3	0.013	0.016	0.009	13.957
Family Members	3	0.012	0.010	0.008	0.293
Cancer Vaccin Taken?	2	0.007	0.031	0.005	4.727
First sex age?	2	0.005	0.005	0.003	1.736
Source of Knowledge?	3	0.003	0.002	0.002	0.010
Marital Status?	2	0.002	0.099	0.001	0.002
How Many Children?	3	0.002	0.002	0.001	0.101
Never Pregnant?	2	0.001	0.005	0.001	0.701
smoker?	2	0.001	0.004	0.000	0.441
Obese?	2	0.000	0.005	0.000	0.253
Take alcohol?	2	0.000	0.000	0.000	0.011
Number of sex partner?	1				nan
Feature 1		nan	nan	nan	nan

Table 7. Comparison between the factors using different algorithms of stress

Classifier Attribute Eval	Correlation Attribute Eval	Gain Ratio Attribute Eval
I. impact on health?	I.stress treat diseases?	I.stress treat diseases?
II.monthly income	II.impact on disorder?	II.impact on disorder?
III.sex	III.impact on depression/scizophrenia?	III.impact on depression/scizophrenia?
IV.family member	IV.stress impact on memory?	IV.impact on anxiety?
V.impact on depression/scizophrenia?	V.impact on health?	V.religion
VI.occupation	VI.stress idea?	VI.impact on health?
VII.religion	VII.event cause stress?	VII.stress impact on memory?
VIII.stress idea?	VIII.family member	VIII.event cause stress?
IX.event cause stress?	IX.monthly income	IX.occupation
X.stress treat diseases?	X.impact on anxiety?	X.stress idea?
XI.impact on anxiety?	XI.occupation	XI.family member
XII.stress impact on memory?	XII.religion	XII.monthly income
XIII.impact on disorder?	XIII.Age	XIII.sex
XIV.Age	XIV.sex	XIV.Age

Table 7.1. Comparison between the factors using different algorithms of stress

Info Gain Attribute Eval	OneR Attribute Eval
I.stress treat diseases?	I.stress treat diseases?
II.occupation	II.impact on disorder?
III.impact on disorder?	III.religion
IV.impact on depression/scizophrenia?	IV.Age
V.religion	V.impact on depression/scizophrenia?
VI.stress impact on memory?	VI.monthly income
VII. impact on health?	VII.sex
VIII.stress idea?	VIII.family member
IX. family member	IX.occupation
X.event cause stress?	X.stress impact on memory?
XI.monthly income	XI.impact on anxiety?
XII. impact on anxiety?	XII.event cause stress?
XIII. sex	XIII.stress idea?
XIV. Age	XIV.impact on health?

Table 7.2. Comparison between the factors using different algorithms of stress

Relief Attribute Eval	Symmetrical Uncert Attribute Eval
I.family member	I.stress treat diseases?
II.impact on depression/scizophrenia?	II.impact on disorder?
III.impact on health?	III.impact on depression/scizophrenia?
IV.stress treat diseases?	IV.occupation
V.occupation	V.religion
VI.sex	VI.stress impact on memory?
VII.impact on disorder?	VII.impact on health?
VIII.monthly income	VIII.stress idea?
IX.religion	IX.event cause stress?
X.stress impact on memory?	X.family member
XI.Age	XI.impact on anxiety?
XII.stress idea?	XII.monthly income
XIII.event cause stress?	XIII.sex
XIV.impact on anxiety?	XIV.Age

With the help of above algorithm I found out the respective flowcharts for the diseases. At last I put all the flowcharts and significant factors together to elicit the superior significant factors. Afterwards combining cervical, ovarian and stress factors I drawn the flowchart for all of them. From those flowcharts and using the significant factors I prepared an application that can be used for cervical and ovarian cancer prediction. Fig 18, Fig 19, Fig 20 shows implementation of my research in an android application.

Table 8. Comparison between the factors using different algorithms of cervical cancer.

Classifier Attribute Eval	Correlation Attribute Eval	Gain Ratio Attribute Eval	Info Gain Attribute Eval
A.family_had_cervical	A.Children	A.Children	A.Children
B. Any_type_of_cancer	B.Age_first_sex	B.Age_first_sex	B.Age_of_husband
C. Social_status	C.Condom_or_diaphragm	C.Condom_or_diaphragm	C.Age
D. Family_income	D.Oral_contraceptives	D.Age_of_husband	D.Age_first_sex
E. Expenses	E.Age_of_husband	E.Oral_contraceptives	E.Condom_or_diaphragm
F. Family_members	F. Age	F.Age	F.Oral_contraceptives
G. Occupation	G.Knowledge_of_cervical_cancer	G.vaccine_carvical_cancer	G.Education_level
H. Education_level	H. eat_fruits	H.Knowledge_of_cervical_cancer	H.family_had_cervical
I. Age	I. Had_hysterectomy	I.Pap_test	I.eat_fruits
J. Marital_status	J.family_had_cervical	J.family_had_cervical	J.Had_hysterectomy
K. Age_of_husband	K. vaccine_carvical_cancer	K.Had_hysterectomy	K.Knowledge_of_cervical_cancer
L. Knowledge_of_cervical_cancer	L.Education_level	L.diethylstilbestrol	L.Occupation
M. Vaccine_carvical_cancer	M. Sex_partner	M.Sexual_infections	M.diethylstilbestrol
N. diethylstilbestrol	N.diethylstilbestrol	N.eat_fruits	N.Area_of_residence
O. Had_hysterectomy	O.Expenses	O.Sex_partner	O.Sex_partner
P. Children	P. Sexual_infections	P.Smoking	P.Family_members
Q. Sexual_infections	Q. Marital_status	Q.Education_level	Q.Sexual_infections
R. Pap_test	R.Area_of_residence	R.Expenses	R.vaccine_carvical_cancer
S. Oral_contraceptives	S. Smoking	S.Any_type_of_cancer	S.Expenses
T. Condom_or_diaphragm	T. Family_income	T.Occupation	T.Pap_test
U. Age_first_sex	U. Family_members	U.Marital_status	U.Marital_status
V. eat_fruits	V.Pap_test	V.Family_members	V.Family_income
W. Smoking	W. Occupation	W.Area_of_residence	W.Smoking
X. Sex_partner	X. Social_status	X.Family_income	X.Social_status
Y. Area of residence	Y. Any type of cancer	Y.Social_status	Y.Any_type_of_cancer

Table 8.1 Comparison between the factors using different algorithms of cervical cancer.

OneR Attribute Eval	Relief Attribute Eval	Symmetrical Uncert Attribute Eval
A.Children	A.Children	A.Children
B.Age_of_husband	B.Age_of_husband	B.Children
C.Age	C.Age_first_sex	C.Age_of_husband
D.Condom_or_diaphragm	D.Age	D.Age
E.Age_first_sex	E.Condom_or_diaphragm	E.Condom_or_diaphragm
F.Oral_contraceptives	F.Education_level	F.Oral_contraceptives
G.Knowledge_of_cervical_cancer	G.Oral_contraceptives	G.Knowledge_of_cervical_cancer
H.Education_level	H.Family_members	H.family_had_cervical
I.Area_of_residence	I.Family_income	I.Had_hysterectomy
J.Family_members	J.Social_status	J.eat_fruits
K.vaccine_carvical_cancer	K.Area_of_residence	K.Education_level
L.Occupation	L.eat_fruits	L.diethylstilbestrol
M.Sexual_infections	M.Occupation	M.Sex_partner
N.Sex_partner	N.Knowledge_of_cervical_cancer	N.vaccine_carvical_cancer
O.Had_hysterectomy	O.family_had_cervical	O.Sexual_infections
P.diethylstilbestrol	P.Had_hysterectomy	P.Occupation
Q.Pap_test	Q.Expenses	Q.Pap_test
R.Marital_status	R.Sex_partner	R.Area_of_residence
S.Smoking	S.diethylstilbestrol	S.Family_members
T.family_had_cervical	T.vaccine_carvical_cancer	T.Expenses
U.Any_type_of_cancer	U.Pap_test	U.Marital_status
V.eat_fruits	V.Sexual_infections	V.Smoking
W.Expenses	W.Marital_status	W.Family_income
X.Social_status	X.Smoking	X.Social_status
Y.Family_income	Y.Any_type_of_cancer	Y.Any_type_of_cancer

Table 9. Comparison between the factors using different algorithms of Ovarian Cancer.

Classifier Attribute Eval	Correlation Attribute Eval	Gain Ratio Attribute Eval
A. Condom/ Diaphragm	A. Problem During Pregnancy?	A. Problem During Pregnancy?
B. Affected By Breast Cancer	B. Abortion?	B. Have Infection in Genital Area
C. Family Members	C. Have Infection in Genital Area	C. Abortion?
D. smoker?	D.. Affected By cervical Cancer?	D. Affected By cervical Cancer?
E. Affected By cervical Cancer?	E. Affected By Breast Cancer	E. Knowledge about ovarian cancer?
F. Height?	F. Take tentative drug?	F. Estrogen Pill taking after Menopause
G. How Many Children?	G. Knowledge about ovarian cancer?	G. Close Relative 24affected by Breast Cancer
H. Ever Had a Hysterectomy	H. Breast Cancer Prior to Age 40	H. Breast Cancer Prior to Age 40
I. Close Relative 24affected by Breast Cancer	I. Close Relative 24affected by Breast Cancer	I. Take tentative drug?
J. Breast Cancer Prior to Age 40	J.Any Birth Control Pill?	J. Affected By Breast Cancer
K. Menopause after 50?	K. Estrogen Pill taking after Menopause	K. Menopause after 50?
L. Any Replacement Therapy Taken	L. Menopause after 50?	L. Condom/ Diaphragm
M. pap test?	M. Ever Had a Hysterectomy	M. Ever Had a Hysterectomy
N. Estrogen Pill taking after Menopause	N. Use Napkin	N. Use Napkin
O. Menopause	O. Regular Exercise?	O. Takes Hormone After Menopause
P. Pregnancy after 35	P. Any Replacement Therapy Taken	P. Any Replacement Therapy Taken
Q. Children after 35	Q. Condom/ Diaphragm	Q. pap test?
R. Oral Contraception	R. Pregnancy after 35	R. Pregnancy after 35
S. Takes Hormone After Menopause	S. pap test?	S. Diet Maintain
T. Social Class	T. Children after 35	T. Children after 35
U. Use Napkin	U. Cancer History In family?	U. Any Birth Control Pill?
V. Diet Maintain	V. Takes Hormone After Menopause	V. Cancer History In family?
W. STI?	W. Diet Maintain	W. Regular Exercise?
X. Cancer History In family?	X. Take adequate fruit	X. STI?
Y. Education	Y. STI?	Y. Take adequate fruit
Z. Obese?	Z. Food contains high fat?	Z. Menopause
AA. BMI?	AA. Oral Contraception	AA. Oral Contraception
AB. Take tentative drug?	AB. Menopause	AB. Food contains high fat?
AC. Take alcohol?	AC. BMI?	AC. Age
AD. Knowledge about ovarian cancer?	AD. Education	AD. Marital Status?
AE. Age of Husband	AE. Menopause End age?	AE. BMI?
AF. Source of Knowledge?	AF. Age of Husband	AF. Age of Husband
AG. Have Infection in Genital Area	AG. Social Class	AG. Menopause End age?
AH. Marital Status?	AH. Age	AH. Education
AI. Take adequate fruit	AI. Height?	AI. Social Class
AJ. First sex age?	AJ. Cancer Vaccine Taken?	AJ. Cancer Vaccine Taken?
AK. Number of sex partner?	AK. Family Members	AK. Height?
AL. Menopause End age?	AL. First sex age?	AL. Family Members
AM. Never Pregnant?	AM. Marital Status?	AM. Never Pregnant?
AN. Cancer Vaccine Taken?	AN. Never Pregnant?	AN. First sex age?
AO. Problem During Pregnancy?	AO. Smoker?	AO. Obese?
AP. Food contains high fat?	AP. Obese?	AP. Smoker?
AQ. Any Birth Control Pill?	AQ. Source of Knowledge?	AQ. Source of Knowledge?
AR. Regular Exercise?	AR. How Many Children?	AR. How Many Children?
AS. Abortion?	AS. Take alcohol?	AS. Take alcohol?
AT. Age	AT. Number of sex partner?	AT. Number of sex partner?

Table 9.1 Comparison between the factors using different algorithms.

Info Gain Attribute Eval	OneR Attribute Eval
A. Problem During Pregnancy?	A. Problem During Pregnancy?
B. Abortion?	B. Abortion?
C. Have Infection in Genital Area	C. Have Infection in Genital Area
D. Menopause	D. Affected By Breast Cancer
E. Affected By cervical Cancer?	E. Any Birth Control Pill?
F. Affected By Breast Cancer	F. Affected By cervical Cancer?
G. Knowledge about overian cancer?	G. Regular Exercise?
H. Take tentative drug?	H. Take tentative drug?
I. Close Relative 25affected by Breast Cancer	I. Knowledge about overian cancer?
J. Breast Cancer Prior to Age 40	J. Breast Cancer Prior to Age 40
K. Estrogen Pill taking after Menopause	K. Take adequate fruit
L. Menopause after 50?	L. Close Relative 25affected by Breast Cancer
M. Ever Had a Hysterectomy	M. Ever Had a Hysterectomy
N. Condom/ Diaphram	N. Cancer History In family?
O. Any Birth Control Pill?	O. Menopause after 50?
P. Use Napkin	P. Menopause
Q. Any Replacement Therapy Taken	Q. Pregnancy after 35
R. Pregnancy after 35	R. Childern after 35
S. pap tesr?	S. Any Replacement Therapy Taken
T. Childern after 35	T. Estrogen Pill taking after Menopause
U. Regular Exercise?	U. pap tesr?
V. Takes Hormone After Menopause	V. Use Napkin
W. Diet Maintain	W. Condom/ Diaphram
X. Cancer History In family?	X. Diet Maintain
Y. Take adequate fruit	Y. Takes Hormone After Menopause
Z. STI?	Z. STI?
AA. Food contains high fat?	AA. Menopause End age?
AB. Age	AB. Education
AC. Menopause End age?	AC. Food contains high fat?
AD. Oral Contraception	AD. Oral Contraception
AE. Education	AE. Social Class
AF. Age of Husband	AF. Age
AG. BMI?	AG. Age of Husband
AH. Social Class	AH. BMI?
AI. Height?	AI. Height?
AJ. Family Members	AJ. Family Members
AK. Cancer Vaccine Taken?	AK. First sex age?
AL. First sex age?	AL. Cancer Vaccine Taken?
AM. Source of Knowledge?	AM. How Many Childern?
AN. Marital Status?	AN. Smoker?
AO. How Many Childern?	AO. Marital Status?
AP. Never Pregnant?	AP. Obese?
AQ. Smoker?	AQ. Number of sex partner?
AR. Obese?	AR. Take alcohol?
AS. Take alcohol?	AS. Never Pregnant?
AT. Number of sex partner?	AT. Source of Knowledge?

Table 9.2 Comparison between the factors using different algorithms.

Relief Attribute Eval	Symmetrical Uncert Attribute Eval
A. Problem During Pregnancy?	A. Problem During Pregnancy?
B. Have Infection in Genital Area	B. Abortion?
C. Menopause End age?	C. Have Infection in Genital Area
D. Abortion?	D. Affected By cervical Cancer?
E. Family Members	E. Knowledge about ovarian cancer?
F. Take tentative drug?	F. Affected By Breast Cancer
G. Menopause	G. Take tentative drug?
H. Knowledge about ovarian cancer?	H. Close Relative affected by Breast Cancer
I. Affected By Breast Cancer	I. Estrogen Pill taking after Menopause
J. Take adequate fruit	J. Breast Cancer Prior to Age 40
K. pap test?	K. Menopause after 50?
L. Affected By cervical Cancer?	L. Ever Had a Hysterectomy
M. Condom/ Diaphragm	M. Condom/ Diaphragm
N. Food contains high fat?	N. Use Napkin
O. Estrogen Pill taking after Menopause	O. Any Replacement Therapy Taken
P. Any Birth Control Pill?	P. pap test?
Q. Menopause after 50?	Q. Pregnancy after 35
R. Takes Hormone After Menopause	R. Any Birth Control Pill?
S. Education	S. Takes Hormone After Menopause
T. Use Napkin	T. Children after 35
U. How Many Children?	U. Diet Maintain
V. Diet Maintain	V. Menopause
W. Age of Husband	W. Regular Exercise?
X. Close Relative affected by Breast Cancer	X. Cancer History In family?
Y. Oral Contraception	Y. Take adequate fruit
Z. Breast Cancer Prior to Age 40	Z. STI?
AA. Regular Exercise?	AA. Food contains high fat?
AB. Any Replacement Therapy Taken	AB. Age
AC. Age	AC. Oral Contraception
AD. Ever Had a Hysterectomy	AD. Menopause End age?
AE. Social Class	AE. Age of Husband
AF. First sex age?	AF. BMI?
AG. STI?	AG. Education
AH. Pregnancy after 35	AH. Social Class
AI. Cancer History In family?	AI. Height?
AJ. Source of Knowledge?	AJ. Cancer Vaccine Taken?
AK. Take alcohol?	AK. Family Members
AL. Children after 35	AL. First sex age?
AM. Height?	AM. Marital Status?
AN. BMI?	AN. Source of Knowledge?
AO. Cancer Vaccine Taken?	AO. Never Pregnant?
AP. Never Pregnant?	AP. How Many Children?
AQ. Smoker?	AQ. Smoker?
AR. Obese?	AR. Obese?
AS. Marital Status?	AS. Take alcohol?
AT. Number of sex partner?	AT. Number of sex partner?

4.1. Knowledge Discovery

A technique to invent knowledge from different sources is known as knowledge discovery [30]. There were seven steps to discover knowledge [30].

- a) Data Cleaning: This step is basically for cleaning the data. WEKA, Orange cleaned the dataset at first before every operation. This is also done manually by sci-kit where I used train function which removes the uncategorized and boisterous data.
- b) Data Integration: Data relevant to each other were integrated. In my research I need to combine data having sub-factors e.g. to analyze stress dataset I used this step.
- c) Data Selection: Three datasets were complete and clear for analysis. The stress, cervical and ovarian datasets were used for my research.
- d) Data Transformation: After performing several analysis within these datasets, these were combined with most significant factors.
- e) Data Mining: These datasets were analyzed with the help of different machine learning models and different data mining algorithms. This is also done manually with the help of Sci-kit learn library of python.
- f) Pattern Evaluation: By analyzing these datasets three flowcharts and combining them one flowchart were prepared.
- g) Knowledge Presentation: Finally my output is described in a mobile application which is made based on a algorithm derived by me. The algorithm was prepared from a equation. This equation is also driven by me.

Table 10. Weightage value of Stress

Parameters	Sub-parameters	Weight/Scores
Stress treat disease	Yoga	9
	Physical Exercise	8
	Book Reading	8.5
	Family Bonding	8
	Recreation	8
	Good Working	8
Impact on disorder	no	7
	yes	8
Impact on depression or scizophrenia	Diarrhoea	7
	Change in appetite	6.8
	Asthma	6.8
	Feeling Sad	6.65
	Fatigue	6.65
	Chest Pain	6.5
	Muscular Tension	6.5
	Anger	6.5
	Headache	6.5
	Lack of Interest	6.5
	Feeling Nervous	6.65
	Dry Mouth	6.65
	Indigestion	6.65
	Coughs	6.5
Occupation	Govt Job	6
	Private Job	6.5
	Business	6.5
	Unemployed/ Housewife	6.25
Stress impact on memory	no	5.5
	yes	6
Impact on health	Schizophrenia	5.5
	Chronary Heart disease	5.25
	Multiple Sclerosis	5.25
	Obesity	5.25
	Sudden death	5.25
	Diabetes	5.25
	Depression	5
	Anxiety	5
	Alzheimer	5
	Heart Disease	5
	Heart Attacks	5
Family Members	4 or below	4.5
	5 to 6	4.75
	8 or above	5
Stress Idea	no	4
	yes	4.5
Monthly Income	21000 or above	3.5
	5000 to 10000	3.75
	11000 to 15000	4
	16000 to 20000	4
Religion	buddah	3
	hindu	3.15
	muslim	3.3
	christian	3.5
Age	Below 35	2.75
	40-50	3
	51 or above	2.5
Impact on anxiety	Yes	2
	no	2.5
Event cause Stress	Family Health/ Family Responsibilities	2
	Economy	1.75
	Serious Injuries	1.5
	Broken relation	1.75
	Loss of job/ Income	1.5
	Death/ Illness	1.5
	Serious Illness	1.5
Gender	Male	1.5
	Female	1

Table 11. Weightage value of Cervical Cancer

Parameters	Sub-parameters	Weight/Score
Children	1 to 2	11
	above 5	11.5
	3 to 5	12
Age first Sex	above 16	10
	below 16	11
Condom or diaphragm	Yes	9.5
	no	10
Education	Undergraduate	9
	Secondary	9.15
	primary	9.3
	Illiterate	9.5
Family had cervical	no	9.5
	yes	9
Vaccine of cervical cancer	yes	8
	no	8.5
Age	below 30	7.5
	31-45	7.75
	46-60	8
	above 60	8
Age of Husband	31-45	7
	below 30	7.5
	above 60	7.5
Pap test	no	6.5
	yes	7
Had hysterectomy	no	6
	yes	6.5
Knowledge of cervical cancer	yes	5.5
	no	6
Oral Contraceptives	no	5
	yes	5.5
diethylstilbestrol	no	4.5
	yes	5
Sex partner	3+	4.5
	1 to 2	4
Sexual Infections	no	3.5
	yes	4
Eat Fruits	yes	3
	no	3.5
Area of residence	rural	2.5
	urban	2.75
	suburb	3
Occupation	housewife	2.3
	govt. service	2
	shop keeper	2.5
	private job	2.15
	Business	2.5
	Unemployed	2.5
Marital Status	Married	1.5
	Widow	2
	Divorce	2
	Separate	1.75
Any type of cancer	yes	1.5
	no	1

Table 12. Weightage value of Ovarian Cancer

Parameters	Sub-parameters	Weight/Score
Problem during pregnancy	yes	17
	no	16
Abortion	yes	16
	no	15
Have Infection in Genital Area	yes	15
	no	14.5
Take tentative drug	no	14.5
	yes	14
Any birth control pill	no	14
	yes	13.5
Regular exercise	no	13.5
	yes	13
Affected by Breast Cancer	yes	13
	no	12.5
Affected by Cervical Cancer	yes	12.5
	no	12
Take adequate fruit	no	12
	yes	11.5
Food contains high fat	yes	11.5
	no	11
Estrogen pill taking after menopause	no	11
	yes	10.5
Cancer history in family	yes	10.5
	no	10
Condom or diaphragm	no	10
	yes	9.5
Knowledge about ovarian cancer	no	9.5
	yes	9
Pap test	no	9
	yes	8.5
Diet Maintain	no	8.5
	yes	8
Any replacement therapy taken	no	8
	yes	7.5
Use Napkin	no	7.5
	yes	7
Breast cancer prior to age 40	yes	7
	no	6.5
Menopause after 50	yes	6.5
	no	6
Close relative affected by breast cancer	yes	6
	no	5.5
Ever had a hysterectomy	yes	5.5
	no	5
Menopause	Late	5
	early	5
	normal	4.5
Pregnancy after 35	yes	4.5
	no	4
Oral Contraception	yes	4
	no	3.5
Takes hormone after menopause	no	3.5
	yes	3
BMI	25-28	3
	>=29	3
	<=24	2.5
Children after 35	yes	2.5
	no	2
Age	63 or more	2
	42-60	1.75
	below 40	1.5
Age of Husband	above 60	1.5
	46-60	1.3
	below 30	1.3
	31-45	1

Table 13. Weightage value of Stress, Cervical and Ovarian Cancer

Parameters	Sub-parameters	Weight/Score
Long term pressure	above 3	15
	less than 3	14
Incident	Sudden death of close relative	17
	self accident	16
	depression	15
Problem during pregnancy	yes	14
	no	13
Abortion	yes	13
	no	12
Have infection in the genital area	yes	12
	no	11.5
Affected by Breast cancer	yes	11.5
	no	10.5
Estrogen pill taken after menopause	no	10
	yes	10.5
Ever had a hysterectomy	no	10
	yes	9.5
Condom or Diaphragm	no	9.5
	yes	9
Any Birth control pill	no	8.5
	yes	9
Use Napkin	no	8.5
	yes	8
Pregnancy after 35	yes	8
	no	7.5
Pap test	no	7.5
	yes	7
Children after 35	yes	7
	no	6.5
Takes hormone after menopause	no	6
	yes	6.5
Take adequate fruit	no	6
	yes	5.5
Age	above 60	5.5
	46-60	5.5
	30-45	5.25
	below 30	5
Oral Contraception	yes	5
	no	4.5
Education	Undergraduate	4
	Primary	4.25
	Secondary	4.25
	Illiterate	4.5
Age of husband	above 60	4
	40-60	3.65
	below 30	3.8
Cancer vaccine taken	below 30	3.5
	no	3.5
	yes	3
First sex age	above 16	2.5
	16 or below 16	3
Marital Status	married	2
	widow	2.5
	divorce	2.5
	separate	2.25
	Unmarried	2
Number of children	1 to 2	1.5
	above 5	1.75
	3 to 5	2
Number of sex partner	1 to 2	1
	3+	1.5

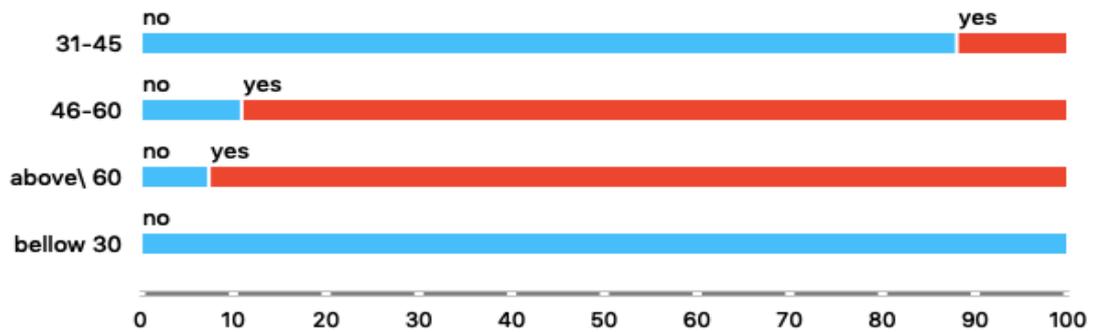


Figure 1. Box plot Age

4.2. Calculation of Risk Level

All of the datasets were scored with a specific value to their factors. Scoring was done with two steps. Firstly, the a weight was given based on the significant level on individual factors than they were scored together based on their importance. For stress the following risk level was obtained,

Low Risk = 60 to 61

Medium Risk = 62 to 63

High Risk = 63 to 65

Very High Risk = greater than or equal to 66

For cervical, I have got,

Low Risk = 115.5 to 117.5

Medium Risk = 118.5 to 120.5

High Risk = 121.5 to 123.5

Very High Risk = greater than or equal to 124.5

For ovarian, I got,

Low Risk = 248 to 251

Medium Risk = 252 to 257

High Risk = 258 to 261

Very High Risk = greater than or equal to 262

From these finding the application was made with the help of my deriving algorithm.

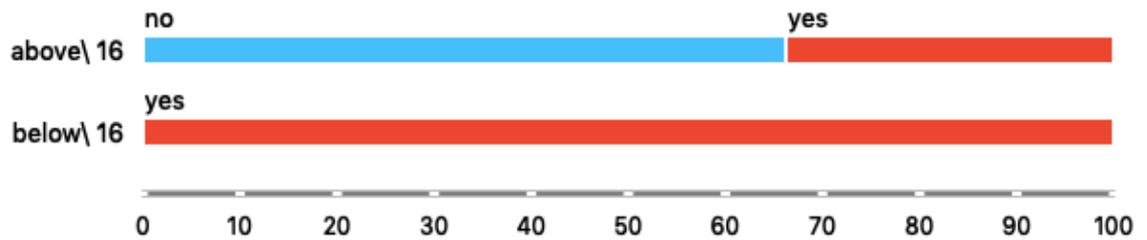


Figure 2. Box plot Age first intercourse

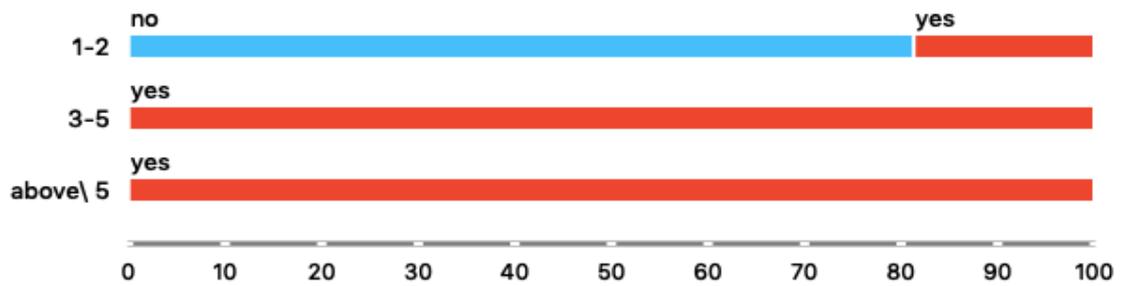


Figure 3 Number of children

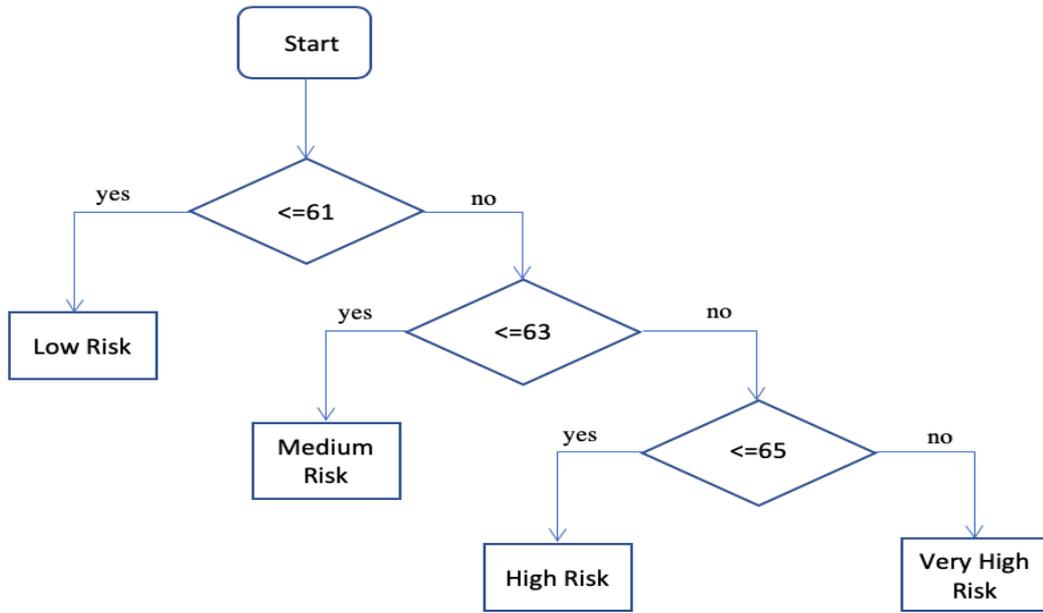


Fig 4: Flowchart of Stress

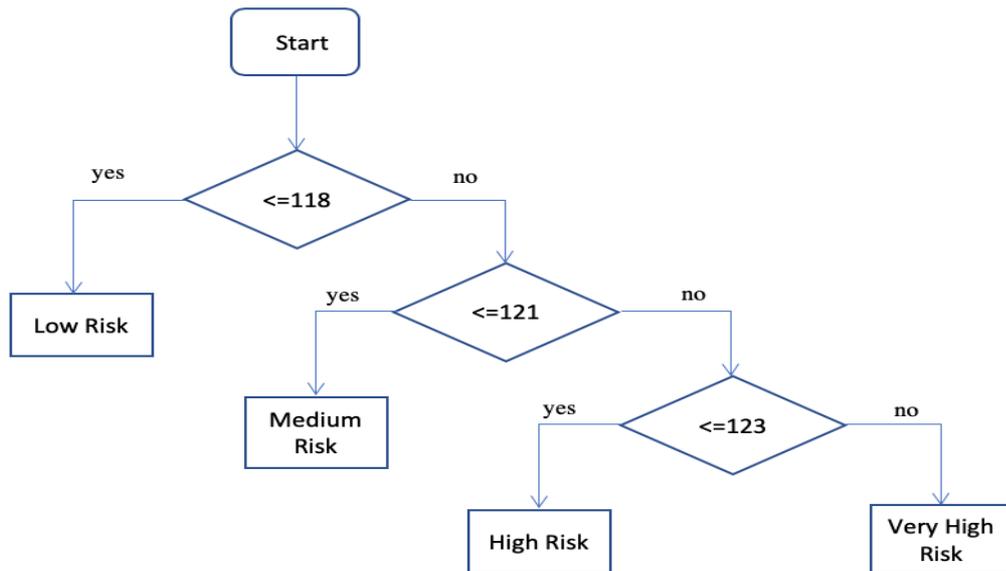


Fig 5: Flow chart of cervical cancer

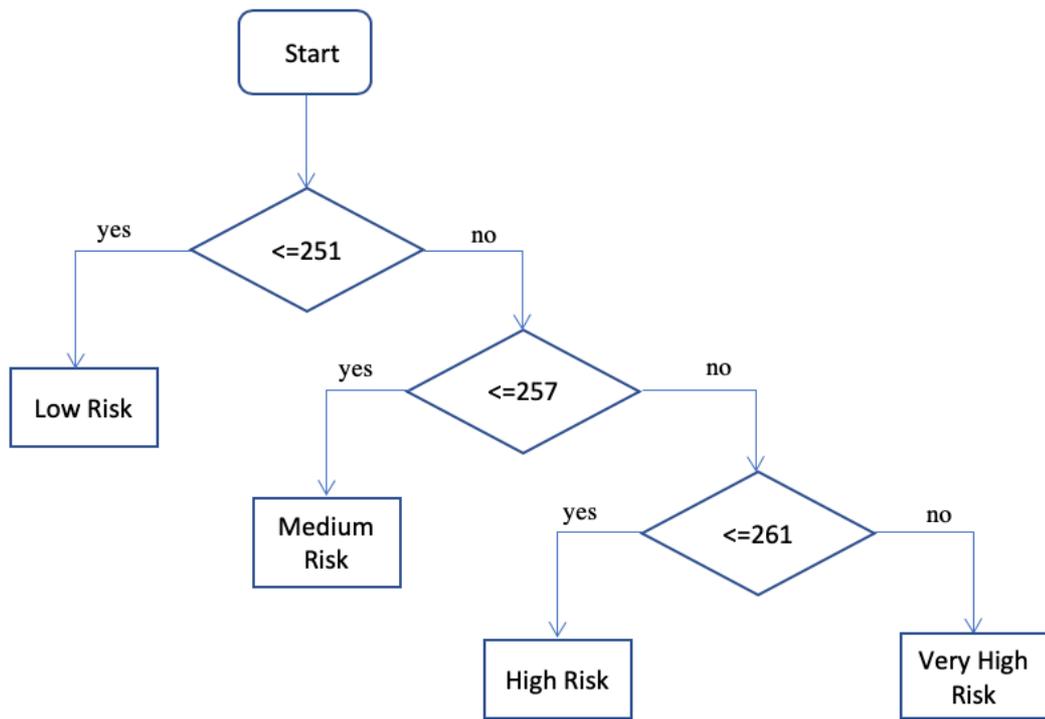


Fig 6: Flowchart of Ovarian Cancer

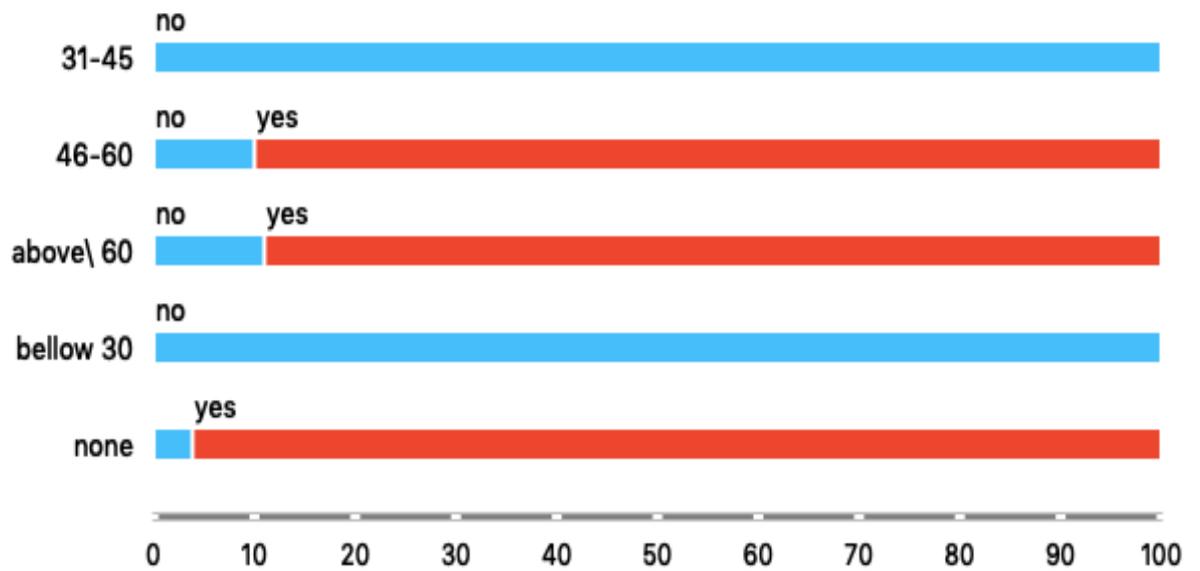


Figure 7. Box plot of Age of husband

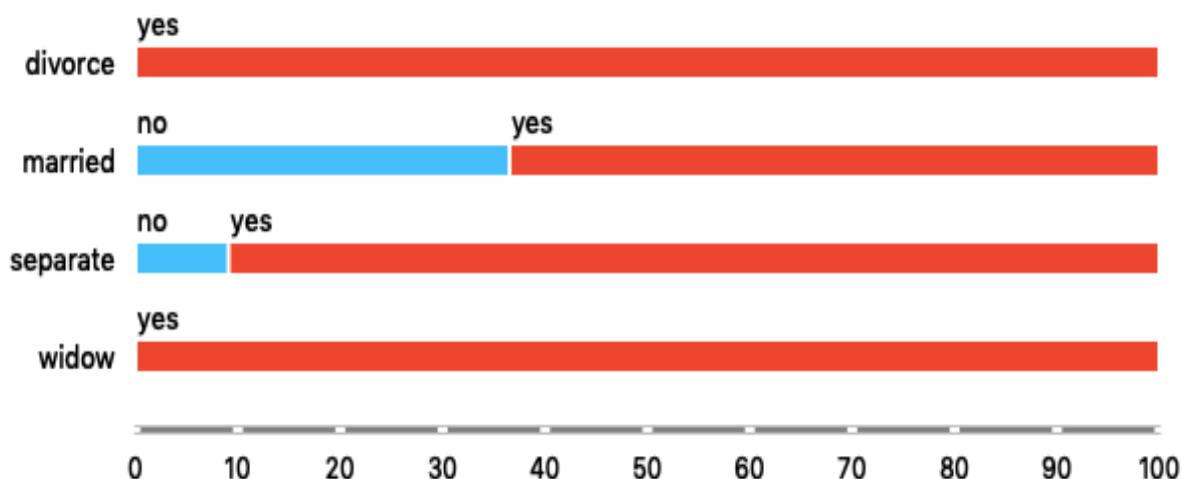


Figure 8. Box plot of Marital Status

Box plot with subgroups of happening cervical cancer were provided in figure 1-5. Women above 60 years or between 46-60 years have highest possibility of getting hit by cervical cancer. Similarly those who had their first intercourse before 16 and those who had taken above 3 children's bears maximum probability. Those women having husband with age above 60 had a probability of happen cervical cancer. Widowed women or divorced women has greater risk of occurring cervical malignancy.

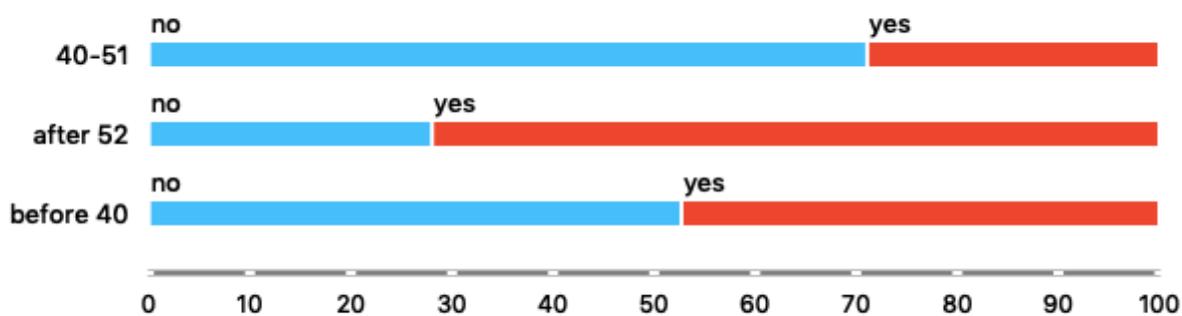


Figure 9. Menopause box plot

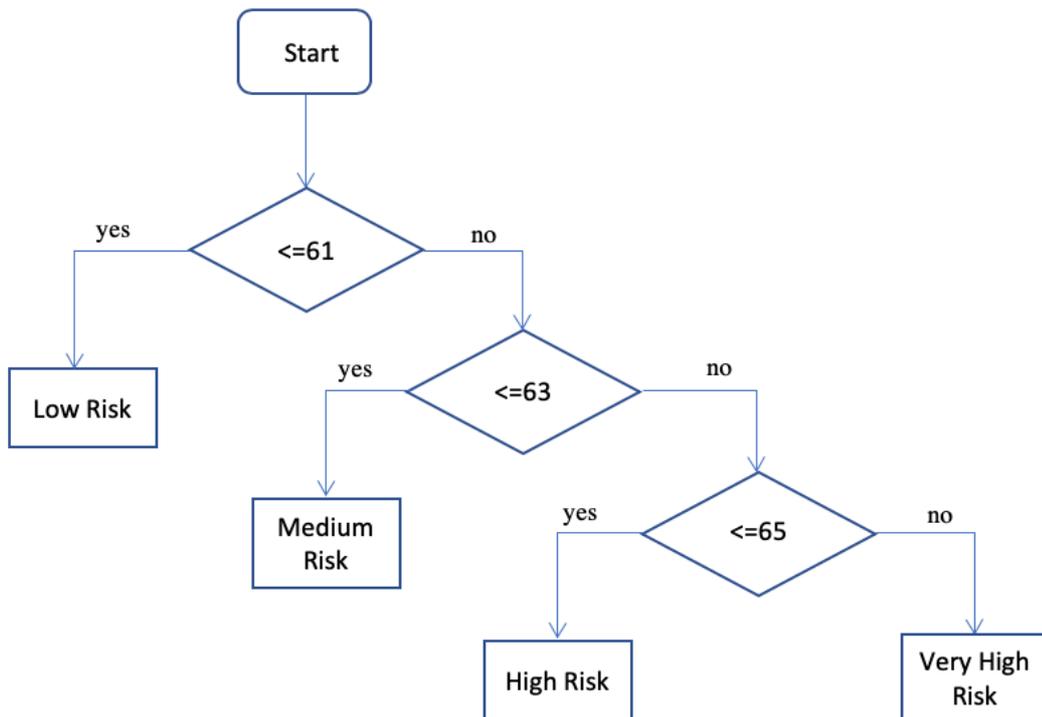


Fig 10: Flowchart of combining cervical, ovarian and stress

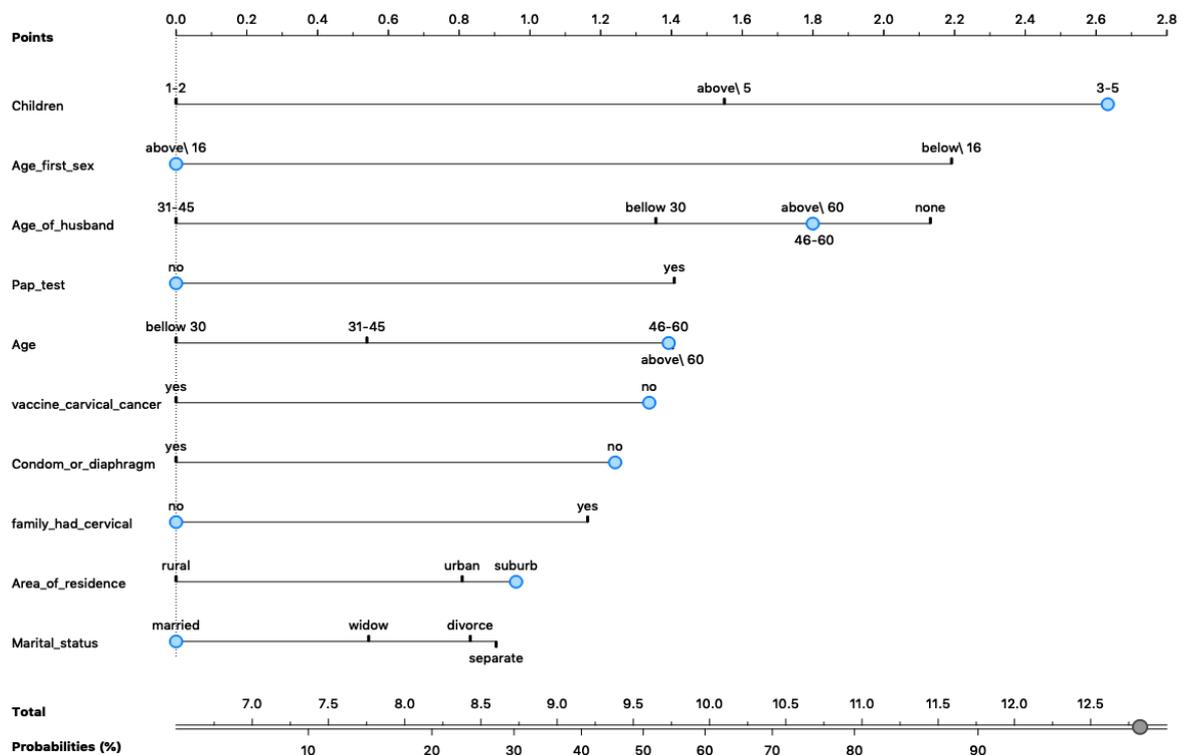


Figure 11. Probabilities of 10 valuable attributes

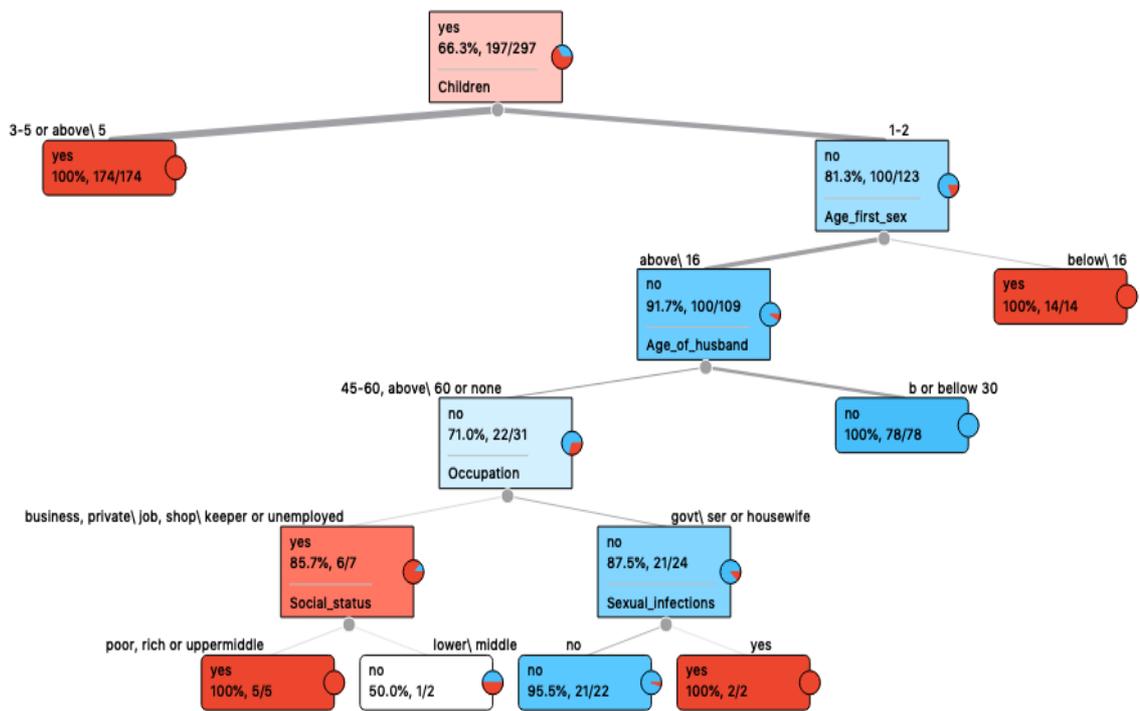


Figure 12. Decision tree between the factors of cervical cancer

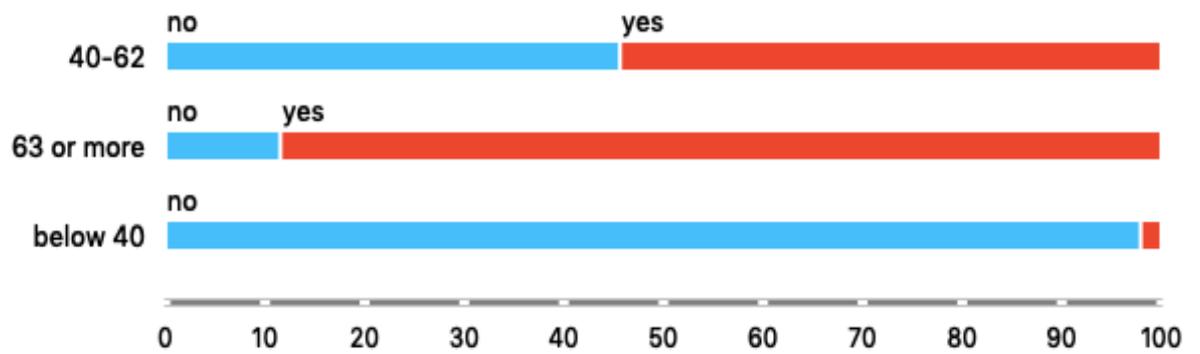


Figure 13. Age box plot

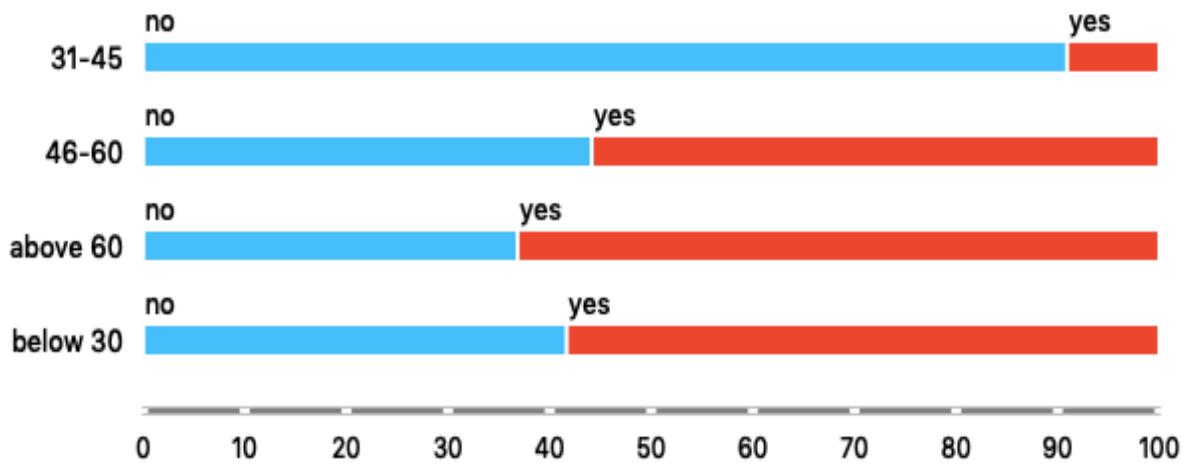


Figure 14. Age of husband box plot

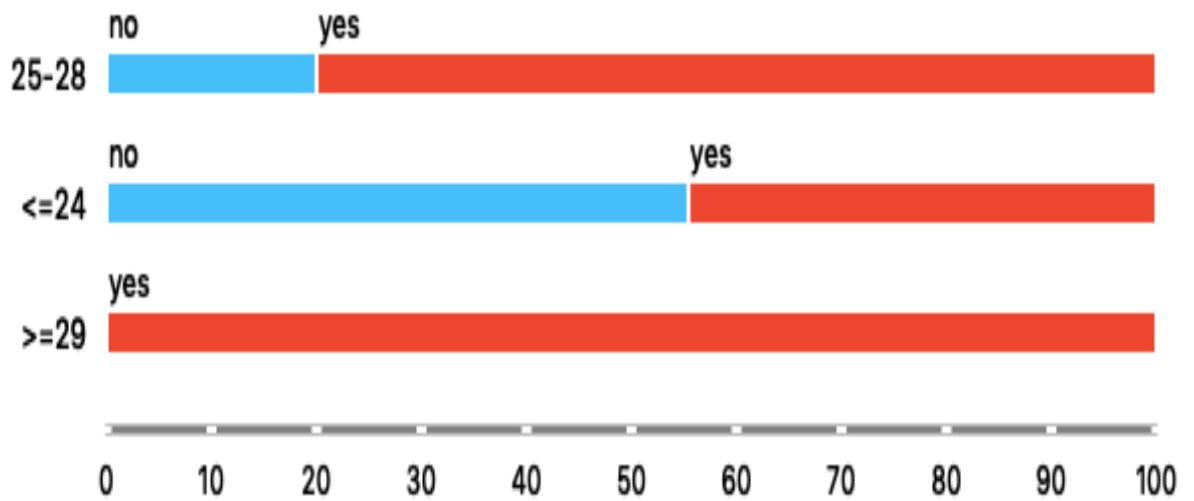


Figure 15. BMI box plot

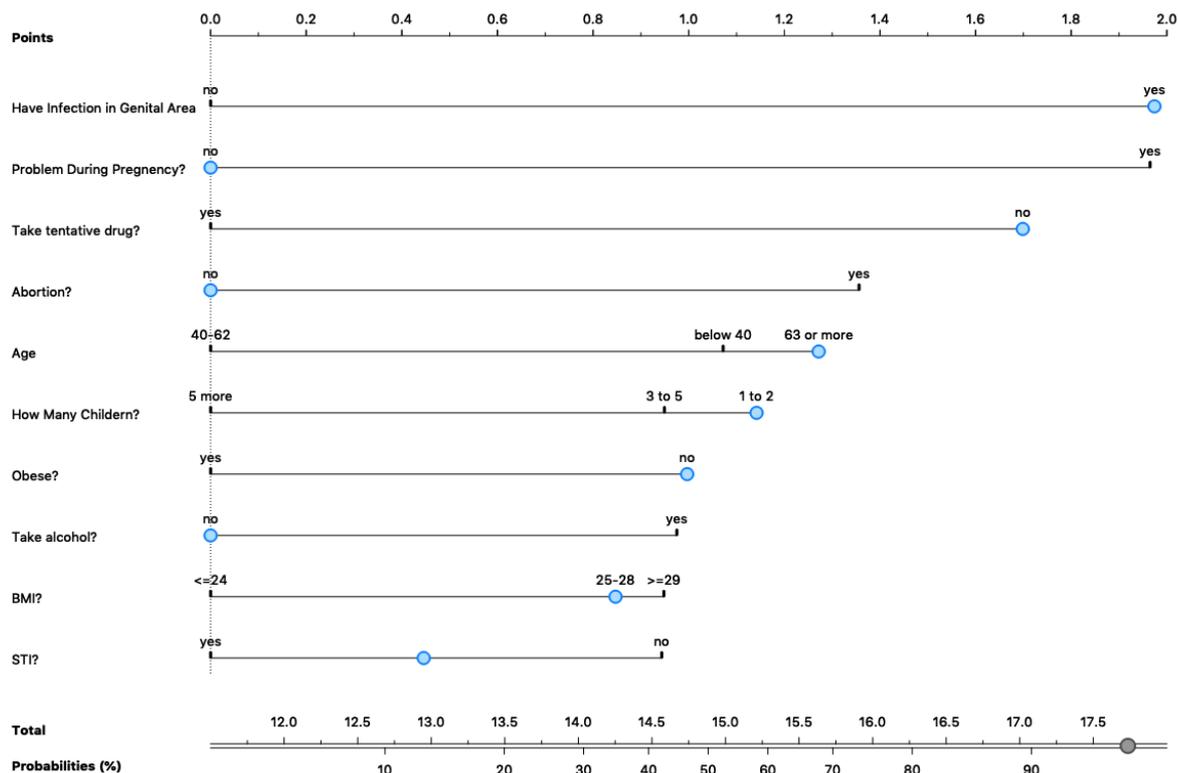


Figure 16. Probabilities 10 highly influenced factors

Figure 6, Figure 9-11 shows the box plots of the 3 manipulating factors e.g. age, age of husband, BMI, Menopause of ovarian cancer. These are drawn with the subfactors of ovarian cancer infected or not. If a female is older than 62 years and menopause ends late after 52 years than female has maximal possibility of happening ovarian cancer.

Figure 7 and 12 shows probability of cancer taking best 10 attributes of cervical and ovarian cancer individually. Probability can be measured with percentage from left to right. Number of children, age of first intercourse, age of husband, pap test, age are the first five attributes those had a great impact behind hit by cervical cancer. On the other hand, genital area infection, pregnancy problem, use of drugs, abortion, number of children are five factors whose values had contact behind the risk of ovarian cancer. The values can help us a way out for prediction at risk of cervical and ovarian cancer.

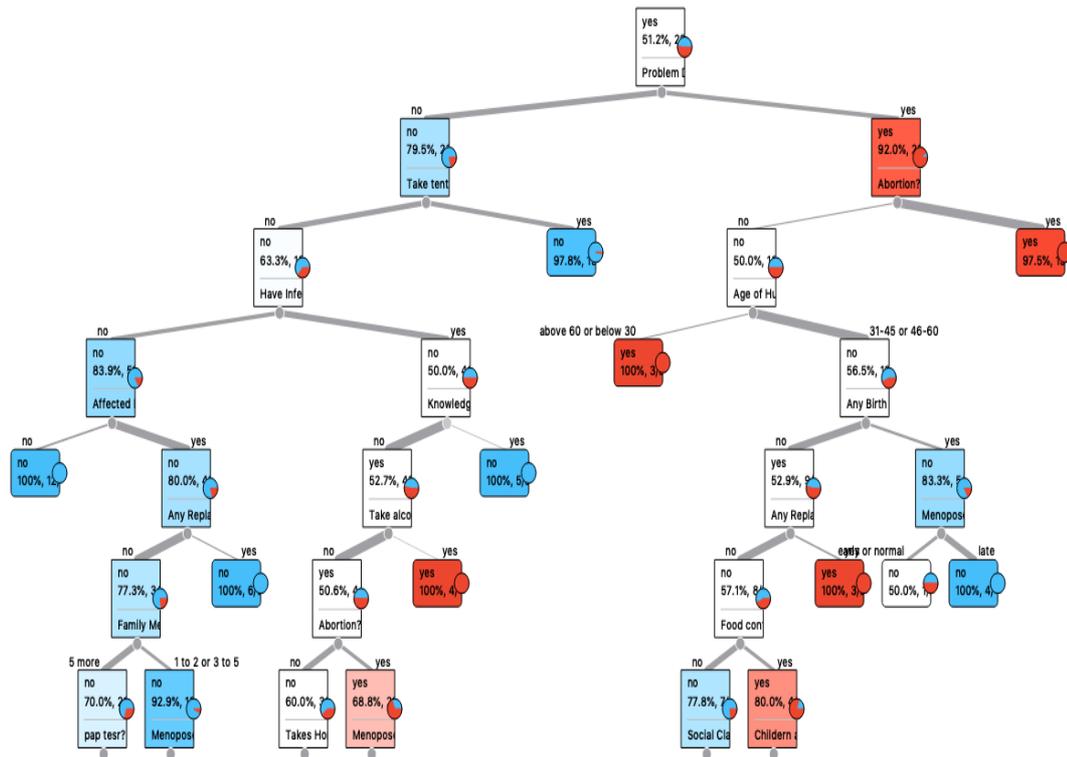


Figure 17. Decision tree between the factors of ovarian cancer

Two decision trees of cervical and ovarian cancer were pointed in figure 7 and 13, there are 100% chances of infected by cervical virus if a women had more than 2 children's. If she had 1-2 children's and her first intercourse was made when she was less than 16 years old than there are also 100% possibility of cervical cancer. Here a decision is made taking 6 precious factors for emerging cervical cancer. Likewise, taking 15 parameters a decision to made to find out the risk of appearing ovarian cancer. The most risky factors are abortion, age of husband, alcohol consumption etc.

4.3. Research Output

The following application was prepared from the findings. With the help of this application it is possible to arise awareness among people of Bangladesh.

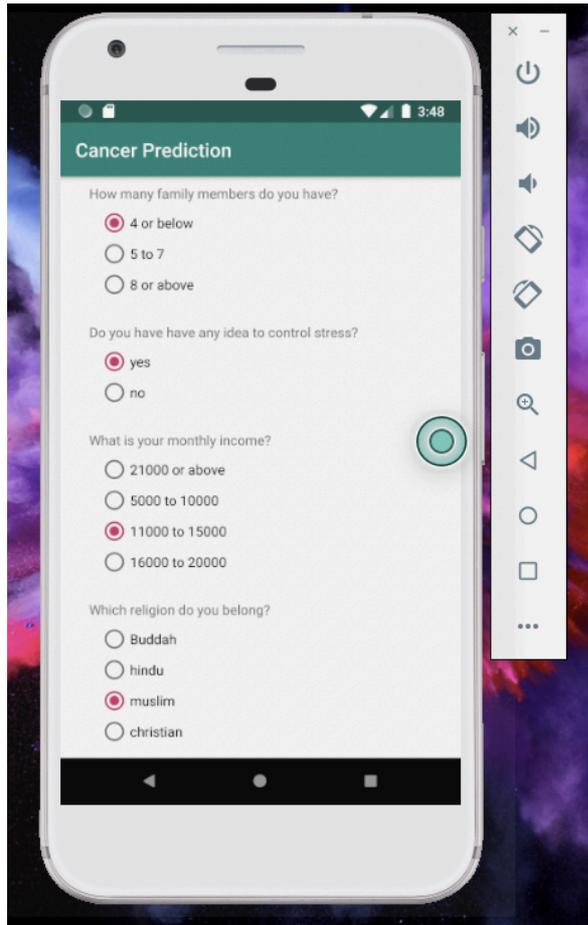


Fig 18: Android application for Cancer prediction

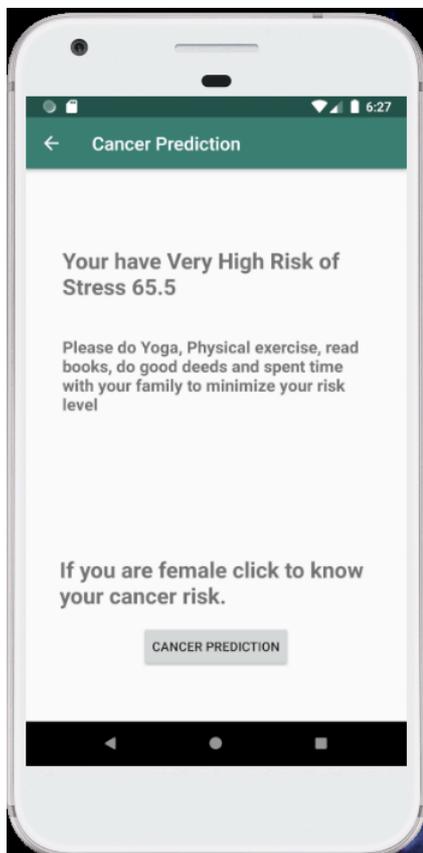


Fig 19: Stress prediction for both male and female

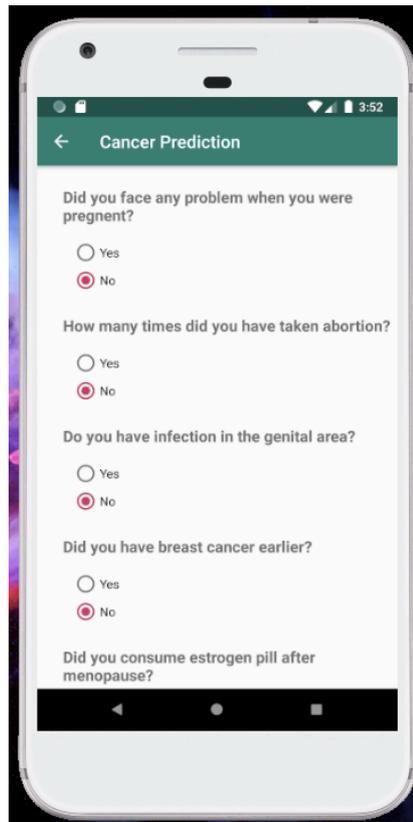


Fig 20: Data collection for Cervical and ovarian cancer prediction

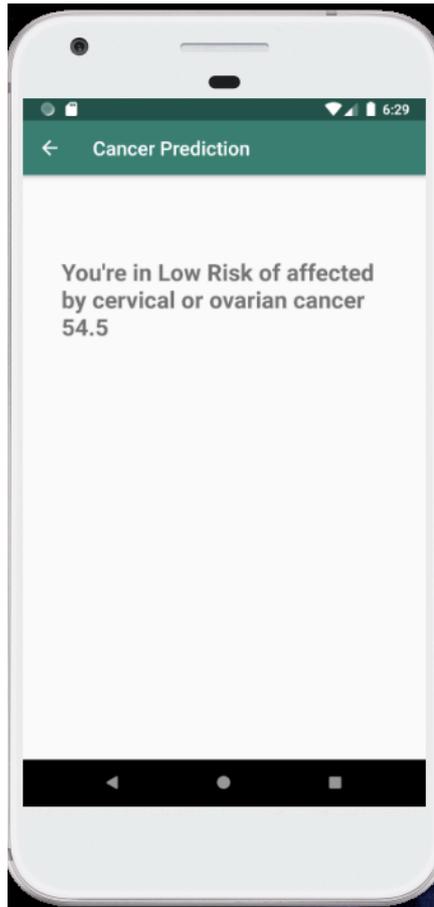


Fig 21: Risk prediction for cervical and ovarian cancer

From the above figures, it is clear that now a days it becomes possible for people to predict deadly diseases earlier and then prevent these disease at an early stage.

5. CONCLUSION

Cervical and ovarian cancers are the dominant causes of women's death in Bangladesh. The majority of the people are unconscious of it. Death is inescapable because of cervical and ovarian cancer. From the findings we got the evidence that immune response may be damaged due stress may even enhance the development of cancer. In this scrutiny, risk factors of cervical and ovarian cancer were analyzed carefully. Stress variable were also founded those have impact on cervical and ovarian cancer. A skillful methodology has been accommodated for the extraction of dominant forms from archives. The noticeable factors related with cervical cancer and ovarian

cancer and their association with stress had been scrutinized. Finally, a connection of stress with cervical and ovarian cancer is pointed out. Afterwards the prediction is made on cancer growth with an application of android.

6. FUTURE WORK

At Present, we have used 298 females data for analyzing cervical cancer, 522 data for ovarian cancer and 161 patient's data for examining stress. We have get an efficient result but using more data will help to get more accurate result. The application will be used to store data of individuals in archive so that it can be used for better prediction.

7. REFERENCES

[1]

H. Zhang, Y. Li, M. Li, and X. Chen, "A randomized controlled trial of mindfulness-based stress reduction for insomnia secondary to cervical cancer: Effects on sleep," *Cancer Reports*, p. e1190, 2019.

[2]

L. Calderón-Garcidueñas, E. Leray, P. Heydarpour, R. Torres-Jardón, and J. Reis, "Air pollution, a rising environmental risk factor for cognition, neuroinflammation and neurodegeneration: the clinical impact on children and beyond," *Revue neurologique*, vol. 172, no. 1, pp. 69–80, 2016.

[3]

C. M. Karch and A. M. Goate, "Alzheimer's disease risk genes and mechanisms of disease pathogenesis," *Biological psychiatry*, vol. 77, no. 1, pp. 43–51, 2015.

[4]

S. Asaduzzaman *et al.*, "Anticipation of the significance of risk factors in cervical cancer for low incoming country: Bangladesh perspective," 2015.

[5]

K. Ahmed, S. Asaduzzaman, M. I. Bashar, G. Hossain, and T. Bhuiyan, "Association assessment among risk factors and breast cancer in a low income country: Bangladesh," *Asian Pac J Cancer Prev*, vol. 16, no. 17, pp. 7507–12, 2015.

[6]

S. Asaduzzaman, F. Al Masud, T. Bhuiyan, K. Ahmed, B. K. Paul, and S. M. Rahman, "Dataset on significant risk factors for Type 1 Diabetes: A Bangladeshi perspective," *Data in brief*, vol. 21, pp. 700–708, 2018.

[7]

J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, "Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk," *Nature genetics*, vol. 50, no. 8, p. 1171, 2018.

[8]

K. Ahmed *et al.*, "Early detection of lung cancer risk using data mining," 2013.

[9]

- M. S. Islam *et al.*, “Early Prevention and Detection of Cancer Risk for Low Income Country using Data Mining Technology: Bangladesh Perspective,” 2016.
- [10]
K. Ahmed, T. Jesmin, and M. Z. Rahman, “Early prevention and detection of skin cancer risk using data mining,” *International Journal of Computer Applications*, vol. 62, no. 4, 2013.
- [11]
M. Chin-Chan, J. Navarro-Yepes, and B. Quintanilla-Vega, “Environmental pollutants as risk factors for neurodegenerative disorders: Alzheimer and Parkinson diseases,” *Frontiers in cellular neuroscience*, vol. 9, p. 124, 2015.
- [12]
R. C. Gardner and K. Yaffe, “Epidemiology of mild traumatic brain injury and neurodegenerative disease,” *Molecular and Cellular Neuroscience*, vol. 66, pp. 75–80, 2015.
- [13]
S. ASADUZZAMAN *et al.*, “Hazardous consequences of polygamy, contraceptives and number of childs on cervical cancer in a low incoming country: Bangladesh,” *Cumhuriyet Üniversitesi Fen-Edebiyat Fakültesi Fen Bilimleri Dergisi*, vol. 37, no. 1, pp. 74–84, 2016.
- [14]
S. Kulkarni, C. D. Bhat, D. Patil, and J. Dara, “Heart disease classification: A case study using machine learning and data mining,” *International journal of computer trends and technology*, vol. 2, no. 4, pp. 36–43, 2018.
- [15]
T. Jesmin, “Investigation of common disease regulatory network for metabolic disorders: A bioinformatics approach,” *Network Biology*, vol. 6, no. 1, p. 28, 2016.
- [16]
B. Halliwell, “Making Sense of Neurodegeneration: A Unifying Hypothesis,” in *Learning To Live Together: Promoting Social Harmony*, Springer, 2019, pp. 115–120.
- [17]
B. T. Hennessy, R. L. Coleman, and M. Markman, “Ovarian cancer,” *The lancet*, vol. 374, no. 9698, pp. 1371–1382, 2009.
- [18]
J. Hunn and G. C. Rodriguez, “Ovarian cancer: etiology, risk factors, and epidemiology,” *Clinical obstetrics and gynecology*, vol. 55, no. 1, pp. 3–23, 2012.
- [19]
J. I. Bisson, “Post-traumatic stress disorder,” *Bmj*, vol. 334, no. 7597, pp. 789–793, 2007.
- [20]
A. L. Roberts, T. Huang, K. C. Koenen, Y. Kim, L. D. Kubzansky, and S. S. Tworoger, “Posttraumatic Stress Disorder Is Associated with Increased Risk of Ovarian Cancer: A Prospective and Retrospective Longitudinal Cohort Study,” *Cancer research*, vol. 79, no. 19, pp. 5113–5120, 2019.
- [21]
V. Chaurasia, S. Pal, and B. B. Tiwari, “Prediction of benign and malignant breast cancer using data mining techniques,” *Journal of Algorithms & Computational Technology*, vol. 12, no. 2, pp. 119–126, 2018.
- [22]

- U. Schmidt-Erfurth *et al.*, “Prediction of individual disease conversion in early AMD using artificial intelligence,” *Investigative ophthalmology & visual science*, vol. 59, no. 8, pp. 3199–3208, 2018.
- [23]
- R. Hossain, S. H. Mahmud, M. A. Hossain, S. R. H. Noori, and H. Jahan, “PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques,” *Procedia computer science*, vol. 132, pp. 1068–1076, 2018.
- [24]
- D. Lu *et al.*, “Psychologic Distress Is Associated with Cancer-Specific Mortality among Patients with Cervical Cancer,” *Cancer research*, vol. 79, no. 15, pp. 3965–3972, 2019.
- [25]
- R. B. Postuma *et al.*, “Risk factors for neurodegeneration in idiopathic rapid eye movement sleep behavior disorder: a multicenter study,” *Annals of neurology*, vol. 77, no. 5, pp. 830–839, 2015.
- [26]
- F. Ma, J. Gao, Q. Suo, Q. You, J. Zhou, and A. Zhang, “Risk prediction on electronic health records with prior medical knowledge,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1910–1919.
- [27]
- E. M. V. Reiche, S. O. V. Nunes, and H. K. Morimoto, “Stress, depression, the immune system, and cancer,” *The lancet oncology*, vol. 5, no. 10, pp. 617–625, 2004.
- [28]
- A. Ascherio and M. A. Schwarzschild, “The epidemiology of Parkinson’s disease: risk factors and prevention,” *The Lancet Neurology*, vol. 15, no. 12, pp. 1257–1272, 2016.
- [29]
- R. Gupta and N. Sen, “Traumatic brain injury: a risk factor for neurodegenerative diseases,” *Reviews in the Neurosciences*, vol. 27, no. 1, pp. 93–100, 2016.
- [30]
- J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.