Identifying risk factors and their impact on Cervical Cancer

A dissertation submitted for the Degree of Master of Information Technology

A.M.R.D. Adikari

University of Colombo School of Computing

2018

DECLARATION

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: A.M.R.D. Adikari

Registration Number: 2015/MIT/001

Index Number: 15550011

_____

Signature: Date:

This is to certify that this thesis is based on the work of

~~Mr.~~/Ms. A.M.R.D. Adikari

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. H.A. Caldera

_____

Signature: Date:

ABSTRACT

Cervical cancer is one of the most common cancers affecting women worldwide. According to medical view, there are several factors which cause this cancer. It is mainly caused due to inappropriate behavior of women. Cervical cancer has become a leading cancer in Sri Lanka by being at the second place out of top ten cancers. Hence, Sri Lankan government has started conducting several awareness programs to minimize the number of effects.

Computational algorithms are the new trend in determining most of the diseases and they are heavily used with cancer data. Since these algorithms are applied to enormous number of actual data, researchers can come up with results that are more accurate. In the meantime, patients can gain a better understanding about themselves prior to a medical checkup.

This research is conducted in order to identify significance of cervical cancer risk factors and their combinational impact. Data mining techniques are applied to identify combinational effect of risk factors. With the application of these techniques, an idea will be provided for medical researchers to conduct their research on this.

Knowledge generated with the application of data mining techniques will be helpful in the medical field since the output of data mining is based on actual data. Doctors can conduct an analysis from their end and enhance the generated knowledge. Therefore, this information will be very useful to doctors in diagnosing their patients and predicting the probability of their patients developing cervical cancer.

ACKNOWLEDGEMENT

I would like to express my gratitude to Dr. Amitha Caldera, course coordinator of Master of Information Technology, University of Colombo School of Computing and Dr. Mindika Premachandra, project coordinator for giving a greater support throughout the degree program.

I would like to pay my special gratitude again to Dr. Amitha Caldera who supervised me in many ways at many times in my difficulties by contributing his valuable time and for guiding me in delivering the project successfully.

I am also thankful to Dr. Aruni Gallage, lecturer at Medical Faculty of Colombo for her support in gaining the domain knowledge. Finally, I thank all those whose names, though not mentioned for their support and encouragement in completing this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

HPV - Human Papilloma Virus

HIV - Human Immunodeficiency Virus

AIDS - Acquired immune deficiency syndrome

CIN - Cervical Intraepithelial Neoplasia

IUD - Intra Uterine Device

HBV - Hepatitis B Virus

PID - Pelvic Inflammatory Disease

RFT - Random Forest Tree

1. INTRODUCTION

1.1.AREA OF STUDY

A disease is an inevitable cause for all living elements within a population. No living being has a 100% perfect immunization system to resist a disease. Similar to the fragility of a product, all living beings experience some kind of a disease at some point of their complex life cycle. A disease is a particular abnormal condition that affects a part or all of an organism not caused by external force and that consists of a disorder of a structure or function, usually serving as an evolutionary disadvantage [8]. A disease is a common experience to all living beings. In humans, a disease is often used more broadly and more specifically to emphasize any medical condition that cause pain, dysfunction, distress, social problems, injuries, disabilities, disorders, syndromes, infections, deviant behaviors or death to the person affected. A disease can be a specific symptom/s or sign/s associated with a certain medical condition. People can get affected physically and psychologically due to a disease. One out of the numerous ways to categorize diseases is as Transmissible/Communicable and Non-Transmissible/Non-Communicable, depending on their genetic transferability.

It is alarming to witness that non-transmissible diseases like stroke, cancer, heart diseases, kidney failures, etc. cause the majority of the human deaths nowadays. Most non-transmissible diseases are fatal due to their non-curable nature. Scientists and Medical Researchers are still struggling to find proper solutions to these diseases that will ultimately benefit millions of people around the world.

1.2.MOTIVATION

Amongst the various cancer types, cervical cancer is a significant challenge to the healthcare system of females. A virus called human papillomavirus (HPV) causes cervical cancer. This virus is transmitted sexually from one person to another. Due to inappropriate personal conduct, cervical cancer has become a burden to the world. It has become the fourth most common cancer among women in the world [44]. Cervical cancer is the most common cancer in women in Eastern and Middle Africa.

However, most of the women in the world are unaware of the risk factors of cervical cancer. According to statistics, most of the cervical cancer incidents are found in less

developed or developing countries [4]. Since this has become a leading type of cancer, researchers are interested in conducting studies to identify the significant impact of its risk factors and thereby increase the awareness about it.

In this study, profiles of cervical cancer patients will be analyzed in order to identify relationships between cervical cancer and risk factors. Cervical cancer occurs when abnormal cells grow out of control on the cervix [1]. The cervix is the lower part of the uterus which is entrance to the womb from the vagina. Mostly a virus called human papillomavirus (HPV) causes cervical cancer [1]. There are two main types of cervical cancers. They are squamous cell carcinoma and adenocarcinoma. The shape of cancer cells distinguishes each type of cancer. In squamous cell carcinoma, the cancer cells develop from exocervix (the part next to the vagina) area and those are of squamous shape where as in adenocarcinoma, the cancer cells develop form gland cells in endocervix (the part of the cervix closest to the body of the uterus) area [2,3]. Figure 1.1 shows the above mentioned parts of a female reproductive system graphically.

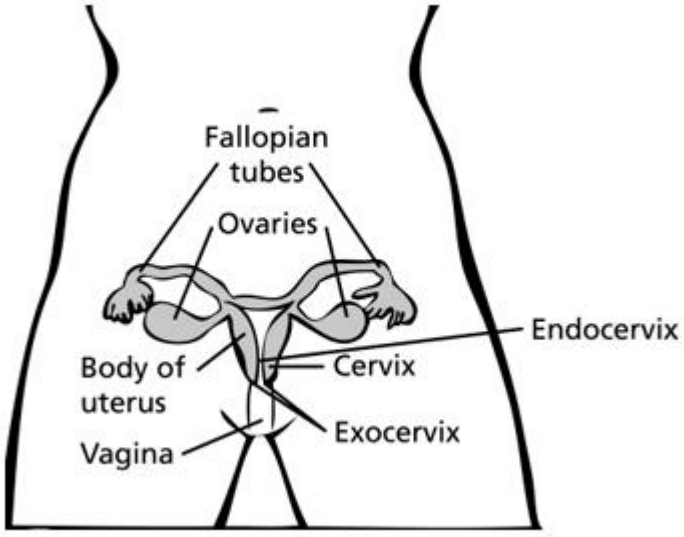

Figure 1.1 : Female Reproductive System

Screening, Pap test and biopsy are medical tests conducted to detect a cervical cancer. However, these medical tests are expensive and inconvenient to patients [49]. Hence, there is a considerable interest in computational techniques to detect cervical cancer. This research is conducted to identify patterns in risk factors of cervical cancer via data mining

techniques and help people in determining their probability of developing cervical cancer before going for a medical test.

1.3.STATEMENT OF THE PROBLEM

This research is conducted to identify associations between cervical cancer and risk factors contributing to its development. HPV infection, HIV infection, herpes, syphilis, condylomatosis, contraceptive usage, smoking and age are some risk factors of cervical cancer that will be analyzed in the study. Since the dataset is taken from Venezuela it can be considered as a representation of a developing country. Most of the time, people have to go for medical tests to identify the probability of developing cervical cancer. However, with the help of computational models people can get an idea about the possibility of getting it without going for medical tests. The relationship between risk factors of cervical cancer helps a lot in making this decision.

Some people with a single risk factor will never develop cervical cancer while people having several risk factors have a high chance of developing cervical cancer. Hence, the association between risk factors need to be identified accurately, in order to decide whether there is a possibility of developing a cervical cancer.

Human analysis without using tools may not make sense when working with multiple attributes and large volumes of data. Therefore, automated computer systems are required to perform intelligent data analysis and find new patterns in data. Both data mining models and statistical models can be used in analyzing data. However, statistical models may not help in analyzing multi factor dependencies of several risk factors. Therefore, data mining models needs to be used to identify relationships between several risk factors.

1.4.AIMS AND OBJECTIVES

Important and significant impact of risk factors may exist as a combination of them rather than individuals and therefore the main aim is to discover such combinational effects. Unidentified associations between cervical cancer and its risk factors can be identified via this. Gathered data will be analyzed to come up with new co-relations between risk factors of cervical cancer with a high significance. The objectives are listed below.

- Identify possible associations between cervical cancer and risk factors.
- Identify associations with highest significance.

- Discover significant impact of risk factors which may exist as combinations rather than as individuals.

1.5.PROJECT SCOPE

The sample consists of women of age 15-52 years and who registered as cervical cancer patients under 'Hospital Universitario de Caracas' in Caracas, Venezuela. It is a publicly owned teaching hospital situated in the premises of Universitario de Caracas. The collected data set has around thirty attributes that are fairly related with cervical cancer. From of the collected thirty attributes, below mentioned attributes of patients will be analyzed under the research.

- Age
- Number of sexual partners
- First sexual intercourse (age)
- No. of pregnancies
- Smoke
- Hormonal Contraceptives
- IUD
- Condylomatosis
- Syphilis
- Genital herpes
- AIDS
- HIV
- Hepatitis B
- HPV
- CIN / Cancer
- PID

In this study, only the patients with complete records of data will be taken into consideration. Patients with missing records of data will be removed from the study as out of the scope records. Since it is not a good practice to fill missing data based on similar records in medical data sets, missing records will not be reflected in analysis.

1.6.STRUCTURE OF THE DISSERTATION

Chapter 1 provides a clear idea about the dissertation and motivates the reader to read it. It defines the problem clearly and briefly to the reader and consists of the motivation to conduct this research project. Objectives and scope of the project also are described in this chapter. Finally, this chapter provides an overall idea about the entire structure of dissertation and content of it.

In chapter 2, an analysis of existing similar projects is provided. Many countries have conducted similar projects on the same problem and they have identified unseen patterns.

This literature review is very helpful in analyzing the dataset and identify new patterns. At the same time, these may be helpful in proving the newly identified patterns with the dataset. When analyzing the literature, risk factors considered in each analysis are studied properly. Except for the main risk factors, others vary slightly from one study to another. Therefore, a considerable number of patterns can be found and different types of data mining algorithms can be learned. At the same time, some algorithms may not be suitable with a certain set of risk factors and those can be identified. This chapter helps in providing a thorough understanding about the problem domain.

Chapter 3 delivers a detailed description about the methodology followed in identifying patterns. This consists of several methodical approaches that could be used and the approach chosen amongst them, with a proper justification. The chapter provides a detailed description of the selected approach of data mining, highlighting the benefits of it. In addition, this includes the initial study conducted with the dataset in identifying unseen patterns. Steps followed in implementing the methodology also are described in this chapter. The problems faced when applying data mining algorithms with technical tools are also explained.

In chapter 4, a detailed analysis of the outcomes of the dataset is described. The results obtained are critically analyzed with the help of diagrams. Required graphs and charts will be provided in verifying the outcomes. At the same time, the chapter explains how the dataset was controlled by applying various methodologies in obtaining results.

Chapter 5 contains the conclusion and future work. This chapter summarizes the work conducted, discusses the findings and points out limitations of the current work. In addition, this chapter discusses the lessons learnt during the study and areas for future development.

2. LITERATURE REVIEW

2.1.LITERATURE ON CANCER

Cancer is at the top most cause of deaths worldwide. Early detection and prevention of cancer is very important in decreasing the number of deaths due to cancer. Most of the findings and literature about cancer have discovered many risk factors associated with the potential patient. A risk factor is any negative aspect/cause that changes and challenges a chance of getting a disease. Different cancers have different risk factors due its diversified and complex nature. Even though this is a preventable disease, this has become a heavier burden to the world. Most of the countries conduct researches on cancer types to identify their root causes in order reduce the number of deaths caused by cancer.

Cancer is caused by uncontrollable growth of cells in part of the body. It may occur at one organ in the body and later spread to other organs. Various studies done in developed countries show that education level is an important risk factor of cancer. In developed countries, there are many breast cancer occurrences while quite a few cancer occurrences in stomach, lungs and uterine cervix. A combination of clustering and decision tree techniques was used in a research to build a cancer risk prediction model, to predict cancers in lung, stomach, blood, breast, oral and cervix. Age, education, living area, habits, occupational hazards, anemia, weight loss and family history of cancer are the attributes considered in this study. This model provides the cancer status of a person by matching his/her data with entire database, the risk score with the patterns mined by decision tree algorithm and type of cancer as a cluster output [51].

Breast cancer is the second leading cause of death and the most common cancer in developed countries. At the same time, this is considered as the leading cause of cancer deaths among women of 40 - 59 years of age. Breast cancer is caused due to a malignant tumor in the breast. Risk factors of breast cancer are stage at diagnosis, age, genetic factors and family history. In this study, eleven attributes were considered and Supervised Learning Algorithms were used to identify the most significant attributes among them. Out of the four types of classified methods used, Decision Tree Algorithm was identified as the most accurate method. With the Decision Tree algorithm, Uniformity of Cell Size, Uniformity of Cell Shape, Clump Thickness and Bare Nuclei were classified as the best set of attributes [52].

2.2.LITERATURE ON CERVICAL CANCER

Cervical cancer is considered as the leading cause of death among women in the low- and middle-income countries. It is estimated that 500,000 women get cervical cancer as new cases worldwide and majority of them are belonging to developing countries [49]. It is highly pointed-out that several major risk factors increase the likelihood of developing cervical cancer, which is an alarming status. According to the extensive surveys and advance researches across the globe it is found that, women without any of these risk factors rarely develop cervical cancer. Although these risk factors increase the tendency and likelihood of developing cervical cancer, many women with these risks do not develop this disease due to genetic or other personal immunization strength. This complex situation leads to unsolved solution to cervical cancers due to the personal diversity and different in-built genetic patterns.

In recent past, it has shown a huge progress in the researches and projects in understanding what happens in cells of the cervix when cancer develops. Several risk factors have been identified as a result. Those increase the likelihood that a woman might develop cervical cancer. It is ascertained that HPV is not the only cause of cervical cancer. Most women with HPV do not get cervical cancer. Certain other risk factors, like smoking and HIV infection, influence to develop cervical cancer for the women who are already exposed to HPV [9].

Cervical cancer is an important cause of mortality among women in developing countries, especially in the Latin America and Caribbean (LAC) region. Infection with human papillomavirus (HPV) has been identified as the primary cause of cervical cancer [11]. Venezuela has 11.34 million of women aged 15 years and older, who are at risk of developing cervical cancer. According to statistics, 4973 women are diagnosed with cervical cancer and 1789 die from it every year. Cervical cancer is the second most frequent cancer among women in Venezuela and the top most frequent cancer among women between 15 and 44 years of age [12].

Based on global statistics, cervical cancer is the fourth most common cancer in women with more than 527,000 new cases being diagnosed every year. The impact of cervical cancer is not shared among the world equally. More than 8 in 10 cervical cancer deaths (85%) are in low to middle-income countries. Women living in Africa, South America,

and parts of Asia are hardest hit by this inequality. People living in those countries have poor access to free healthcare and the nearest affordable hospital can be many miles away with little means of transport. In the meantime, low-income countries are struggling with financial problems in setting up an effective national cervical screening programme [13].

It is identified that the reasons for the high incidence of cervical cancer in sub-Saharan Africa include lack of awareness of cervical cancer among the population. Raising awareness about cervical cancer risk factors including young age at first sexual intercourse, multiple male sexual partners, infections with the human papillomavirus, young age at first full-term pregnancy, prolonged use of oral contraceptives and HIV infections will be helpful in reducing the number of cervical cancer incidents [5]. In Sub-Saharan Africa, cervical cancer covers 22.5% of all cancer cases in women and the majority of those women live in rural areas, which means they are not educated or aware about cervical cancer and its impact. According to statistics, Rwanda has a population of 11 million with 2.72 million women at a risk of developing cervical cancer. At the same time, cervical cancer ranks as the most frequent cancer among women in Rwanda. In sub-Saharan countries, prolonged HPV infections and HIV/AIDS are considered as key risk factors of cervical cancer. In addition to above risk factors, sexual activity before age of 20 years old, multiple sexual partners, tobacco smoking and oral contraceptive pill use for more than 5 years also are considered as risk factors [6].

There are both controllable and uncontrollable risk factors for cervical cancer. Age is considered as an uncontrollable risk factor whereas HPV, AIDS, no. of sexual partners and smoking can be considered as controllable factors. These risk factors increase the chance of getting cervical cancer. According to studies conducted, it has identified that only a single factor may not cause cervical cancer. A combination of several factors may cause this. Therefore, the co-relation between risk factors needs to be identified to minimize or avoid incidents of cervical cancer. Researchers have identified that women who have had three or more full-term pregnancies have an increased risk of developing cervical cancer. In the meantime, Women who were younger than 17 years when they had their first full-term pregnancy are almost 2 times more likely to get cervical cancer later in life [7].

Cervical cancer has become a burning problem in in developing country like Bangladesh, India and Pakistan causing over 88% of women death. Cervical cancer was identified as

the leading cause of cancerous death in Bangladesh by year 2012. A research was conducted to analyze significance of cervical factor risk factors in Bangladesh women with both statistical and data mining approach. According to this study, about 10 factors have been identified as highly significant risk factors. Some of them are premature chronicle of cancer, no. of sexual partners, first sexual intercourse below age 16, use of oral contraceptives, no. of children, previous cancer history and affected by sexually transmitted disease. Risk factor precedence also has been calculated in this study using ranker algorithm with attribute evaluators like OneRAttributeEval, ReliefFAttributeEval and CorrelationAttributeEval [43].

According to statistics, cervical cancer is becoming a huge problem in developing countries like Sri Lanka. Cervical cancer ranks as the second leading cause of female cancer in Sri Lanka. Figure 2.1 shows that 10% of all female cancers are cervical cancer. Not only HPV but also there are some other co-factors, which help in causing cervical cancer. Tobacco smoking, long-term hormonal contraceptive usage and co-infection with HIV are some of those established cofactors. In the meantime, sexual intercourse is been identified as the primary route of transmission of HPV [4].

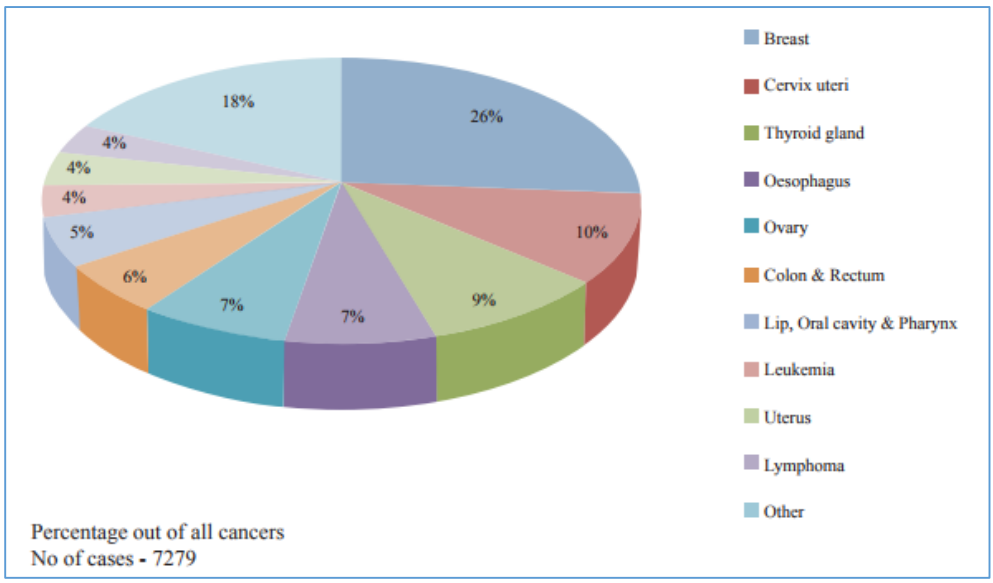

Figure 2.1: Percentage out of all cancers in Sri Lanka

In India, cancer is among the top ten causes of death. Based on the statistics provided by Indian National Cancer Registry, cancer in oral cavity, lungs, esophagus and stomach in men and cervix, breast and oral cavity in women. From these, cancer in cervix and breast causes 50% of female deaths [49]. According to this study, risk factors consists of

inadequate screening, HPV, multiple sexual partners, male sexual behavior, sexual intercourse at young age, tobacco consumption, living habits and homostatic factors. In the meantime, it was identified that coexistence of risk factors increases the risk of getting cervical cancer. Risk factors were categorized into four main groups based on the association with cervical cancer. Category I consists of most important risk factors with highest correlation such as early marriage, first sexual intercourse at early age and multiple sexual partners. Category II contains risk factors with a likely correlation such as malnutrition, male sexual behavior and living conditions. Category III includes risk factors associated with increased cervical cancer rate like psychological factors. Cancer risk can be decreased by modifying those factors. Category IV is made with risk factors associated with increased cervical cancer rate which cannot be influenced such as age and family history [49]. In this study, K-Means algorithm is used to generate above groups of risk factors.

A study was conducted to apply advanced data mining techniques for recurrent cervical cancer with medical records provided by Chung Shan Medical University Hospital, Taiwan. In this study twelve variables including age, cell type, tumor grade, tumor size, pT, pStage, surgical margin involvement, LNM, number of fractions of other RT, RT target summary, sequence of Locoregional therapy and LSVI were analyzed with two advanced data mining techniques. They are multivariate adaptive regression splines (MARS) and C5.0 classifier model. This study found that pStage and pT are independent prognostic factors whereas cell type and RT target summary are significantly related to recurring of cervical cancer [50].

A research shows that RFT and K-mean algorithm can be easily used in medical data mining. RFT method plays a major role in classifying large data sets accurately. In the meantime, K-means algorithm was used to cluster the dataset into four different clusters and identify patterns among them. Based on above two algorithms, significant risk factors were identified and a weight was given for each risk factor. Literacy, marriage life and poor hygiene are the factors that lie at the top most among the risk factors according to the analysis [53].

Data mining techniques are increasingly becoming to play a vital role in assisting the physician in making sense out of massive data. As discussed in the research paper medical diagnosis process can be improved by applying the correct technique at the right

condition. Furthermore, this survey paper focused on a variety of data mining approaches like classification model, clustering model and bio inspirational model. This study reveals that each model differs in their performance depending on the type of dataset used. For instance predicting a disease with labeled dataset the classification model is well suited. The clustering model suits for pattern recognition among the several methods. In order to increase the performance of dataset with more optimization, the bio inspirational based techniques are suitable. The aim of this survey is to identify the usage of data mining techniques in diagnosing diseases in their early stages [10].

A research was done in order to predict normal cervix or cancer cervix with the help of data mining algorithms. The Regression Tree Algorithm was used for prediction. In this study, combination of two algorithms is used to validate the accuracy of outputs. Accurate prediction of occurrence of cervical cancer has been the most challenging and toughest task in medical data mining because of the non-availability of proper data set. Random Forest Tree algorithm achieves the predicted output by integrating independently distributed vectors of random type contained as a collection of tree-structured classifiers. RFT algorithm was associated with K-MEANS Learning. This study was an attempt to find out the solution for cervical cancer and give awareness to the women regarding the health issues. The research has described the prediction of cervical cancer in two stages i.e. Benign or Malignant of women with data mining algorithms, with reasonable accuracy [14].

3. METHODOLOGY

3.1.DATA MINING ALGORITHMS

Data mining techniques help in analyzing an enormous set of data and identifying previously unknown patterns or relationships among them. There are several data mining techniques such as association rule mining, clustering and classification to mine datasets. Amongst these, association rule mining is the most suitable algorithm for this research. It is because this project is interested in finding relationships between risk factor attributes.

Association Rule mining

Apriori algorithm is used in discovering interesting relations under association rule mining technique. This algorithm is mostly used in mining biological data, customer buying patterns, etc. Apriori uses a bottom up approach in generating subsets by extending one item at a time [42]. When selecting interesting rules from possible association rules, minimum thresholds on support and confidence will be used.

Support

The support, supp(X) of an itemset X is defined as the proportion of transactions in the data set which contains the itemset [45].

$$supp(X) = \text{no. of transactions which contain the itemset } X \text{ / total no. of transactions}$$

Itemsets above the threshold support are considered as significant itemsets.

Confidence

Confidence is an estimate of the probability P(Y | X), the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.:

$$conf(X \rightarrow Y) = supp(X \cup Y)/supp(X)$$

This says how likely item **Y** is there when item **X** also is there. However, this measurement misrepresents the importance of association since it only considers **X**. Therefore, another measurement, which considers the association, also needs to be analyzed

Lift

Lift is the performance of targeting association rules at classifying cases.

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(Y) * supp(X)}$$

This says how likely item **Y** is there when item **X** also is there while controlling how popular item **Y** is.

The Association Rule Generation process consists of two main steps. They are [45]:

1. Minimum support is applied to find all frequent itemsets in a database.
2. Frequent itemsets and the minimum confidence constraint are used to form rules.

Association rules with higher support than the threshold support, with high confidence values and with lift greater than or equals to one are taken into consideration when selecting interesting patterns.

3.2.RESEARCH METHODOLOGY

Knowledge Discovery Process

The knowledge discovery process is followed to mine the cervical cancer dataset and extract patterns in it. Figure 3.1 provides a pictorial representation of stages in this process.

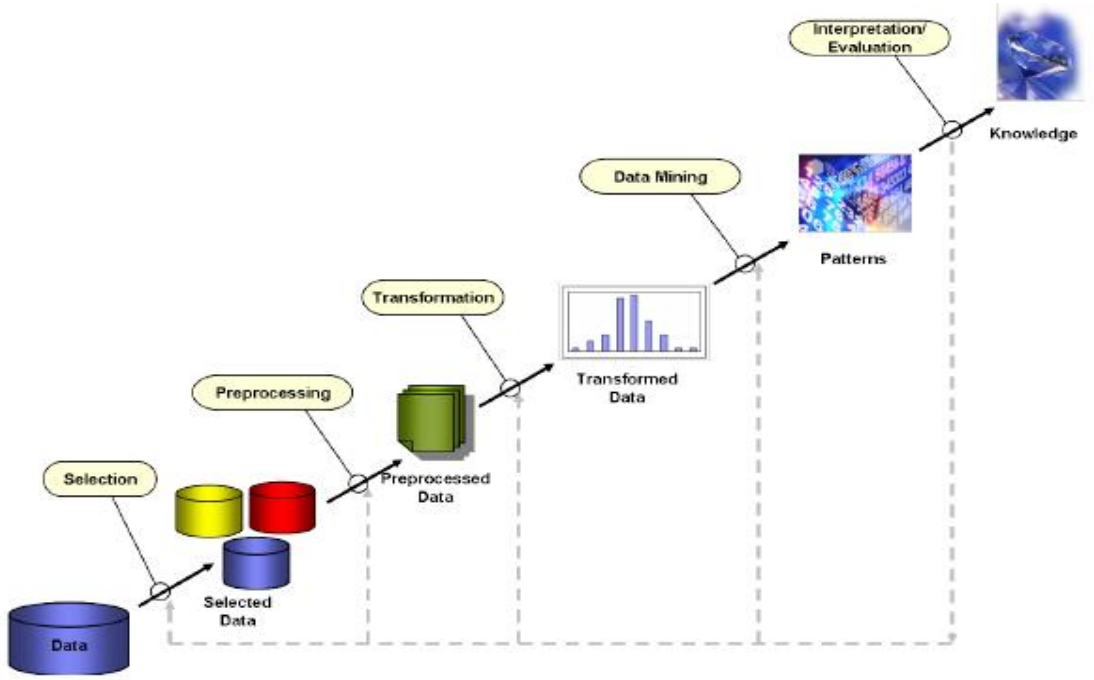

Figure 3.1: Steps in Knowledge Discovery Process

1. Selection

A comprehensive literature review was done in order to get a thorough understanding about the domain. Literature review was conducted mainly focusing on medical research and data driven approaches used with medical data. After getting a proper understand about the domain, the dataset was selected with considerable set of attributes relating to the problem domain.

2. Preprocessing

Preprocessing was conducted to remove the anomalies in the dataset. Since this is a medical dataset, there were a lot of missing data. That is mainly because some patients are reluctant to expose their demographic and habitual data to outside world. Applying data mining techniques to a dataset with missing values will not give a proper model. Those records were excluded from the study since it will not contribute in the pattern

recognition process and if they were there, computational algorithms will be misled.

3. Transformation

In the meantime, there were some attributes, which needed to be ignored from the study because it could be evaluated from another attribute. Such attributes were removed when finalizing the data set. Therefore, the data set was pruned initially by removing some attributes and anomalies.

In the dataset, there are three separate attributes to capture smoking habitual details. They are whether the respondent smokes, no. of years she has been smoking and no. of packets smoked per year. In this analysis, only the no. of packets smoked per year was considered. When analyzing medical data, smoking factor is considered in pack years. Below equation is used to calculate pack years.

$$\textit{Number of pack-years} = \textit{Packs smoked per day} \times \textit{No. of years as a smoker}$$

There are two individual attributes in the dataset for Intra Uterine Device (IUD) usage as well. One of them is whether respondent has used IUD or not and the other one is for how long she has been using IUD. The no. of years respondent has been using IUD will be considered in the study since the other attribute is already being covered by this. Hormonal contraceptives usage also has two separate attributes as in IUD. In this, how long she has been using hormonal contraceptives will be taken into consideration.

Several attributes are there related to Sexually Transmitted Diseases (STD) as well. They are whether the respondent has STD or not, no. of STD she has got and separate attribute for each STD stating whether the respondent has the disease or not. STD attributes such as human papilloma virus, hepatitis B virus, syphilis and HIV were taken into consideration individually and the other two attributes are ignored. Data set to be analyzed was pruned by removing above attributes.

Attributes in the pruned dataset

1. Age
2. Syphilis
3. Number of sexual partners
4. Genital herpes
5. First sexual intercourse (age)
6. AIDS
7. No. of pregnancies
8. HIV

9. Smoke (packs/year)
10. Hepatitis B
11. Hormonal Contraceptives (years)
12. HPV
13. IUD (years)
14. CIN
15. Vulvo-perineal Condylomatosis
16. PID
17. Cervical Condylomatosis
18. Molluscum Contagiosum
19. Vaginal Condylomatosis
20. Cancer

4. Data mining

Out of the three data mining techniques mentioned under the previous topic, the association rule mining technique will be used in this study in order to identify interesting patterns and relationships between risk factors. With this data mining algorithm, frequently seen risk factors in patients can be identified. Association between risk factors can be explained with the help of calculated measure values.

5. Pattern recognition

Frequent item sets will be appeared together in association rules. The process of association rule mining is used to find out the presence of which objects in a frequent item set, leads to the presence of other objects belonging to the same frequent item set.

This algorithm will be applied repetitively by changing the minimum support with the aim of identifying interesting patterns. Different set of combinations are generated with changed values and the set of associations with the best and interesting relationships are taken as valid output. Figure 3.2 provides a diagram with steps that were followed in identifying interesting patterns.

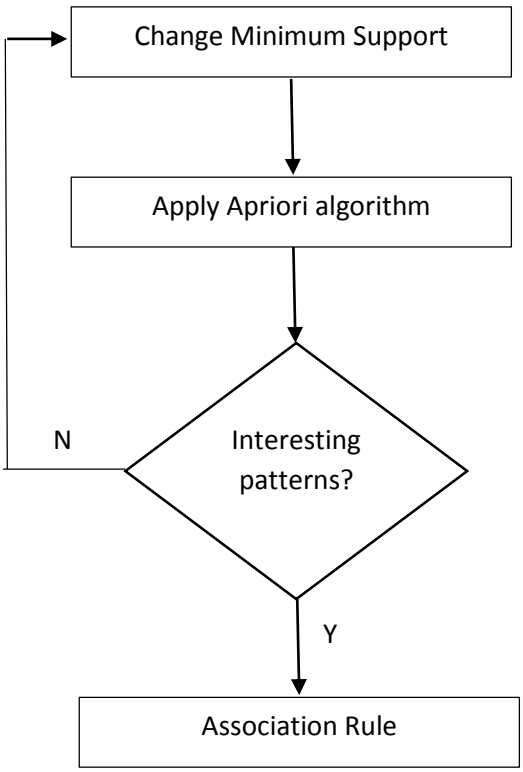

Figure 3.2: Process of Apriori Algorithm

In accordance with the process, first set minimum support to a higher value initially and then apply apriori algorithm. Check whether interesting patterns were generated for that particular minimum support value. If there are no any interesting pattern, change the minimum support value to a smaller value and apply apriori algorithm again. These should be done repetitively until an interesting pattern is generated. Once interesting pattern got generated, repetition process needs to be stopped. Then these generated association rules will be interpreted with the help of measure values calculated, under chapter 4.3.

4. EVALUATION AND RESULTS

4.1.EVALUATION PROTOCOL

Results obtained will be evaluated with the help of research papers, publications and discussions done with medical doctors. At the same time, a statistical package will be used in order to verify the relationships generated with Apriori algorithm. Bivariate correlation will be used to analyze two risk factors together and partial correlation will be used to analyze few variables together.

4.2.INDIVIDUAL RISK FACTOR ANALYSIS

Human Papilloma Virus (HPV)

Genital human papilloma virus (HPV) infection has been identified as one of the most possible causes of development of cervical cancers as well as cervical pre-cancerous changes. HPV collectively includes more than 150 viruses. HPV can spread from one person to another during skin-to-skin contact. It spreads via sex including vaginal, anal and oral [18]. HPV is the most common sexually transmitted infection. Almost all the cervical cancers are caused by HPV-16 and HPV-18 [19].

In accordance with figure 4.1, all the patients who suffer from cervical cancer are not infected with HPV. However, most of the patients have infected their selves with HPV. More than half of the respondents are infected with HPV.

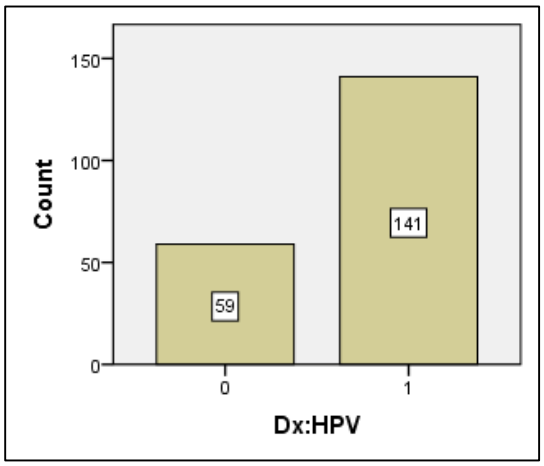

Figure 4.1: Individual factor contribution: HPV

Human Immunodeficiency Virus (HIV)

HIV kills or damages the cells in immune system. It spreads through unprotected sexual intercourse with an infected person or through contact with the blood of an infected person [21]. Women with HIV have a high risk of developing cervical cancer as well as cervical intraepithelial neoplasia (CIN) which is the pre-cancerous stage of cervical cancer [20]. According to a study, it is identified that women living with HIV in low- and middle-income countries are five times more likely to develop cervical cancer than HIV-negative women [22]. Hence, HIV can be considered as a significant risk factor for cervical cancer.

It is visible from figure 4.2 that more than half of the patients in the dataset are infected with HIV. However, there are a considerable number of people who do not suffer from HIV even though they are victims of cervical cancer. Therefore, HIV can be considered a risk factor which gives an average level significance.

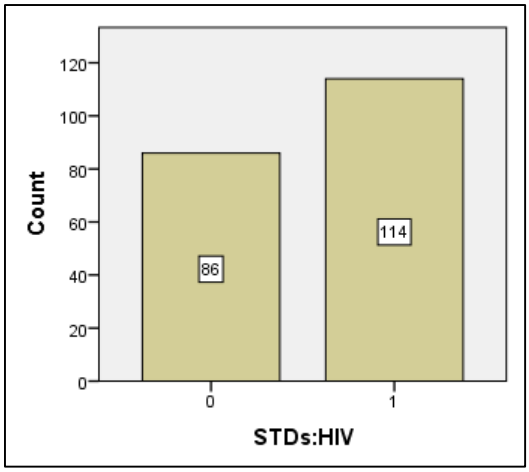

Figure 4.2: Individual factor contribution: HIV

Acquired immune deficiency syndrome (AIDS)

AIDS is a disease of the immune system caused by infection with HIV. It consists of group of signs and symptoms such as skin rashes or bumps, recurring fever, chronic diarrhea, persistent white spots on tongue or in mouth, persistent fatigue and soaking night sweats. The immune system of women with AIDS is severely damaged and there is a high chance of them developing cervical cancer.

Most of the respondents in the dataset are infected with AIDS according to figure 4.3. However, some patients do not have AIDS but suffer from cervical cancer. Since that count is comparatively small, AIDS can be considered a significant risk factor.

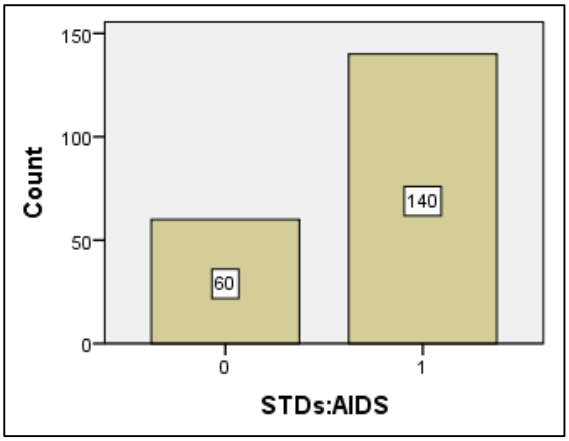

Figure 4.3: Individual factor contribution: AIDS

Cervical Intraepithelial Neoplasia (CIN)

Cervical intraepithelial neoplasia (CIN) is a precancerous condition in which abnormal cells grow on the surface of the cervix. CIN is known as cervical dysplasia as well. This is very common among women of age 25-35. CIN occurs when a woman is infected with HPV [28]. There are some common risk factors for both CIN and cervical cancer such as having multiple sexual partners, HIV infection and smoking [29]. Women who are not been treated for CIN have a high risk of developing cervical cancer.

Figure 4.4 shows that more than half of the respondents in the dataset suffer from CIN. There are few respondents, who do not have CIN but suffer from cervical cancer. Since the majority of the respondents have CIN, the dataset can be considered as behaving in a normal way.

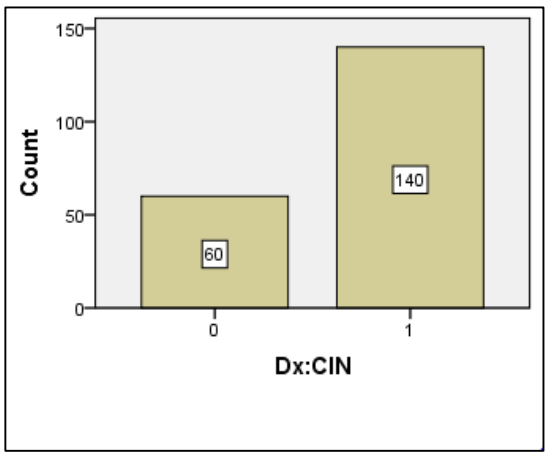

Figure 4.4: Individual factor contribution: CIN

Number of sexual partners

Having sex with many different partners increases the chance of encountering with a person who is infected with HPV and other sexually transmitted diseases. According to studies, having multiple different sexual partners can be considered as a high potential risk factor for cervical cancer [23]. Researchers have identified that having many sexual partners in young age increases the risk of cervical cancer. It is because the cervix changes during puberty and those changes make it more vulnerable to damage [24].

It is clearly shown from figure 4.5 that having sexual intercourse with multiple partners is a risk factor for cervical cancer. According to the dataset, having intercourse with 2-5 people is more than enough to be infected and later suffer from cervical cancer. Most of the respondents have had three sexual partners.

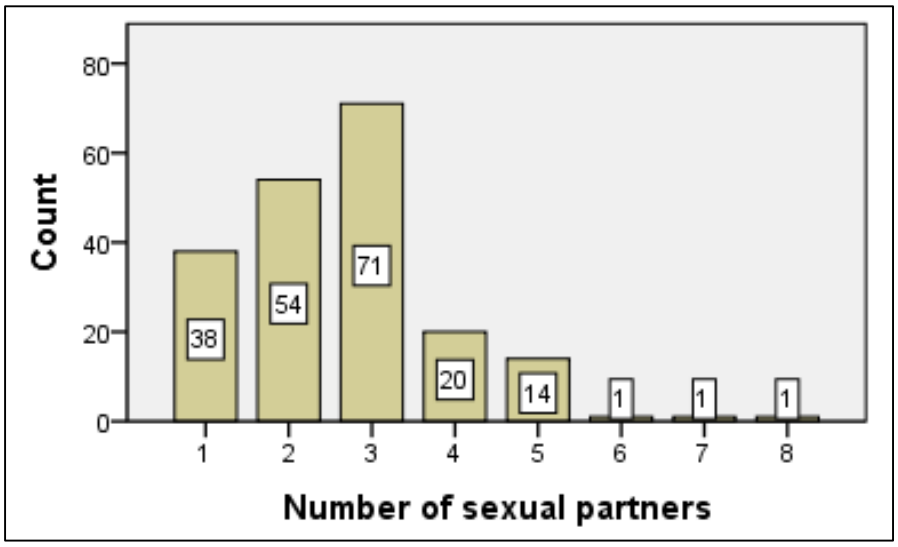

Figure 4.5: Individual factor contribution: No. of sexual partners

First sexual intercourse

First sexual intercourse at young age is an important risk factor of cervical cancer. Based on a study conducted in 2011, it was found that there is an increasing risk of cervical cancer with 95% confidence level in women who have had their first sexual intercourse at early age [25]. Most of the young women in developing countries have their first sexual intercourse when they are of age 16-19 [30]. Most probably, the first infection with HPV occurs soon after the first sexual intercourse [25].

It can be agreed with figure 4.6 that having first sexual intercourse at young age makes the way to cervical cancer. According to the below chart majority of women have had their first sexual intercourse at the ages of 14-18 which has a high risk of developing cervical cancer.

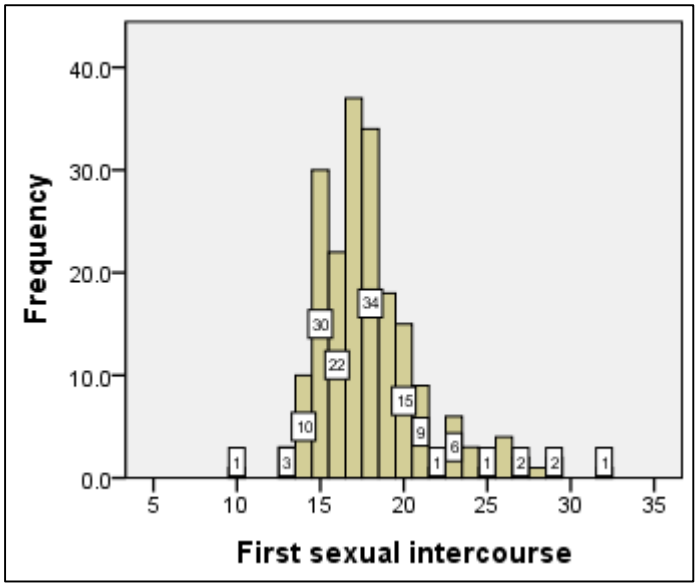

Figure 4.6: Individual factor contribution: Age

Number of pregnancies

There is a high risk of cervical cancer with women who have got pregnant several times. Women who have got pregnant three or more times have an increased risk of developing cervical cancer. Due to high levels of sex hormones present during pregnancy, there is a high chance of HPV infection and developing cervical cancer [24]. Pregnant women might have weaker immune systems which allows HPV infection. At the same time, these women may be exposed to HPV several times and get infected [18].

In accordance with figure 4.7, getting pregnant for 1-5 times can be considered as a high risk of developing cervical cancer. Majority of the women have got pregnant twice and are suffering from cervical cancer.

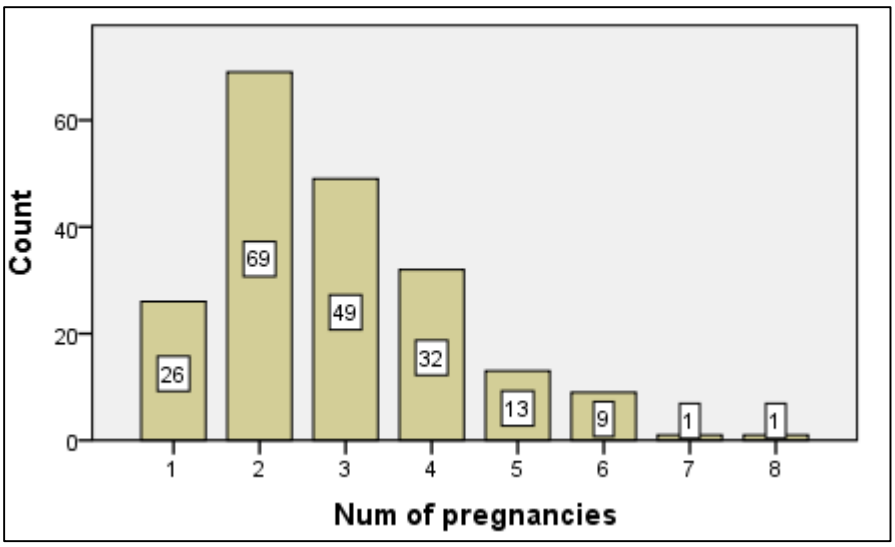

Figure 4.7: Individual factor contribution: No. of Pregnancies

Hormonal contraceptives

Hormonal contraceptives are used in birth controlling. Oral contraceptives fall under hormonal contraceptives and it means having birth control tables. Women who takes birth control tablets for 5 or more years have a high chance of getting cervical cancer. These women may not use any other protection method and have sex with partners who are infected with sexually transmitted diseases [26]. It is identified that women who had used oral contraceptives for 10 years or longer have four times higher risk of getting cervical cancer compared to women who had not used [27]. The hormones in oral contraceptives ease the HPV infection to cause cervical cancer [26].

It is agreed with the figure 4.8 that having hormonal contraceptives increases the risk of getting cervical cancer. Close around 2/3 of patients are using hormonal contraceptives and they are victims are slow sound only. At the same time, consuming hormonal contraceptives for prolonged period also increases the risk of cervical cancer. According to figure 4.9, using hormonal contraceptives for even few months causes cervical cancer.

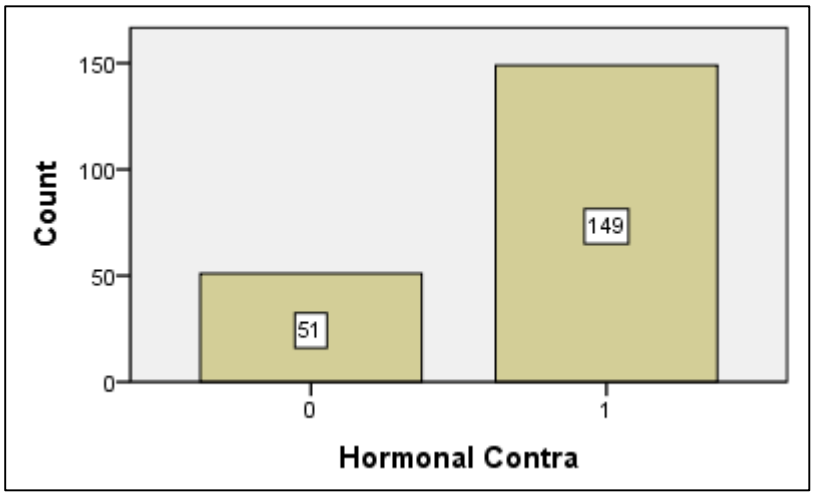

Figure 4.8: Individual factor contribution: Hormonal Contraceptives

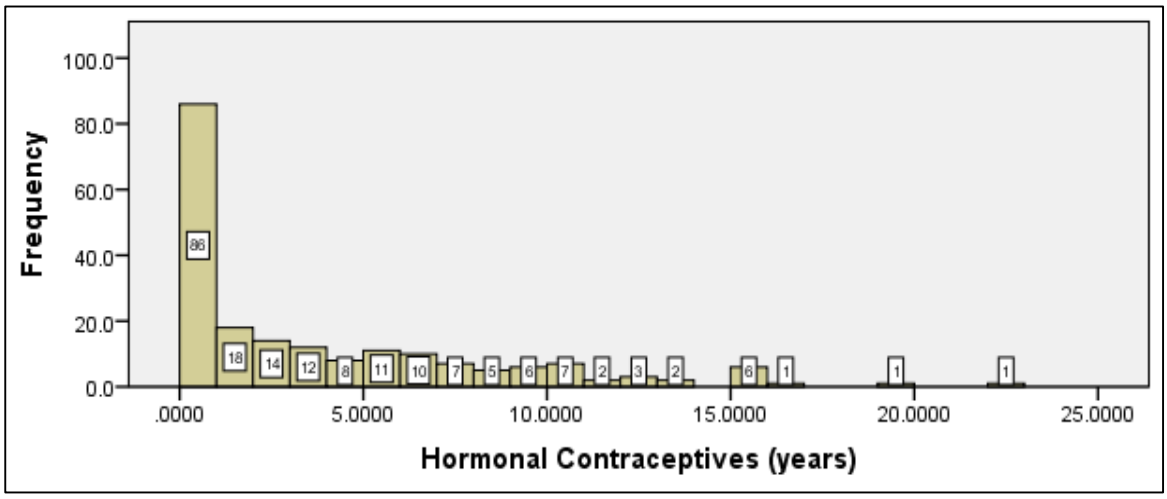

Figure 4.9: Individual factor contribution: Hormonal Contraceptives (Years)

Intra Uterine Device (IUD)

Women who use IUD as a birth control method have a lower risk of developing cervical cancer and the risk is been reduced by almost half [26]. The protective effect remains with the women who have used IUD for even less than a year [31]. The procedure to insert or remove an IUD may destroy HPV-related wounds before they become cancerous. Another way of reducing cancer effect is that hormone-targeting IUDs may affect the natural history of HPV infection [31].

According to figure 4.10, more than half of the women use IUD. According to medical research, it was revealed that using IUD has a lower risk of getting cervical cancer. However, according to the dataset, most of the patients use IUD and they suffer from

cervical cancer. Therefore, this dataset behaves in a different way compared to the normal behavior.

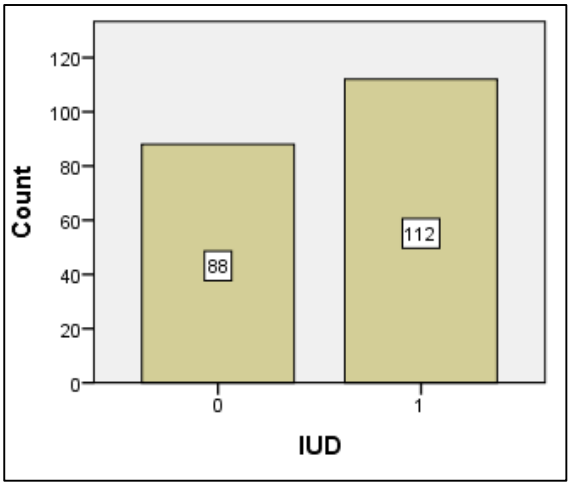

Figure 4.10: Individual factor contribution: IUD

Hepatitis B Virus (HBV)

Women with HBV may have higher risk of developing cervical cancer. Hepatitis B is a virus which spreads through contact with the blood and body fluids of an infected person. This is considered as one of the most common sexually transmitted diseases and is caused by having sex with an infected person or by having sex with multiple partners [32]. If a woman is having HBV, there is a high probability of that person been exposed to HPV as well. Hence, such women are more likely to get cervical cancer.

According to figure 4.11, more than half of the women in the dataset are infected with HBV. The dataset also proves that hepatitis B has a positive impact on development of cervical cancer.

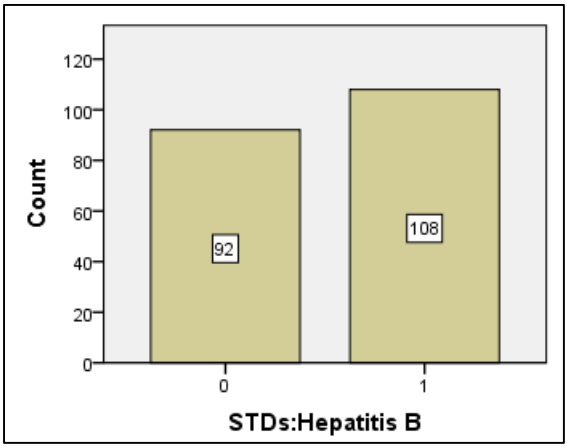

Figure 4.11: Individual factor contribution: Hepatitis B

Syphilis

Syphilis is a sexually transmitted disease (STD) which is caused by an infection with bacteria called Treponema pallidum. This is spread through by having sexual contact with an infected person [33]. According to research, it has identified that Syphilitic women develop carcinoma at an average age of 47 years, as compared to 51 years in non-syphilitic women in this series [34]. It is identified that syphilis, as other STDs increases chance of getting cervical cancer.

As shown in figure 4.12, most of the patients are suffering from syphilis. Therefore, it can be considered as a significant risk factor which in developing cervical cancer.

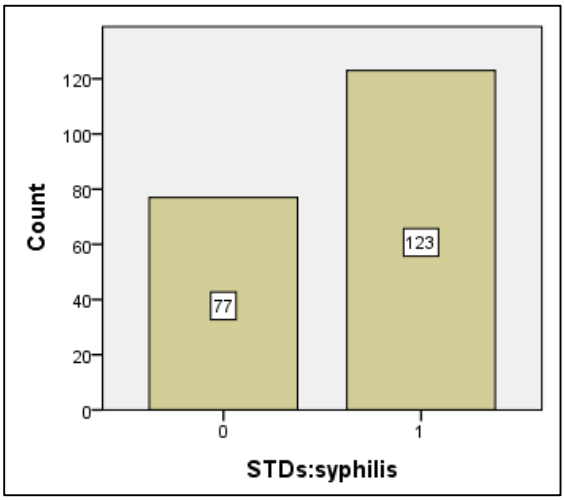

Figure 4.12: Individual factor contribution: Syphilis

Condylomatosis

Condylomatosis is genital warts transmitted via sexual contact. These warts may occur at cervix, vagina or vulvo-perineal area. Cervix is the lower part of the uterus in the human female reproductive system. Vagina is an elastic, muscular canal with a soft, flexible lining that provides lubrication and sensation [40]. Vulvo-perineal is the area outside and around the vagina. As other STDs, condylomatosis also increases the probability of getting cervical cancer.

According to figure 4.13, most of the patients are not suffering from any type of condylomatosis though they are victims of cervical cancer. Therefore, condylomatosis does not play a significant role in developing cervical cancer according to this dataset.

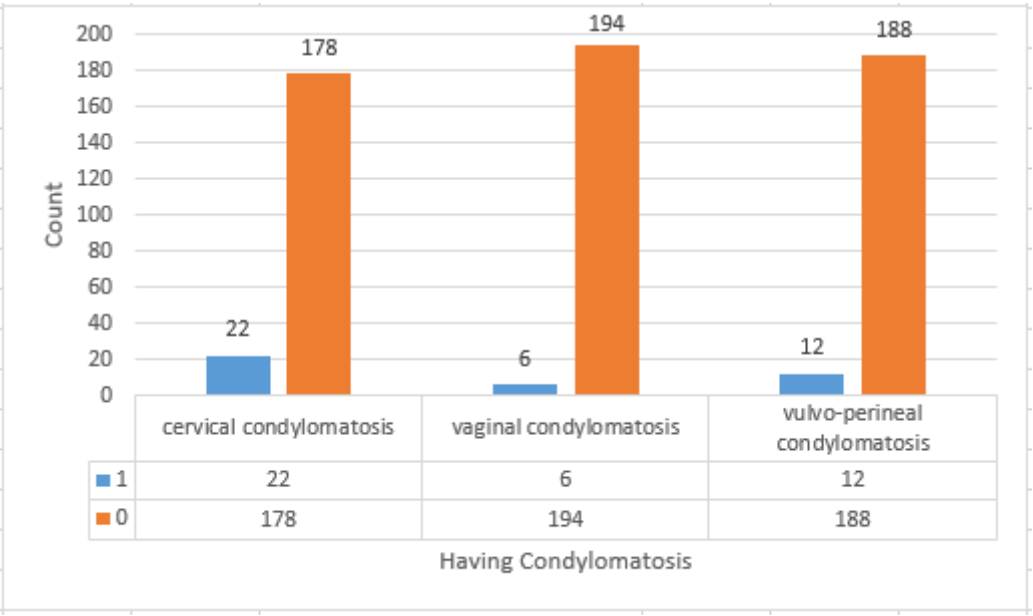

Figure 4.13: Individual factor contribution: Condylomatosis

Pelvic Inflammatory Disease (PID)

Pelvic inflammatory disease is an infection of the organs of a women's reproductive system which caused by a sexually transmitted infection like chlamydia or gonorrhea [35]. In addition, this can be caused by normal bacteria found in the vagina and on the cervix [36]. According to a research result, HPV occurrence was 33.74% in patients with PID and 26.40% in the group of women without PID [37]. Therefore, it can be said that PID has some sort of significance in getting cervical cancer.

With respect to figure 4.14, most of the cervical cancer patients are suffering from PID. Therefore, according to the dataset, it can be considered as a risk factor with average significance.

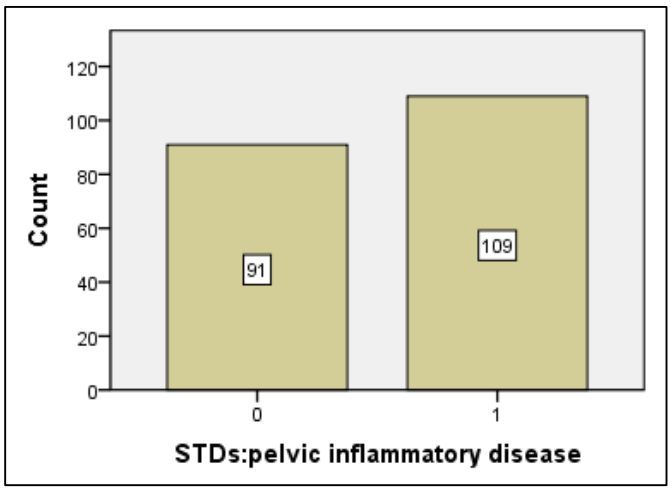

Figure 4.14: Individual factor contribution: PID

Genital Herpes

There is a relationship between genital herpes and development of cervical cancer. Herpes simplex virus-2 which causes genital herpes is infected through sexual contact. Herpes virus was detected in nearly half of women with cervical cancer. It was found that women infected with both HPV and HSV-2 were two to three times more likely to get cervical cancer. Herpes simplex virus-2 woks with HPV in boosting risk of cervical cancer [38].

According to the dataset, genital herpes can be considered as a significant risk factor of cervical cancer. Figure 4.15 shows that more than half of the patients in the dataset have genital herpes.

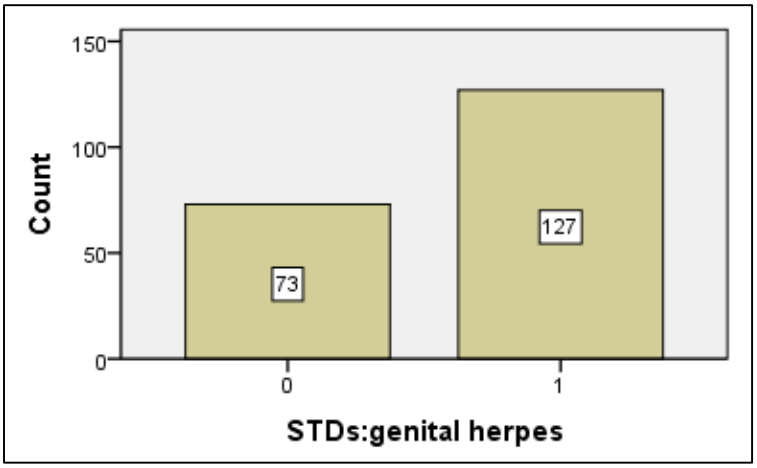

Figure 4.15: Individual factor contribution: Genital Herpes

Molluscum Contagiosum

Molluscum contagiosum is a viral infection of the skin that results in small, raised, pink lesions with a dimple in the center. This is called water warts as well. This virus is spread by either direct contact with an infected person or shared objects [39]. The virus is infected to people who are infected with HIV or are having sexual contact with infected people. Therefore, there is no direct relationship between molluscum contagiosum and cervical cancer. Hence, this factor can be considered as a non-significant risk factor.

According to figure 4.16 only a very few patients in the dataset has molluscum contagiosum. More than 90% of the patients do not have this disease. Therefore, it is proved that molluscum contagiosum can be considered a non-significant risk factor.

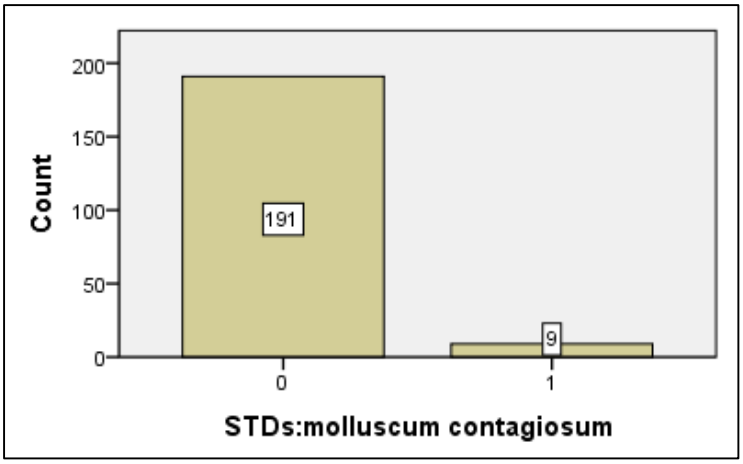

Figure 4.16: Individual factor contribution: Molluscum Contagiosum

Smoking

Smoking interferes with frequency of HPV infection and is strongly associated with cervical cancer. Although cervical cancer is caused primarily by HPV, cigarette smoking is considered as a cofactor. It is because certain types of HPV and cancer-causing chemicals related to smoking may work together to increase the possibility of developing cancer. According to a study, women who were exposed to three or more hours of smoke a day have about three times the risk of cervical cancer compared to women who do not smoke. Smoking prevents body's immune system from effectively fighting against HPV [17].

In accordance with the dataset most of the patients do smoke. Figure 4.17 shows that more than half of the respondents smoke and they are suffering from cervical cancer. Therefore, smoking can be considered a significant risk factor in developing cervical cancer.

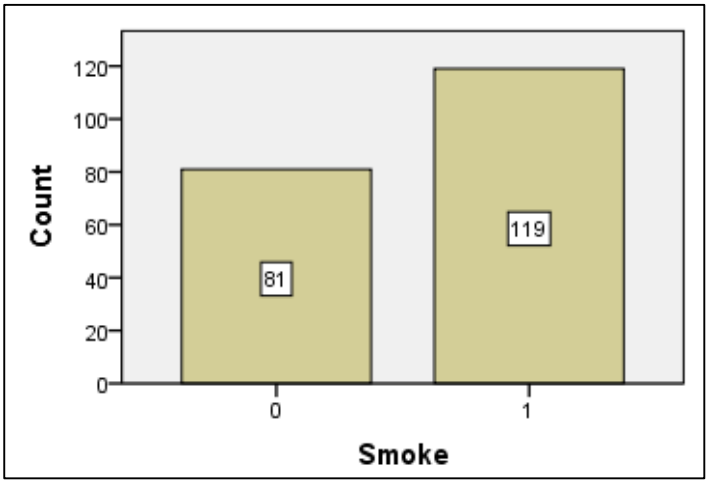

Figure 4.17: Individual factor contribution: Smoke

Age

Age is not a significant risk factor for cervical cancer. However, the risk of having this type of cancer increases between the late teens and mid-30s, which is considered as the reproductive age [28]. According to a study conducted in UK, Age-specific cervical cancer incidence rates rises suddenly from age 15-19 and peaks in the age 25-29 [29]. Figure 4.18 shows the behavior of age of the cervical cancer patients in UK. In the meantime, women over 40 years of age are at a risk and they need to have regular cervical cancer screenings.

The dataset consists of majority of women from 22-38 years of age who suffer from cervical cancer. Although age is not a significant risk factor figure 4.19 shows that there is a high risk of developing cervical cancer in late 20s.

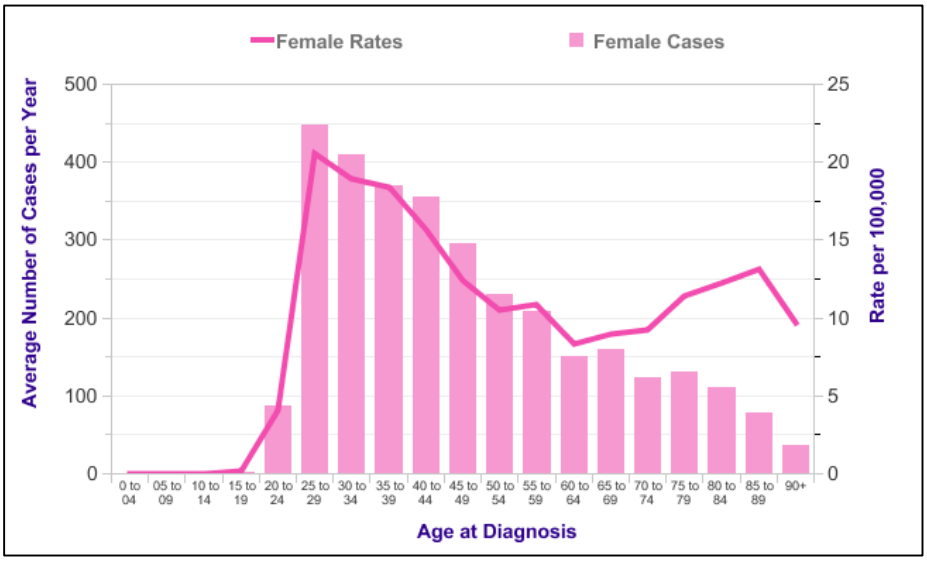

Figure 4.18: Age-Specific Incidence Rates, UK, 2012-2014

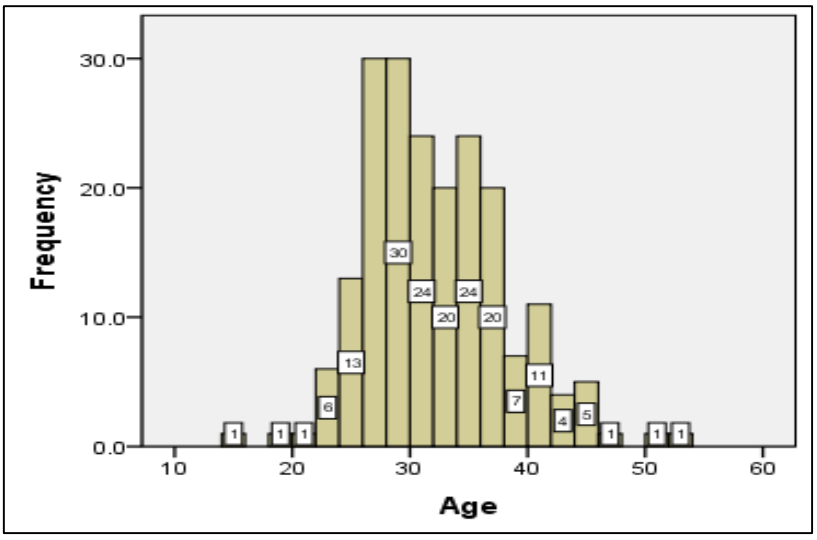

Figure 4.19: Individual factor contribution: Age

4.3.COMBINATIONAL RISK FACTOR ANALYSIS

Age and HPV

According to the dataset most of the patients of 25-37 years of age are infected with HPV. It shows that women in their mid 20's and 30's have a high chance of getting HPV may be due to their unethical sexual behaviors. Since this age group is the reproductive age of women, there is a high probability in increasing the spread of HPV if they are infected. This happens mainly due to the hormonal change in the women's body. Figure 4.20 generated for the dataset shows that most of the patients are concentrated between ages 25-37.

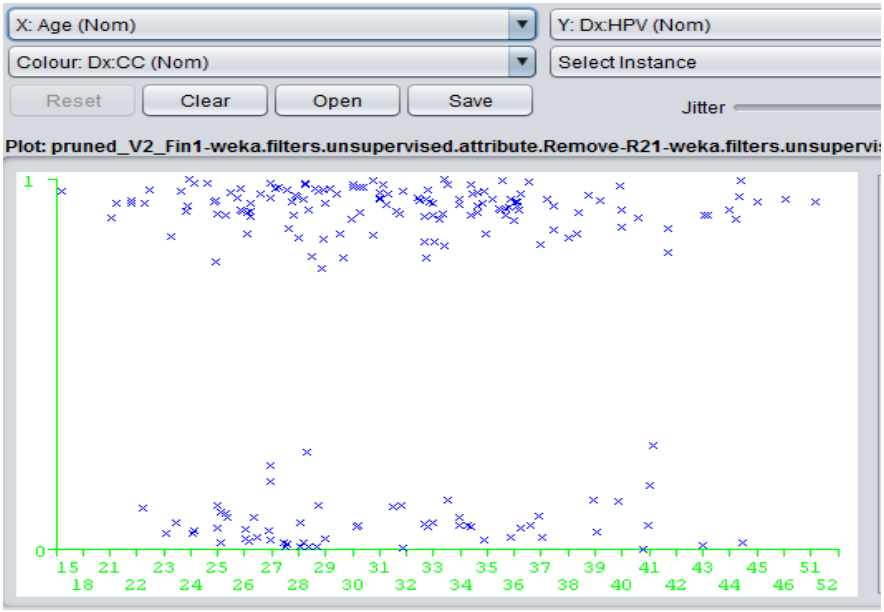

Figure 4.20: Age and HPV Combination

Hormonal Contraceptives and HPV

Figure 4.21 shows that women who are infected with HPV have a high probability of getting cervical cancer if they are using hormonal contraceptives at the same time. The area marked with a circle has high number of cervical cancer occurrences and it represents having HPV while using hormonal contraceptives. Hormonal contraceptives contains combined oestrogen and progestogen which affect the behavior of reproductive system of women. Therefore, women with HPV eases the spread of HPV by consuming hormonal contraceptives. It is clear that there is a positive relationship between HPV and hormonal contraceptives towards getting cervical cancer.

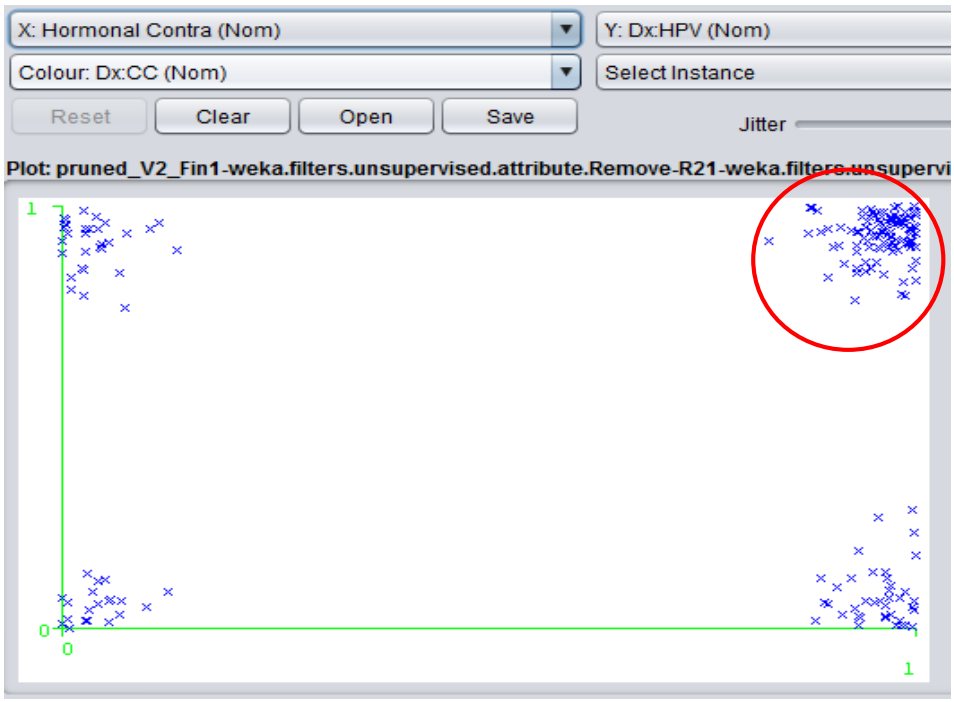

Figure 4.21: HPV and Hormonal Contraceptives

With figure 4.22, it is clear that irrespective of how long women are using hormonal contraceptives, they get cervical cancer if they are infected with HPV. In chapter 4.2, it is stated that according to medical research there is a high chance of getting cervical cancer if women are using hormonal contraceptives for more than 5 years. However, it is identified with the dataset that this prolong usage is hidden when they are infected with HPV while using hormonal contraceptives.

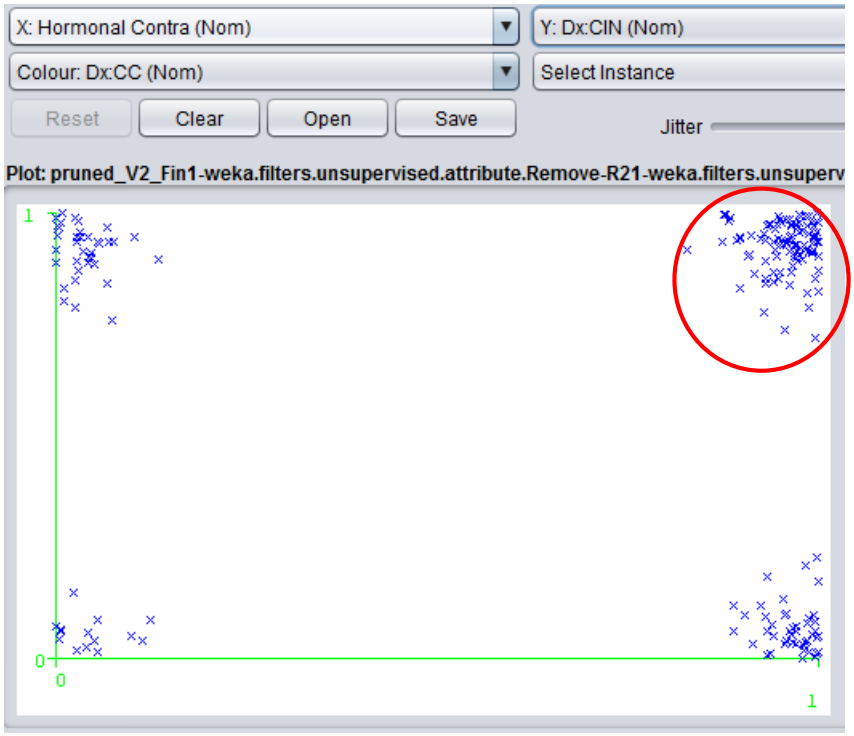

Figure 4.22: HPV and Hormonal Contraceptives (Years)

Hormonal Contraceptives and CIN

Figure 4.23 illustrates that there are large number of incidents under the circled are which denotes taking hormonal contraceptives while having CIN, the pre-cancerous stage of cervical cancer. In chapter 4.2, it is mentioned that CIN individually contributes the increase of cervical cancer. But the below diagram shows that the combination of using hormonal contraceptives while having CIN has a higher probability in getting cervical cancer than the individual factors. Hormones contains in hormonal contraceptives impact the growth of initial cancer cells in cervix and the chance in getting cervical cancer will improve due to this.

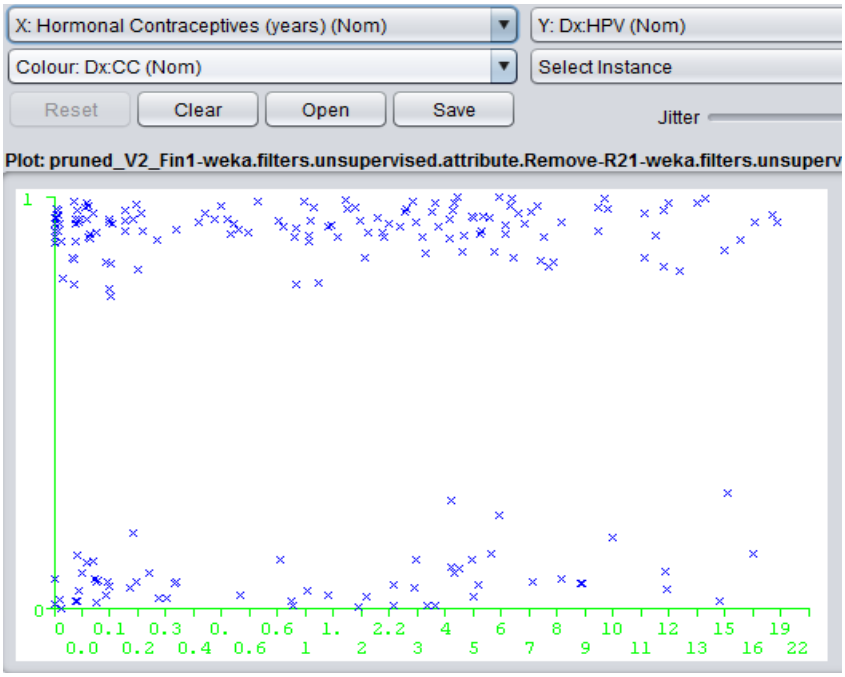

Figure 4.23: Hormonal contraceptives and CIN

Number of Sexual Partners and HPV

Women infected with HPV, having sexual intercourse with multiple partners have a higher chance of getting cervical cancer. According to the dataset, multiple partners does not mean 5-10 sexual partners. Concerning figure 4.24, only 2-4 partners are sufficient for a women with HPV to get cervical cancer. Most of the cervical cancer occurrences are concentrated around 1-4 sexual partners. When women infected with HPV having sexual intercourse with one or more infected sexual partners the possibility of getting cervical cancer increases a lot.

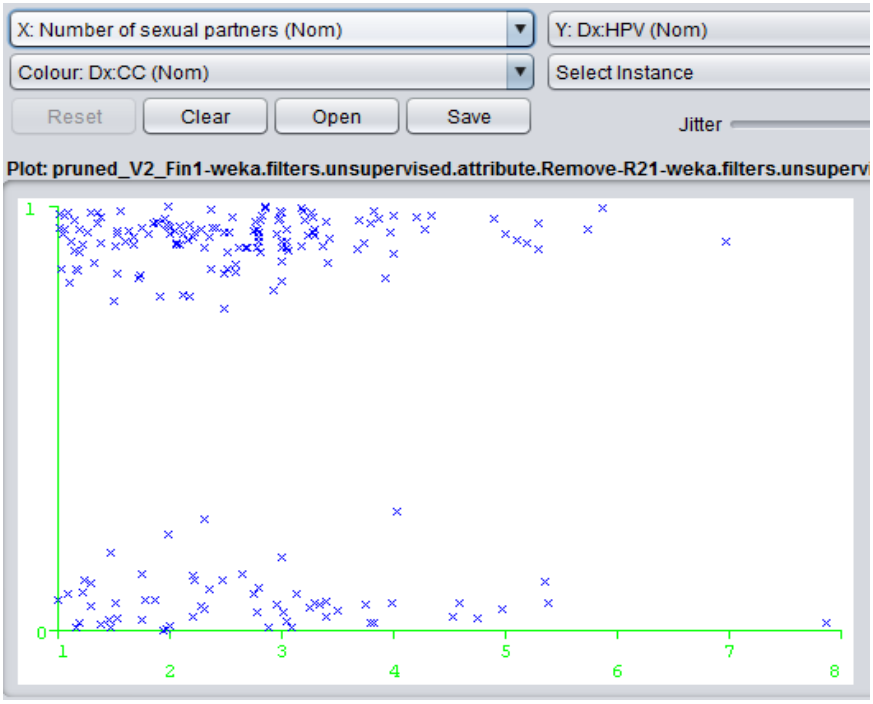

Figure 4.24: No. of Sexual partners and HPV

IUD and HPV

In accordance with figure 4.25, it is visible that women infected with HPV who use IUD as a birth control mechanism has a high chance of getting cervical cancer. Even though the chapter 4.2 stated that the possibility of getting cervical cancer is reduced by using IUD, when it is combined with HPV risk factor, the total possibility of getting cervical cancer is increased. This means that IUD acts in reverse way when it is associated with HPV.

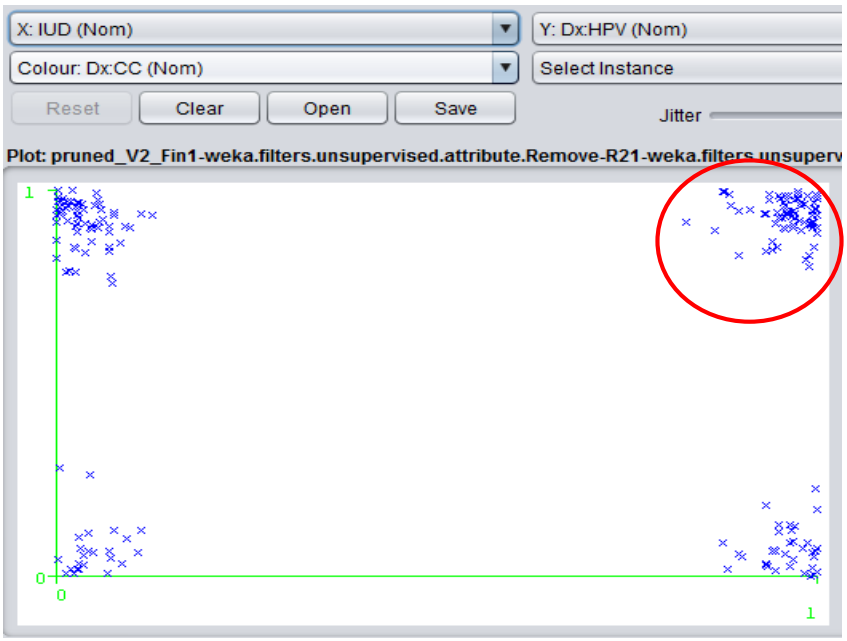

Figure 4.25: IUD and HPV

First sexual intercourse and Hormonal contraceptives

Women who have had their first sexual intercourse at teen age and using hormonal contraceptives meantime, have a high chance of developing cervical cancer. With figure 4.26, it is clearly shown that above combination of risk factors work positively in increasing the possibility of getting cervical cancer. Hormones contained in hormonal contraceptives helps in changing the conditions of reproductive system and when it combines with sexual intercourses at young age, the risk of getting cervical cancer rises. Teen age is considered as the reproductive age of women and when hormone concentrations are changed at that age cervical cancer is developed easily.

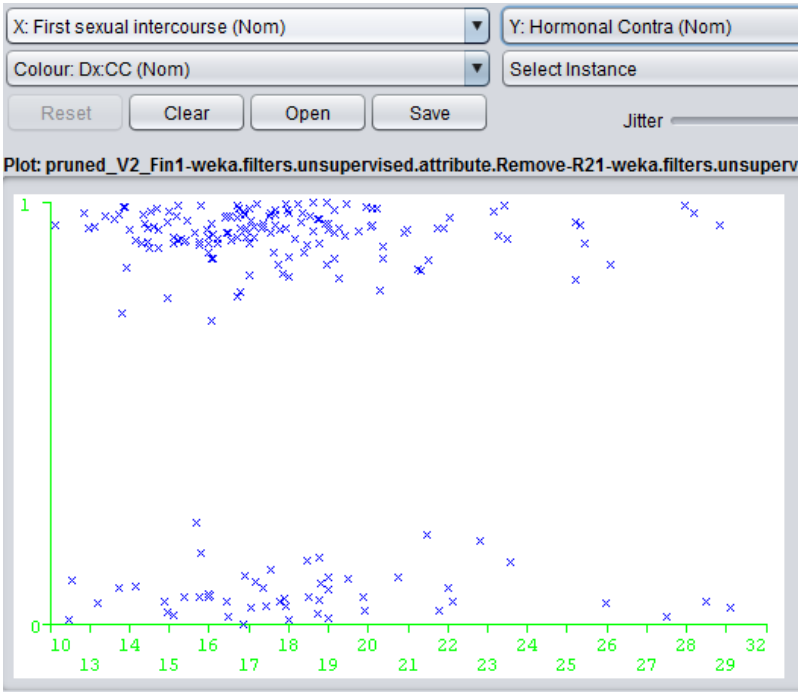

Figure 4.26: First sexual intercourse and Hormonal contraceptives

Number of pregnancies and hormonal contraceptives

There is a positive relationship between no of pregnancies and hormonal contraceptives in getting cervical cancer. The figure 4.27 shows that when number of pregnancies are 1- 5 and use hormonal contraceptives at the same time, the probability of getting cervical cancer increases. When women have got pregnant for several times the hormone concentration of the body changes and when they use hormonal contraceptives on top that hormone concentration changes further. This will affect the development of cervical cancer severely. Therefore getting pregnant multiple times and consuming hormonal contraceptives has a combinational influence the development of cervical cancer.

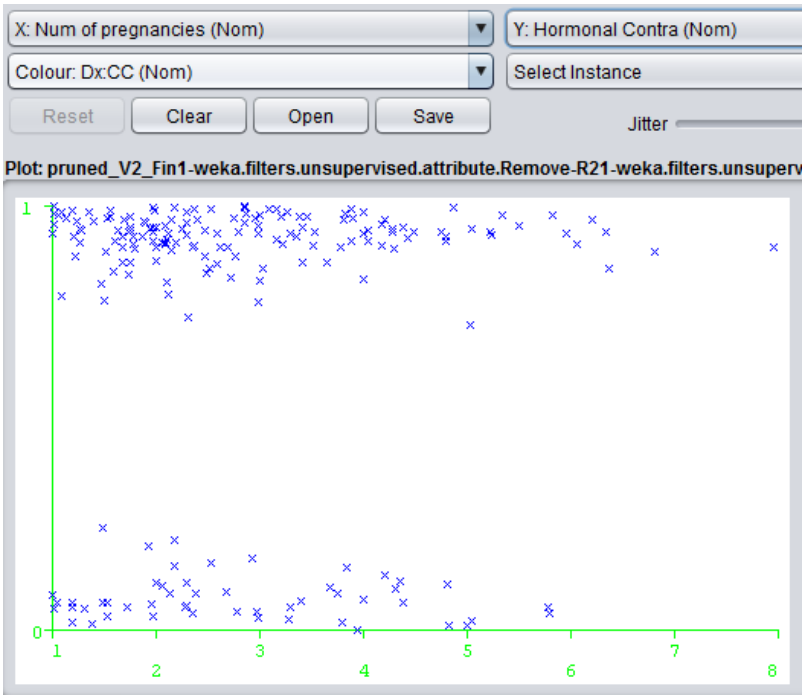

Figure 4.27: Number of pregnancies and hormonal contraceptives

Number of sexual partners and Hormonal Contraceptives

According to figure 4.28, women who have had multiple sexual partners while taking hormonal contraceptives have a high possibility of getting cervical cancer. When having more than one sexual partner, sexual hormone consistency of the body changes frequently and hormonal contraceptive usage contributes to that. With this combination, there is a high chance of getting cervical cancer in them,

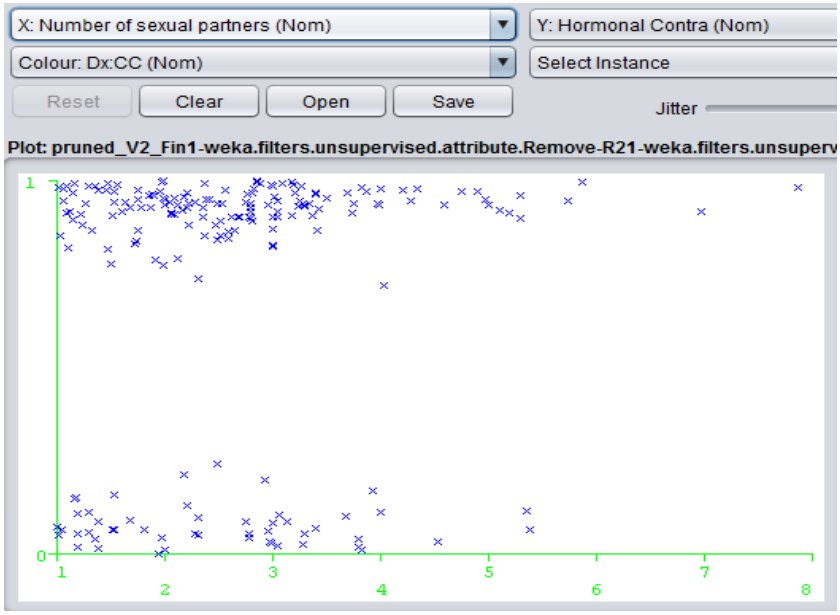

Figure 4.28: Number of sexual partners and Hormonal Contraceptives

Cervical condylomatosis and vaginal condylomatosis

In accordance with figure 4.29, cervical condylomatosis and vaginal condylomatosis negatively affect the development of cervical cancer. Chapter 4.2 stated that both of them increase the development of cervical cancer but according to this dataset women who do not have both cervical condylomatosis and vaginal condylomatosis have developed cervical cancer. Therefore, combination of cervical condylomatosis and vaginal condylomatosis does not influence in getting cervical cancer.

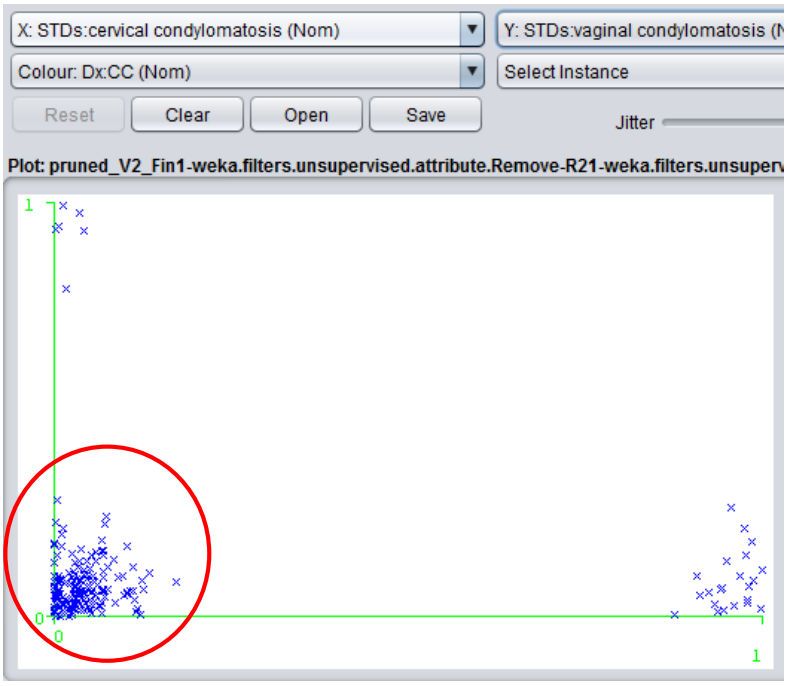

Figure 4.29: Cervical condylomatosis and Vaginal condylomatosis

4.4.RESULTS

Apriori algorithm was used to mine association rules among risk factors of cervical cancer. Association rules that satisfy the support threshold value are taken into consideration. In the meantime, generated rules are analyzed further in order to identify interesting patterns.

Support threshold for generated association rules is 0.7. All the rules which satisfy the support threshold are taken into consideration.

Mined association rules

STDs:HIV=1 114 ==> STDs:vaginal condylomatosis=0 112 conf:0.98 lift:1.01

In the dataset there are 114 patients who have HIV and 112 out of them do not have vaginal condylomatosis. Since the lift value is greater than 1, there is an association between HIV and vaginal condylomatosis. However, it is a negative relationship. 98% of the HIV victims do not have vaginal condylomatosis, which means these two risk factors are negatively associated in developing cervical cancer. Although both of these diseases are sexually transmitted, there is a high chance of HIV infected people not getting vaginal condylomatosis.

Dx:HPV=1 141 ==> STDs:vaginal condylomatosis=0 Dx:CC=1 138 conf:0.98 lift:1.01

Out of 141 patients who are infected with HPV, 138 do not have vaginal condylomatosis. Therefore, there is a negative association between HPV and vaginal condylomatosis according to the dataset. This association matches 98%. As explained in chapter 4.2, both of these diseases positively affect in getting cervical cancer. However, according to the data set HPV and vaginal condylomatosis have a negative association. That means there is a high chance of getting cervical cancer when women infected with HPV but do not have vaginal condylomatosis

STDs:AIDS=1 140 ==> STDs:molluscum contagiosum=0 135 conf:0.96 lift:1.01

According to the above association rule out of the 140 women infected with AIDS, 135 are not suffering from molluscum contagiosum and it is 96% when considered as

percentage. Similar to above association rules, this also has a negative association. As discussed in chapter 4.2, molluscum contagiosum does not impact directly to the development of cervical cancer. This association also proves the negative association with AIDS. Above association rule explains that patients who suffer from AIDS but do not suffer from molluscum contagiosum have a high chance of developing cervical cancer.

STDs:cervical condylomatosis=0 Dx:HPV=1 128 ==> STDs:vulvo-perineal condylomatosis=0 122 conf:0.95 lift:1.01

The above association rule states that out of 128 patients who are infected with HPV but do not have cervical condylomatosis, 122 do not have vulvo-perineal condylomatosis. This association has 95% of confidence. According to this HPV, cervical condylomatosis and vulvo-perineal condylomatosis has a negative association in developing cervical cancer. However, in chapter 4.2, it says that all these three attributes have a positive impact towards getting cervical cancer. In contrast to what is mentioned in chapter 4.2, the dataset shows that patients who are infected with HPV but do not have cervical condylomatosis have a high chance of not having vulvo-perineal condylomatosis and such patients are suffering from cervical cancer.

Hormonal Contra=1 Dx:HPV=1 110 ==> Dx:CC=1 110 conf:1 lift:1

According to above association rule, all the patients who are infected with HPV and using hormonal contraceptives are suffering from cervical cancer. As discussed in chapter 4.3, there is a positive relationship between HPV and hormonal contraceptives. It is proved in this chapter with the above association rule. As stated in chapter 4.3, when women infected with HPV uses hormonal contraceptives it will affect the reproduction system and eases the development of cervical cancer.

Smoke=1 Hormonal Contra=1 86 ==> Dx:CC=1 86 conf:1 lift:1

In the dataset, there are 86 patients who smoke and uses hormonal contraceptives and all of them suffer from cervical cancer. Therefore, there is a positive relationship between smoking and hormonal contraceptive usage towards cervical cancer. As mentioned in chapter 4.3, the sexual hormone concentration of the body changes when using hormonal contraceptives. However, when they smoke on top of that those hormones work together with chemicals infected when smoking and boost up the cervical cancer cells.

IUD=1 STDs:AIDS=1 77 ==> Dx:CC=1 77 conf:1 lift:1

In the dataset, 77 patients are infected with AIDS and uses IUD, and all of them are victims of cervical cancer. Under chapter 4.2, it is mentioned that IUD reduces the chance of getting cervical can by healing the wounds in cancerous cells. However, with the above association rule it is proved that both IUD and AIDS together contribute positively to the development of cervical cancer.

This can be considered as an interesting relationship since the medical research says that IUD reduces the risk of getting cervical cancer but with the data driven approach it is found that IUD positively contribute to the development of cervical cancer. This relationship does not consider about individual factor. Therefore, this does not mean that all the women who use IUD has a high risk of developing cervical cancer or all the women who suffer from AIDS have a high risk. However, when it comes to the combination of AIDS and usage of IUD there is a high risk of developing cervical cancer.

Num of pregnancies=2 Hormonal Contra=1 56 ==> Dx:CC=1 56 conf:1 lift:1

In the dataset there are 56 women who have got pregnant 2 times and uses hormonal contraceptives, and all of them are victims of cervical cancer. There is a positive relationship between number of pregnancies and hormonal contraceptives towards cervical cancer. In chapter 4.2 even though it is mentioned that there is a high risk in developing cervical cancer when number of pregnancies are getting increased, the dataset shows that the count will not be that effective and getting pregnant twice would be enough for them to develop cervical cancer.

Smoke=1 IUD=1 Dx:CIN=1 Dx:HPV=1 31 ==> Dx:CC=1 31 conf:1 lift:1

Above-mentioned association rule is also an interesting rule since it shows a combination of several risk factors. According to the association rule, there are 31 women who smoke, use IUD, suffering from CIN and affected with HPV and all of them are victims of cervical cancer. Although the medical research says that using IUD reduces the risk of getting cervical cancer, IUD acts as a positive factor when it is combined with other risk factors. When a women is infected with HPV and she is smoking, the chemicals in cigarettes work together to increase the spread of HPV. On top of that, when the women has CIN and uses IUD there are wounds in the cervix which may turn into cervical cancer cells. Consequently, IUD positively influences developing cervical cancer when it is with other risk factors like smoking, CIN and HPV.

4.5.RESULTS EVALUATION

Results obtained with data driven approach are verified via a statistical approach. If both the methods provide the same answer, the selected data driven approach can be considered as an accurate tool.

According to table 1, the output given by statistical approach and data driven approach are same. For instance, when binomial correlation was applied on HIV and vaginal condylomatosis it gives a negative correlation value. Also according to the data driven approach, there is a negative association between HIV and vaginal condylomatosis**.**

Risk factors	**Statistical Approach**	**Data-driven Approach**
	Bivariate/Partial correlation	**Association**
HIV and vaginal condylomatosis	-0.084	negative
HPV and vaginal condylomatosis	-0.079	negative
AIDS and molluscum contagiosum	-0.068	negative
HPV, cervical condylomatosis, vulvo-perineal condylomatosis	-0.086	negative
HPV and hormonal contraceptives	0.125	positive
Smoke and Hormonal contraceptives	0.062	positive
AIDS and IUD	0.031	positive
No. of pregnancies and hormonal contraceptives	0.026	positive
Smoke, CIN, HPV, IUD		positive
CIN, HPV, IUD	0.036	
Smoke, CIN, HPV	0.015	
Smoke, HPV, IUD	0.114	

Table 4.1 : Results of data-driven and statistical approaches

Bivariate correlation was applied to relations with two risk factors and they are,

- HIV and vaginal condylomatosis
- HPV and vaginal condylomatosis
- AIDS and molluscum contagiosum
- HPV and hormonal contraceptives
- Smoke and Hormonal contraceptives
- AIDS and IUD
- No. of pregnancies and hormonal contraceptives

Partial correlation was applied on below relations.

- HPV, cervical condylomatosis, vulvo-perineal condylomatosis
- Smoke, CIN, HPV, IUD

As mentioned in table 1, first four relations give a negative correlation as well as negative association where as next five relations give a positive correlation as well as positive association.

5. CONCLUSION AND FUTURE WORK

5.1.SUMMARY

The research project was conducted in order to identify risk factors and their impact on cervical cancer. A critical literature review was conducted in order to get familiar with the study domain. At the same time, discussions were conducted with medical doctors in order to clarify confusing areas.

Dataset was preprocessed and cleaned before applying analytical algorithms. Individual factor analysis was conducted manually in order to identify the spread of the dataset. Then a combinational factor analysis was done in order to get an understanding about the behavior of the dataset further more.

Data mining techniques were used to identify relationships among risk factors. Among the data mining methodologies, the most suitable methodology was identified and it was applied on the data set. Interesting association rules were selected out of the set of rules generated. Statistical approach was used to verify the results generated from data mining techniques.

5.2.RESEARCH FINDINGS

Individual risk factor analysis was conducted with the aim of identifying the significance of each factor towards cervical cancer. It was identified that HPV, Condylomatosis, Syphilis, Genital herpes, Hepatitis B, AIDS, HIV, CIN, PID, Cancer, hormonal contraceptives, smoking, no. of sexual partners, no. of pregnancies and age at first sexual intercourse have a significant impact in causing cervical cancer whereas IUD and molluscum contagiosum are not that significant according to the general study conducted on risk factors of cervical cancer. Though the chapter 4.2 says that condylomatosis is a significant risk factor in accordance with the general study on risk factors, it was not very significant according to the dataset. In the meantime, according to the dataset analysis done under chapter 4.2, it was found that IUD could be considered as an average significant risk factor though the general risk factor study says that it is not a significant risk factor.

Combinational risk factor analysis also was done in order to find the combinational behavior of risk factors. Age and HPV, hormonal contraceptives and HPV, hormonal

contraceptives and CIN, No. of sexual partners and HPV, IUD and HPV, first sexual intercourse (age) and hormonal contraceptives, No. of pregnancies and hormonal contraceptives, No. of sexual partners and hormonal contraceptives, cervical condylomatosis and vaginal condylomatosis are the combinations analyzed under chapter 4.3. Amongst many combinations only above were taken into consideration since they show significant relationships as combinations. Out of these selected combinations, IUD and HPV was found as an interesting relationship since when both factors were present, the risk of getting cervical cancer increased, though the literature says that IUD reduces the risk of getting cervical cancer.

Apriori algorithm was applied on the dataset to find significant and interesting association rules. Several interesting relationships were identified from the association rules. Some of them are HPV and vaginal condylomatosis, AIDS and molluscum contagiosum, AIDS and IUD and Smoke, CIN, HPV, IUD. However, HPV individually has a positive relationship with cervical cancer when it is combined with vaginal condylomatosis, the relationship towards cervical cancer has become negative. The combination of AIDS and molluscum contagiosum also reduces the risk of getting cervical cancer though they have a positive impact when considered individually. In the meantime AIDS and IUD combination positively affect the development of cervical cancer although IUD individually is considered as a risk factor which reduces the chance of getting cervical cancer. Smoke, CIN, HPV and IUD together increase the possibility of getting cervical cancer irrespective of their individual behavior.

When considering the combinational risk factors, some factors that were considered as significant individually do not have a considerable impact on getting cervical cancer when combined with some other factor. At the same time, some factors which do not have a significant impact individually, have a considerable significance when combined with some other risk factors.

5.3.LIMITATIONS

It was very difficult to find a proper dataset to apply data mining techniques. Since it is a medical dataset there were many problems with the dataset such as missing values and those values needed to be handled. Handling missing values using algorithms is not suitable when working with medical data. Therefore, records with missing values needs to be removed from the study dataset. Once the missing values were removed, the total

record count was reduced. Meantime it is better to apply data mining algorithms to a very large dataset in order to get the most accurate results. Due to missing values dataset was reduced and was not able to conduct data mining process on a large dataset.

The dataset consists of several cervical cancer causing risk factors and the research was limited to those attributes. Therefore there is no opportunity to check the behavior of cervical cancer risk factors with other factors, which may be considered as unseen patterns based on the behavior of factors. Hence the research is limited to general risk factors of cervical cancer.

In order to get most accurate results for a general population data mining algorithms needs to be applied to several datasets from several populations. Due to unavailability of data research was conducted with a data set belonging to Venezuela, which is a developing country. Hence the research findings are proved for developing countries.

5.4.FUTURE WORK

The study was conducted with a dataset which belongs to Venezuela and it needs to be conducted on several datasets in order to get a global understanding. Datasets needs to be find from developed countries and non-developed countries as well and apply data mining algorithms on them. With this, interesting patterns in cervical cancer risk factors can be found in a global aspect.

A study needs to be conducted with a Sri Lankan dataset as well mainly because cervical cancer is becoming the second most common cancer among women in Sri Lanka. In the meantime HPV spread is on the rise in Sri Lanka, thus people need to be made aware of the damage it can cause. Therefore, the Sri Lankan aspect of cervical cancer and HPV distribution needs to be studied.

There may be some other hidden factors which may contribute significantly in developing cervical cancer but do not affect it directly. A study needs to be conducted to identify factors which are does not have a direct impact on the development of cervical cancer. A dataset should be found with both significant and indirect risk factors to apply data mining techniques on it. With such factors, unseen patterns can be identified and later can be redefined as a new knowledge.

At the same time indirect risk factors may not contribute to development of cervical cancer as individuals. They may increase the probability of getting cervical cancer combined with one or two general risk factors of cervical cancer. These type of patterns needs to be identified by applying data mining algorithms. In order to conduct this study, a proper dataset with required attributes needs to be analyzed.

6. REFERENCES

[1] Cervical Cancer. (2017, Jul.). Cervical Cancer – Topic overview. WebMD. [Online]. Available: http://www.webmd.com/cancer/cervical-cancer/cervical-cancer-topic-overview#1

[2] Cancer Research UK . (2017, May). Cervical Cancer. [Online]. Available: http://www.cancerresearchuk.org/about-cancer/cervical-cancer/stages-types-grades/types-and-grades

[3] Cancer Research UK . (2014, Jan). Small cell cancer of the cervix. [Online]. Available: http://www.cancerresearchuk.org/about-cancer/cervical-cancer/stages-types-grades/small-cell-cancer-cervix

[4] Information Centre on HPV and Cancer SL. (2017, Jul.). Human Papillomavirus and Related Diseases Report. [Online]. Available: http://www.hpvcentre.net/statistics/reports/LKA.pdf

[5] A.D. Mwaka et al. (2016, Aug). Awareness of cervical cancer risk factors and symptoms: cross-sectional community survey in post-conflict northern Uganda. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4957614/

[6] C.A. Opoku et al. (2016, Jun). Perception and risk factors for cervical cancer among women in northern Ghana. [Online]. Available: http://www.panafrican-med-journal.com/content/article/22/26/full/#.WXvFtFUjHDc

[7] American Cancer Society. (2016, Nov). What Are the Risk Factors for Cervical Cancer? [Online]. Available: https://www.cancer.org/cancer/cervical-cancer/causes-risks-prevention/risk-factors.html

[8] Wikipedia. (2018, Mar.). Disease. [Online]. Available: https://en.wikipedia.org/wiki/Disease

[9] World Health Organization. (2016, Jun.). Human papillomavirus (HPV) and cervical cancer. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs380/en/

[10] K. Lokanayaki et al. (2013, Sep.). Exploring on various prediction model in data mining techniques for disease diagnosis. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.402.8670&rep=rep1&type=pdf

[11] CancerIndex. (2017, Mar.). Venezuela. [Online]. Available: http://www.cancerindex.org/Venezuela

[12] Venezuela. (2017, Jul.). Human Papillomavirus and related cancers, Fact sheet 2017. [Online]. Available: http://www.hpvcentre.net/statistics/reports/VEN_FS.pdf

[13] Cancer research UK. (2017, Feb.). World Cancer Day 2017 : how to prevent cervical cancer cases around the globe. [Online]. Available: http://scienceblog.cancerresearchuk.org/2017/02/08/world-cancer-day-2017-how-to-prevent-cervical-cancer-cases-around-the-globe/

[14] R. Vidya et al. (2016, Aug.). Prediction of cervical cancer using hybrid induction technique: A solution for human hereditary disease patterns. [Online]. Available: http://www.indjst.org/index.php/indjst/article/viewFile/82085/72318

[15] American society of clinical oncology. (2017, Jul). Cerival Cancer: Risk Factors. [Online]. Available: http://www.cancer.net/cancer-types/cervical-cancer/risk-factors

[16] Cancer research UK. (2014). Cervical Cancer incidence statistics. [Online]. Available: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/cervical-cancer/incidence#heading-One

[17] M.R. Vann. (2009, Feb.). Stop smoking and reduce your cervical cancer risk. [Online]. Available: https://www.everydayhealth.com/cervical-cancer/smoking-risk.aspx

[18] Cancer research UK. (2016,Dec.). What are the risk factors for cervical cancer. [Online]. Available: https://www.cancer.org/cancer/cervical-cancer/causes-risks-prevention/risk-factors.html

[19] E.R.Bahmanyar et al. (2012, Dec.). Prevalence and risk factors for cervical HPV infection and abnormalities in young adult women at enrolment in the multinational PATRICIA trial. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0090825812007263

[20] J.Frederick. (2017, Mar.). Certain HPV Strains Increase Cervical Cancer Risk in HIV-Infected Women, Study Shows. [Online]. Available: https://cervicalcancernews.com/2017/03/24/certain-hpv-strains-increase-cervical-cancer-risk-women-hiv/

[21] L.Gregory. (2016, Sep.). Journal of HPV and Cervical Cancer. [Online]. Available: https://www.omicsonline.org/cervical-cancer.php

[22] Kaliane. (2017, Jan.). The Double Burden: HIV and Cervical Cancer Webinar with the International AIDS Society. [Online]. Available:

http://pinkribbonredribbon.org/the-double-burden-hiv-and-cervical-cancer-webinar-with-the-international-aids-society/

[23] Z.C.Liu. (2016). Multiple Sexual Partners as a Potential Independent Risk Factor for Cervical Cancer: a Meta-analysis of Epidemiological Studies. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25987056

[24] Cervical cancer. (2017). Risk factors for cervical cancer. [Online]. Available: http://www.cancer.ca/en/cancer-information/cancer-type/cervical/risks/?region=on

[25] M.Plummer et al. (2011, Aug.). Time since first sexual intercourse and the risk of cervical cancer. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/ijc.26250/pdf

[26] Cervical cancer. (2013, Dec.). Risk factors. [Online]. Available: http://www.nytimes.com/health/guides/disease/cervical-cancer/risk-factors.html?mcubz=3

[27] V.Moreno et al. Effect of oral contraceptives on risk of cervical cancer in women with human papillomavirus infection: the IARC multicentric case-control study. Lancet 2002; 359(9312):1085–1092. [PubMed Abstract]

[28] Cleveland clinic. (2014). Cervical Intraepithelial Neoplasia (CIN). [Online]. Available: https://my.clevelandclinic.org/health/articles/cervical-intraepithelial-neoplasia

[29] MedlinePlus. (2017, Oct.). Cervical dysplasia. [Online]. Available: https://medlineplus.gov/ency/article/001491.htm

[30] K.S. Louie. (2009, Apr.). Early age at first sexual intercourse and early pregnancy are risk factors for cervical cancer in developing countries. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2670004/

[31] S. Boyles. (2017). IUDs for Birth Control May Cut Cervical Cancer Risk. [Online]. Available: http://www.webmd.com/sex/birth-control/news/20110912/iud-for-birth-control-may-cut-cervical-cancer-risk#1

[32] J.Marcin. (2017, Mar.). Hepatitis B. [Online]. Available: http://www.healthline.com/health/hepatitis-b

[33] Sexual conditions. (2017). Syphilis. [Online]. Available: http://www.webmd.com/sexual-conditions/syphilis#1

[34] W.G. Harding. (2017, Oct.). The Influence of Syphilis in Cancer of the Cervix Uteri. [Online]. Available: http://cancerres.aacrjournals.org/content/canres/2/1/59.full.pdf

[35] Women's health. (2017). What Is Pelvic Inflammatory Disease? [Online]. Available: https://www.webmd.com/women/guide/what-is-pelvic-inflammatory-disease

[36] Breaking the stigma. (2017, Oct.) Pelvic Inflammatory Disease – PID – A Serious Bacterial STD in Women. [Online]. Available: https://www.thestdproject.com/pelvic-inflammatory-disease-pid-serious-std-women/

[37] J.Skapinyecz. (2003). Pelvic inflammatory disease is a risk factor for cervical cancer. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/14584656

[38] S.Kirchheimer. (2017). Herpes Virus Linked to Cervical Cancer. [Online]. Available: https://www.webmd.com/genital-herpes/news/20021105/herpes-virus-linked-to-cervical-cancer#1

[39] Wikipedia. (2017, Sep.). Molluscum contagiosum. [Online]. Available: https://en.wikipedia.org/wiki/Molluscum_contagiosum

[40] M.Hoffman. (2017). Picture of the Vagina. [Online]. Available: https://www.webmd.com/women/picture-of-the-vagina#1

[41] Wikipedia. (2017, Oct.). Decision tree learning. [Online]. Available: https://en.wikipedia.org/wiki/Decision_tree_learning

[42] Wikipedia. (2018, Jan.). Apriori Algorithm. [Online]. Available: https://en.wikipedia.org/wiki/Apriori_algorithm

[43] S. Asaduzzaman et al. (2015, Nov.). Anticipation of the Significance of Risk Factors in Cervical Cancer for Low Incoming Country: Bangladesh Perspective. [Online]. Available: https://www.ijser.org/researchpaper/Anticipation-of-the-Significance-of-Risk-Factors-in-Cervical-Cancer-for-Low-Incoming-Country-Bangladesh-Perspective.pdf

[44] Cervical Cancer. (2017, Oct.). Estimated Incidence, Mortality and Prevalence Worldwide in 2012. [Online]. Available: http://globocan.iarc.fr/old/FactSheets/cancers/cervix-new.asp

[45] Mining Frequent Itemsets – Apriori Algorithm. [Online]. Available: http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab8-Apriori.pdf

[46] Big data made simple. (2015, Apr.). What is clustering in data mining? [Online]. Available: http://bigdata-madesimple.com/what-is-clustering-in-data-mining/

[47] K-means clustering. [Online]. Available: http://www.saedsayad.com/clustering_kmeans.htm

[48] S. Polamuri. (2016, Sep.). Classification and clustering algorithms. [Online]. Available: http://dataaspirant.com/2016/09/24/classification-clustering-alogrithms/

[49] K. Thangavel. (2006). Data mining approach to cervical cancer patients analysis using clustering technique. [Online]. Available: http://docsdrive.com/pdfs/medwelljournals/ajit/2006/413-417.pdf

[50] C. Chang et al. (2013, Feb.). Prediction of recurrence in patients with cervical cancer using MARS and classification. [Online]. Available: http://www.ijmlc.org/papers/276-LC010.pdf

[51] P. Ramachandran et al. (2014, Jul.). Early detection and prevention of cancer using data mining techniques. [Online]. Available: https://pdfs.semanticscholar.org/5e7c/60bc801d87000e7bd271a4d816e4e2fc482f.pdf

[52] S. Shajahaan et al. (2013, Nov.). Application of data mining techniques to model breast cancer data. [Online]. Available: https://pdfs.semanticscholar.org/85f6/a72dfc83ee65597c844c779bd67ccac36f84.pdf

[53] R. Vidya et al. (2015, Mar.). A pioneering cervical cancer prediction prototype in medical data mining using clustering pattern. [Online]. Available: https://www.scribd.com/document/261697578/A-Pioneering-Cervical-Cancer-Prediction-Prototype-in-Medical-Data-Mining-using-Clustering-Pattern