Application of Data Mining Techniques for Avoiding Underestimation of an Event

Avijit Kumar Chaudhuri
c.avijit@gmail.com

Dr. Deepankar Sinha
dsinha2000@gmail.com

Dr. Anirban Das
anirban-das@live.com

Dr. Dilip K. Banerjee
dkbanrg@gmail.com

***Abstract*: Medical records comprise varied data types; artificial intelligence and data-mining methods (DMTs) are useful to draw insights and patterns. Several scholars claim that there is no universal way of addressing diagnosis issues, and a mixed model is desirable to resolve these concerns. In this paper, the authors compare the proven approaches and propose a framework to integrate the findings from various techniques to evade Type 2 and Type 1 errors. The dataset chosen for this purpose includes medical data on HPV disease. Two sets of dataset – disease and treatment dataset and features found significant from ensemble method – the random forest were used and to predict the disease. The results show that traditional methods such as Logistic Regression(LR) performed better with features found significant using Random Forest(RF). However, this approach fails when the dichotomy of data (i.e., disease or no disease) is not distinct. Decision Tree(DT) analysis shows consistent performance across all variants of the dataset chosen in this paper. The paper suggests an amalgamation of association rules and a prediction approach (with or without integration) that provides higher accuracy.**

Keywords: Data mining techniques; cervical cancer; Type 1 and Type 2 errors; integrated-approach; under estimation

I. INTRODUCTION

Artificial intelligence and data mining techniques (DMTs) have often been used to extract useful information from massive data sets and patterns (Liao et al., 2012 [1]. There is evidence of widespread use of DMT in the diagnosis of diseases. However, no single method has shown consistent outcomes, and thus, researchers have proposed a hybrid approach (Jothi & Husain, 2015) [2]. In most cases, the results suffer from the over-fitting or under-fitting of data affecting predictions. Incorrect treatments are irreversible and have long term impacts. There are instances of missed treatment (Type 1 error) or treating the wrong ones (Type 1 error). The individual methods showed varying levels of accuracy, and diagnosis with the highest accuracy levels, say 75% or so, is used for suggesting treatments.

In some cases, the accuracy can be around 90% or more. Thus, the question arises as to how can the prediction levels be enhanced with existing methods? The authors aim to ensemble DMTs in this paper to explore the possibility of increasing the accuracy of the prediction. Two databases - patient and treatment database were used to identify the occurrence of cervical cancer. Three supervised learning methods - decision tree (DT), random forest (RF), and logistics regression, have been used to identify the importance of variables. The prediction was carried using original datasets, and revised datasets comprising variables found significant. A comparison of the outcomes across two datasets was made to conclude the cause of the disease.

Several kinds of literature (detailed in the next section) on cervical cancer point out that HPV OF certain types leads to warts in patients, but all warts do not show signs of cancer either in the short term or in the long run. This uncertainty is so because initial detection of symptoms (such as the formation of warts) does not mark cancer in a patient, as this is temporal. Warts so observed may, after some time, turn out to be carcinoid. There are instances where warts prevalent in a human body even after six months or so may not require cancer treatment, whereas a 3-month old wart may turn out to be cancerous. Thus, there are possibilities of incorrect estimation of HPV in many cases due to limitations in the study of this disease. In such cases, analysis requires mining the patients' data set (with warts) diagnosed with or without cancer and patient (with warts) who underwent treatment. Unfortunately, there is a lack of a centralized database to arrive at the right diagnosis. Different databases, one about patients detected with warts and other - patient (with warts) undergoing treatment, are available. There are few attempts to derive association rules combining patient characteristics and treatment response for firming up the right diagnostics.

In this article, diagnosis of information available in the World-Wide-Web was carried out to develop the information hierarchy (of cervical cancer) to facilitate analysis. In this paper, the authors aim to combine data mining methods to explore the possibility of increasing the accuracy of the forecast.

The paper has been organized into seven sections. The next section gives a brief overview of the disease - cervical cancer and the Human papillomavirus virus (HPV); section 4 discusses the different DMTs used in this study; section describes the dataset, and section 6 analyses the data. Section 7 discusses the results of the analyses, and section 8 concludes the research work.

II. HUMAN PAPILLOMA VIRUS (HPV)

A. General Description

The human papilloma virus (HPV) is the most widely recognized sexually transmitted viral infection. Sexually dynamic individuals are most prone to this infection(Koutsky, 1997; Gabbey & Jacquelyn, 2017) [3-4]. The disease's primary issue is the time gap between the actual time of infection and the time the virus gets manifested, following which the treatment begins. This time gap is because warts that develop due to HPV infection has no signs or side effects in the infected person. As a result, the disease unconsciously passes on to their sexual accomplices. In the long run, the infection causes cervical cancer.

Developing countries witness women's maximum death on account of cervical cancer while worldwide, it ranks second as a terminal disease. Study shows that sexual transmission of human papilloma virus (HPV) causes cervical Intra-epithelial neoplasia and invasive cervical cancer(Gabbey & Jacquelyn, 2017) [4]. Every year 510,00 new cases and

288 000 deaths are reported worldwide (Shouman et al., 2012) [5]. The uniqueness lies in the fact this particular malignancy is detected at the productive age of women. The disease gets detected at an average age 30 to 34 years, and the tops at the age of 55 to 65. The mid-age being 38. On average, sexually dynamic women get infected with genital HPV by 50 years of age (Sankaranarayanan & Ferlay, 2006) [6].

India accounts for around one-third of global deaths due to cervical cancer(http://www.who.int/hpvcentre) [7] against 6.6% of global infection. The primary types include serotypes 16 and 18, accounting for 76.6% of cases. The women developing warts account for around 2 – 25% of sexually communicated diseases (http://www.who.int/hpvcentre) [7].

Cervical cancer is detected using the Pap test followed by colposcopy test. The serious problem lies in the fact that there is no recommended conventional treatment for HPV infection. Doctors resort to wait-watch on the symptoms to convert to the pre-cancerous stage, i.e., observe for wart and or pre-cancerous changes of the cervix. In this paper, the authors propose to apply data mining tools to derive meaningful insights for the treatment of cervical cancer.

B. Cervical cancer – Profile Analysis

Profile analysis serves as a convenient way to represent information and has been used by researchers to describe information about groups and families of sequences. Some authors(Gribskov et al., 1987) [8] used this approach for the study of proteins. HPV virus-infected persons develop warts in the throat or genitals and can cause cancer in these regions or in the head and neck. Warts appear benign in the initial stage but later manifest to be malign, and the significant problem is that cancer is in a later phase of development at this stage. Patients who undergo periodic Pap tests may have early detection of the disease. These findings can improve perspective and increase chances of survival(Gabbey & Jacquelyn, 2017) [4].

HPV types 16 and 18 are the cause of about 70% of all cervical cancer cases worldwide. HPV antibodies that forestall HPV 16 and 18 infections are currently accessible and can decrease the incidence of cervical and other anogenital cancers.

A report citing the different perspective of this disease provides HPV related insights; and mentions factors adding to cervical cancer; cervical cancer screening rehearses; HPV antibody presentation; and other significant vaccination pointers. Study shows underestimations of the prevalence of HPV in many cases due to limitations in methods of research of this disease (Schmitt et al., 2010). The chance of this infection increases with the severity of the lesion. Cervical lesions between 41% and 67% of high grade and 16%-32% of low grade contribute to 70% of cases, namely HPV 16 and 18(Shouman et al., 2012) [5]. 20% of other cervical cancer types include the HPV types – 31, 33, 35, 45, 52, and 58.

The most challenging aspect of this infection is that in many cases, HPV goes away on its own, so there is no treatment required. Instead, the doctor would preferably want the patient to go for repeated testing on a half-yearly or yearly basis to check for the infection's persistence. In such a case, six months may be too late for the treatment.

The US Food and Drug Administration (FDA) accepted the principal DNA test for HPV in 2014. Updated rules suggest that women have their first Pap test, or Pap smear, at age 21 and be tested for HPV simultaneously, paying little mind to the beginning of sexual activity(Gabbey & Jacquelyn, 2017) [4]. After that, women between the ages of 21 to 29 ought to have a Pap test at regular intervals. Standard Pap tests help recognize anomalies in cell structure, which serves as the caution against cancer growth or any such severity.

The doctors follow the general convention of screening women in the age group of 30 to 65 at an interval of five years based on Pap and HPV diagnosis. For age less than 30 years, such tests are prescribed if the smear test shows inconsistent results. For example, if patients have any of the 15 HPV strains, the likelihood of cancer is higher. The frequency of screening, in such cases, increases as in many cases, it takes ten years to get malignant. Such FDA affirmed tests for men are not present(Gabbey & Jacquelyn, 2017) [4].

The periodicity of the test, so far, recommended may have high variation depending on a host of factors. The availability of patient data and the application of data mining techniques can lead to a more precise recommendation of the test's periodicity and predict the probability of cancer occurrence from warts detected.

C. Web Information Diagnosis

The authors carried out a diagnosis of information available in the World-Wide-Web using Sem-Rush software. The result highlights two aspects – age and HPV type 16; besides culls out various variants associated with age, types, cancer, warts, gender, sex, and related facets.

Information hierarchy of cervical cancer

1) Information hierarchy of cervical cancer

Figure 1 shows the author's ontology of cervical cancer from the study of literature and web-information analysis.

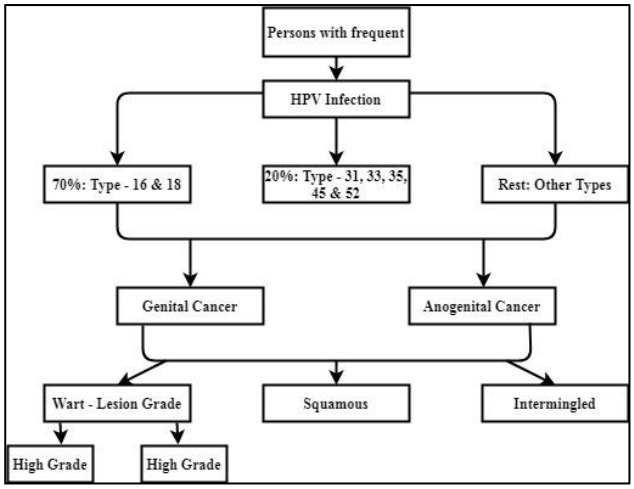

Fig. 1. Profile of Cervical Cancer

Therefore, the research question is how to minimize type 1 and type 2 errors in cervical cancer treatment. In other words, the sub-questions are:

i. What are the major causes of this disease?

ii. When to begin treatment after detection of warts or abnormality in Pap test?

iii. What is the probability of correct diagnosis given the age, gender, the time elapsed before treatment (month), the

number of warts, types of the wart (Count), Surface area of warts(mm2), and other factors?

D. Data mining in the analysis of diseases

1. Cervical cancer mostly affects women, with 570,000 new cases in 2018. The majority of these cases are from developing countries. The detection of cancer at the right time; effective screening and treatment programs can make cancer patients' lives better(https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/) [9]. The limited availability of resources screening and diagnosis from a Computer-Aided Diagnosis (CAD) point of view becomes difficult. Also, sometimes, patients do not participate in routine testings. Thus the correct, timely diagnosis and estimation of specific risks of each patient are an assurance of proper treatment. The doctor's experience and abstract choice affect the results in many of these screening techniques(Singh, 2005) [10]. A survey could be applied to patients to decide the riskiest groups and to reduce pointless screenings. Hence the patients can resort to testing based on the likelihood of cancer.

There are several references to different DMT use to discover numerous investigations about the assessment of some therapeutic screenings. Sen & Das(2013) [11] in their paper, have introduced an artificial neural network in pancreatic ailment diagnosis dependent on a set of symptoms. The authors found that detection using neural networks had higher accuracy than other manual methods(Gabbey & Jacquelyn, 2017) [4]. Fernandes et al. displayed a regularization-based TL way to exchange the contribution type for each component on direct models. Author, namely

Kalyankar & Chopde(2013) [12] demonstrated the use of Genetic Algorithms, Artificial Neural Network, Hierarchical Clustering, Neuro-Fuzzy framework, Raman Spectra. To arrive at an overview of various kinds of cancer, for example, skin cancer, bosom cancer, pancreatic cancer, In this study, classification performance varied in the range of 80.5% and 95.8% depending upon the cancer type and grouping. In 2016, Kanimozhi & Karthikeyan(2016) [13] showed the use of mining procedures to investigate the various coronary illness. The authors utilized different databases and found that the achievement rates have fluctuated between 45% to 99.1% contingent upon arrangement strategies and traits in the different databases. Fatima & Pasha(2017) [14] exhibited a similar examination of various machine learning calculations for diagnosing various maladies, for example, coronary illness, diabetes infection, liver sickness, dengue ailment, and hepatitis ailment. They demonstrated that different techniques showed different accuracy levels for a particular disease, while the best method for a specific ailment did not perform equally for a separate category of illness or a database. Sankaranarayanan & Ferlay, 2006 [6] showed that Naive Bayes had an accuracy of 97% while Neural Network demonstrated 70 percent achievement rates. Thus, the diagnosis based on the best result may lead to both type 1 and 2 errors. This issue gets resolved by integrating the techniques and identifying some association rules derived from the disjoint set of the DMT. The researchers have so far not used this approach in diagnosing diseases, especially cervical cancer.

III. DATA MINING MODELS

This study includes K-means clustering, an unsupervised model and supervised machine learing techniques – Decision tree (DT), Random forest (RF) and Logistics regression (LR).

A. Kmeans

K-means is an unsupervised machine learning method of clustering. This algorithm's essential impact is its lucidity and speed, which enables it to keep running on huge datasets. With an enormous number of factors, K-Means might be computationally quicker than various hierarchical clustering (when K is less).

B. Random forest (RF)

This method falls under supervised machine learning algorithms, aiming to minimize overfitting and reduce variance (Breiman, 2001) [15]. In this method, the data set is divided into subsets formed by random sampling with replacement. This method of sampling is also referred to as bootstrapping. A decision tree algorithm is run on each sample, and outcomes are noted. RF follows the principles of bagging. In this principle, each sample's outcome is noted, and the best result is proposed by the method of voting. The importance of variables is determined using an out-of-bag approach. The relationship between variables identified through different samples is tested using a test set comprising records left out in training.

The prediction accuracy depends on the way samples are selected randomly, that is, the kind of randomness to maximize accuracy. The creation of trees is to be done to minimize prediction errors (Chen & Ishwaran, 2012) [16].

C. Decision trees(DT)

Authors propose the DT technique among the other types of Data mining techniques because of the following criteria:

- DT filtration techniques are easy to implement and easily understandable.
- It is a systematic and widely used data mining method.
- In data mining, DT demonstrates a very large-scale achievement in comparison with other techniques.
- Tree-like models are used to make decisions in this decision support system.
- DT techniques can be treated as the proven mechanism in knowledge discovery fields.
- DT classifiers are the most used in data and text mining, machine learning, information extraction, and pattern recognition.
- Input data such as nominal, numeric, and text can be handled by this method. The processing of datasets with missing values is also possible.

Thus DT is a supervised classification function that can be used for accomplishing the value of a dependent factor given the values of the independent variables (Shouman, Turner & Stocker, 2012) [5]. This method is not affected by linearity in data, missing values, types of data, and outliers. The results are easy to understand and interpret.

D. Logistic- Regression (LR)

LR enables prediction of a disease outcome , a dependant variable, using independent variables of multiple types. This method is suitable where outcome is a categorical variable

and the finds its place in medical research since (Hall & Round, 1994) [17].

IV. DATA SETS

Table 1 depicts the attributes in the dataset obtained from the University of California, Irvine – the data related to patients in Hospital Universitario de Caracas in Caracas, Venezuela. It captures 35 characteristics of cervical cancer of 858 patients (UCI, 2019). The attributes comprise information associated with demography, habits, and historical medical records (UCI, 2019).

The second dataset relates to the treatment of patients with common wart types, sourced from dermatology clinic of Ghaem Hospital in Mashhad during the period January 2013 to February 2015. Table 2 presents the dataset that captures eight features of patients who underwent treatment using the immunotherapy method. The class attribute in these datasets is the Response to Treatment feature.

TABLE I. ATTRIBUTE INFORMATION OF FIRST DATASET

Feature	Type	Feature	Type
Age	Int	STDs:pelvic inflammatory disease	bool
# of partners	Int	STDs:genital herpes	bool
Age of 1st intercourse	Int	STDs:molluscumcontagiosum	bool
# of pregnancies	Int	STDs:AIDS	bool
Smokes	bool	STDs:HIV	bool
Smokes years	Int	STDs:Hepatitis B	bool
Smokes packs/year	Int	STDs:HPV	bool
Hormonal Contraceptives	bool	STDs: Number of diagnosis	Int
Hormonal Contraceptives years	Int	STDs: Time since first diagnosis	Int
IUD	bool	STDs: Time since last diagnosis	Int
IUD years	Int	Dx:Cancer	bool
STDs	bool	Dx:CIN	bool
STDs number	Int	Dx:HPV	bool
STDs:condylomatosis	bool	Dx	bool
STDs:cervicalcondylomatosis	bool	Hinselmann: target variable	bool
STDs:vaginalcondylomatosis	bool	Schiller: target variable	bool
STDs:vulvo-perinealcondylomatosis	bool	Cytology: target variable	bool
STDs:syphilis	bool	Biopsy: class or target variable	bool

TABLE II. ATTRIBUTE INFORMATION OF THE SECOND DATASET

Feature name	Values	Mean ± SD
Response to treatment	Yes or No	
Gender	41 Man, 49 Woman	
Age (year)	15–56	31.04 ± 12.23
Time elapsed before treatment (month)	0–12	7.23 ± 3.10
The number of warts	1–19	6.14 ± 4.2
Types of the wart (Count)	1– Common (47), 2– Plantar (22), 3– Both (21)	
Surface area of the warts(mm2)	4–750	85.83 131.73

TABLE III. K-MEANS ANALYSIS OF DATASET RELATED TO TREATMENT

	Cluster				
	1	2	3	4	5
Sex	2	1	2	2	1
Age	32	15	26	32	35
Time	6.25	3.00	8.29	7.17	9.25
Number_of_Warts	8	2	4	6	4
Type	2	3	1	2	1
Area	350	900	177	49	504
induration_diameter	6	70	11	15	6
Result_of_Treatment	1	1	1	1	1

V. VARIABLE CLUSTERING AND IMPORTANCE

The authors used K-meansfor identifying the clusters. They used DT and RF techniquesto identify importance of variables suggested by these methods respectively. To predict the results, the authors used Logistics-regression (LR) on the original and revised data set and compared the results to see the change in prediction accuracy. The revised set consisted of data on important variables obtained from RF and DT analysis respectively.

A. K – means Clustering

The Disease Data Set Analysis (Table XI. – Appendix A):

Analysis of the disease dataset using this method indicated that factors, namely, patient's age, their age of first sexual intercourse, smoking habits, the period of use of contraceptives, prevalence of STD, and the time since STD was first diagnosed and last indications are the crucial factors.

The Treatment Data Set Analysis: The K-means cluster analysis (Table 3) of dataset related to the treatment of patients indicate that the cluster where 76 out of 90 cases clustered indicate that number of warts (6) in women patients of age 32 years showed signs of cancer in around seven months. However, there are four cases where cancer was detected in approximately six months, even though the number of warts is low (2 numbers). There is one case where cancer got detected in three months after detection of warts. Table 4. shows the number of cases in each cluster.

B. Random Forest (RF) Analysis

RF analysis on original desease dataset shows Age, Number of sexual partners, Hormonal Contraceptives (years), IUD (years), STDs: Time since the first diagnosis, Dx:CIN and Dx:HPVas relatively important variables.

TABLE IV. NUMBER OF CASES IN EACH CLUSTER

Cluster	1	4.000
	2	1.000
	3	6.000
	4	76.000
	5	3.000
Valid		90.000
Missing		.000

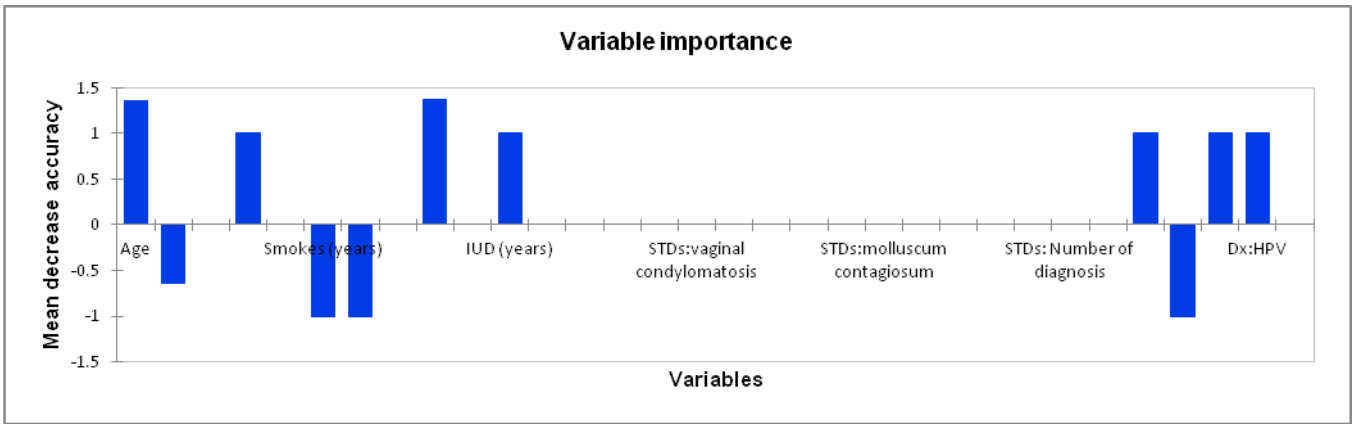

Fig. 2. RF analysis of the data on disease dataset

Similar exercise on treatment dataset predicts Time, age, Number_of_Warts, and Area as relatively significant ones.

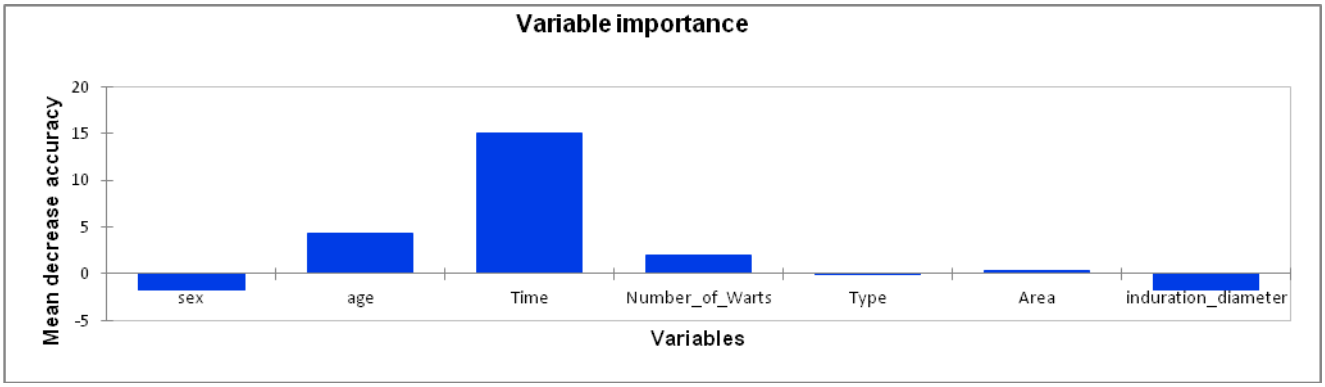

Fig. 3. RF analysis of the data on treatment dataset

C. Decision Tree

Disease dataset

The results from DT analysis of disease dataset is given in figure.4.

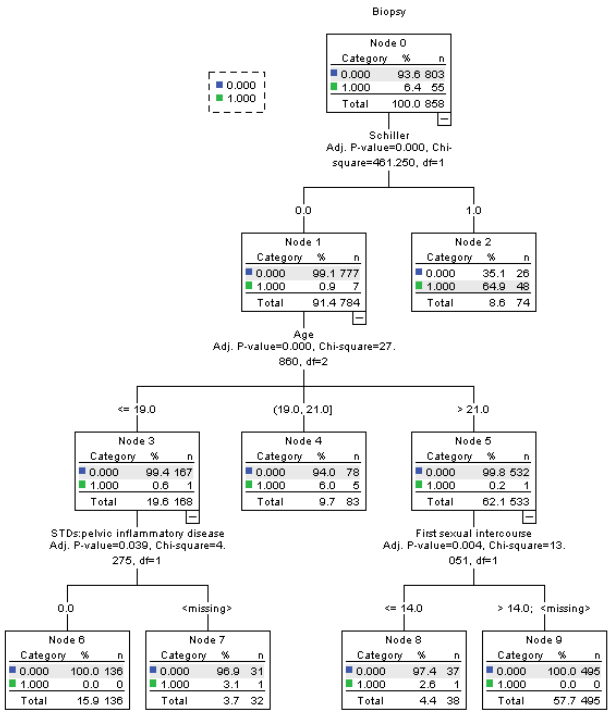

Fig. 4. Outcome of the DT analysis carried out on all the variables of disease dataset.

The results suggests the association rules as:

Schiller's test result is the significant factor for detection of cervical cancer.

Schiller's test result is the significant factor for detection of cervical cancer.
For the Schiller's test result = 1, 64.9 % of women have cervical cancer.
For the Schiller's test result = 0, next best predictor is age
If age > 19 and <= 21 then, 6 % of women have cervical cancer
For the age <= 19, next best predictor is STDs : pelvic inflammatory disease
For STDs : pelvic = 0.0, 0.0 % women are attacked with cervical cancer
For STDs : pelvic = missing, 3.1% of women have cervical cancer
For the age > 21, next best predictor is First sexual intercourse
For First sexual intercourse <= 14, 2.6% of women have cervical cancer
For First sexual intercourse > 14, 0.0% of women have cervical cancer

The accuracy of findings is to the tune of 96.2% as shown in Table 5

TABLE V. ACCURACY LEVEL OF DT ANALYSIS.

Classification			
	Predicted		
Observed	0	1	% Correct
0	777	26	96.8%
1	7	48	87.3%
Overall Percentage	91.4%	8.6%	96.2%
Growing Method: CHAID			
Dependent Variable: Biopsy			

DT analysis shows that variables Schiller Test, STDs:pelvic inflammatory disease and First sexual intercourse) are important and differ results of RF analysis excepting Age being common in both the findings.

Exercise will provide relative accuracy when further extended to the revised data set consisting only of the variables identified as important from the RF classification.

Treatment dataset

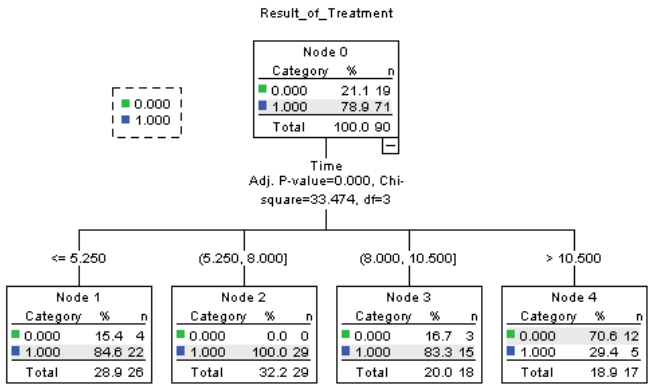

Fig. 5. Outcome of the DT analysis carried out on all the variables of treatment dataset

The DT with all the variables of treatment dataset predicts the association rules as follows

The time factor is the only and best factor for the treatment of warts.

For the time <= 5.250, 84.6 % of women got treated.
Time > 5.250 and <= 8.000 then, 100% of women were treated
If Time > 8.000 and <= 10.500 then, 83.3% of women were treated
For the time > 10.500, 29.4 % of women got treated.

Table 6 shows the extent of accuracy. The DT analysis on all variables predicts the accuracy of 86.7%.

DT application on treatment dataset shows that time is the only important variable and thus differ from results ofRF analysis that predicts time, age, number of warts and area as significant. Time is common finding from both the analysis.

The excercise when further extended on revised treatment-dataset comprising only variables found important from RF test will provide the relative accuracy levels.

DT Analysis of Relatively Important Variables as determined by RF Analysis

Disease Data set

The DT analysis of variables identified as relatively significant using the RF analysis shows that Dx: HPV is the critical factor for the disease (Figure 6).The approach predicted with 93.6% accuracy (Table 7). However, it failed to predict the occurrence of illness using this approach. The cause being the lesser number of data points (6.4%) relating to patients with biopsy equal to 1. This method correctly predicted the instances of HPV patients who did not have cancer, i.e., with no type 2 error. This approach proved inferior to the results of DT analysis considering all variables in the data set.

TABLE VI. PREDICTION ACCURACY

Classification			
Observed	Predicted		
	0	1	% Correct
0	12	7	63.2%
1	5	66	93.0%
Overall Percentage	18.9%	81.1%	86.7%
Growing Method: CHAID			
Dependent Variable: Result_of_Treatment			

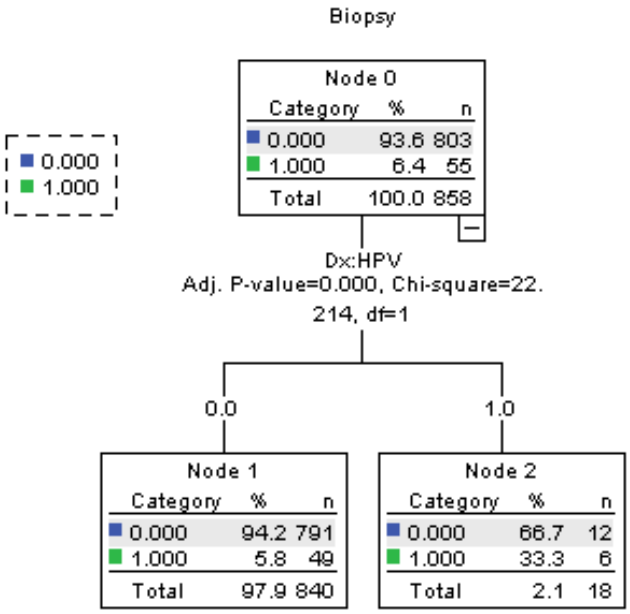

Fig. 6. Outcome of the DT analysis carried out on all the relatively important variables from disease data set

TABLE VII. PREDICTION ACCURACY

Classification			
Observed	Predicted		
	0	1	% Correct
0	803	0	100.0%
1	55	0	0.0%
Overall Percentage	100.0%	0.0%	93.6%
Growing Method: CHAID			
Dependent Variable: Biopsy			

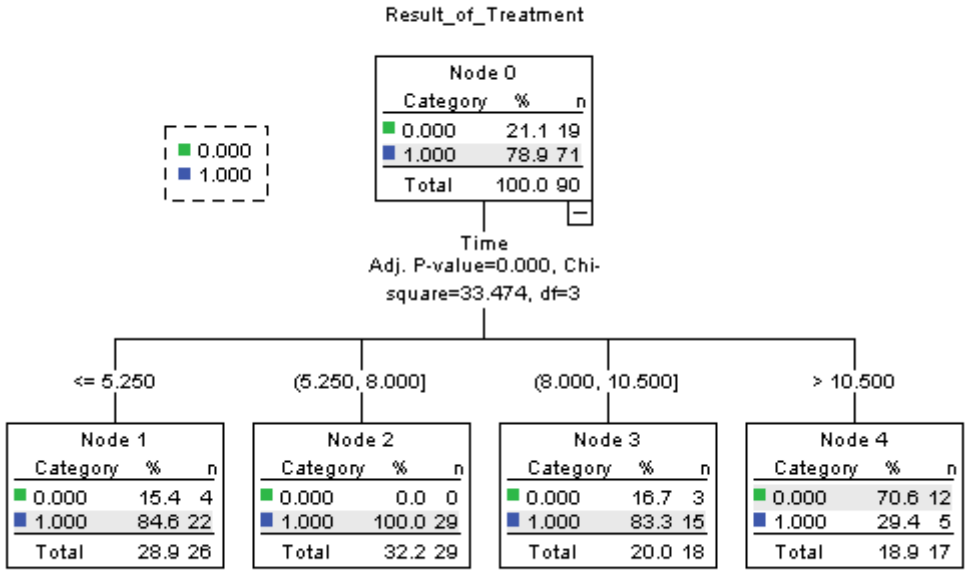

Fig. 7. Outcome of the DT analysis carried out on all the relatively important variables from the treatment data set

Treatment Data set

The application of DT on this dataset indicated time (since first detected) as the most significant factor (Figure 7). The approach predicted with 86.7% accuracy (Table 8). The occurrence of cancer had a prediction accuracy of 93% as against 63.2% accuracy of HPV patients without cancer (biopsy=0). Thus the type 2 error was found to be higher than the type 1 error. The results from this approach did not differ from the DT analysis carried out, taking all variable in the treatment data set. In this data set the two categories of classification comprised data records in the ratio of 79:21 (Biopsy = 1: Biopsy = 0) percentage.

D. Logistics Regression

Disease Data set

The Logistics regression on variables of disease dataset shows that none of the variables are significant in predicting the disease.

Treatment dataset

The Logistics regression on variables of disease dataset shows that Time is the significant variable along with the constant value. With Time as a significant variable, Equation 1 enables the prediction of the disease.

$$p = \frac{1}{1 + e^{-(5.151 - 0.353 \times \text{Time})}} \quad (1)$$

The results showed an accuracy of 85.6% (Table 9).

LR classification of revised dataset comprising important-variables determined from DT classification

In this approach, as well, none of the variables were found significant. The results obtained from DT analysis (on the complete data set) proved to be the best so far with an accuracy level of 96.2%

Revised Dataset (comprising variables found crucial from RF analysis)

This exercise also revealed time as the significant factor and the prediction accuracy was higherthan the other approaches – table 10.

TABLE VIII. PREDICTION ACCURACY

Classification			
Observed	Predicted		
	0	1	% Correct
0	12	7	63.2%
1	5	66	93.0%
Overall Percentage	18.9%	81.1%	86.7%
Growing Method: CHAID			
Dependent Variable: Result_of_Treatment			

TABLE IX. CLASSIFICATION TABLE

Classification Table[a]

Observed			Predicted		
			Result_of_Treatment		% Correct
			0	1	
Step 1	Result_of_Treatment	0	6	13	31.6
		1	0	71	100.0
	Overall Percentage				85.6

a. The cut value is .500

TABLE X. CLASSIFICATION TABLE

Observed			Predicted		
			Result_of_Treatment		Percentage Correct
			0	1	
Step 1	Result_of_Treatment	0	9	10	47.4
		1	0	71	100.0
	Overall Percentage				88.9

a. The cut value is .500

The accuracy level of the LR on RF determined variables shows a superior result, for predicting the treatment, compared to all approaches discussed so far.

VI. RESULTS AND DISCUSSIONS

The cause of the disease was determined using the disease dataset. The common factors across the different approaches include the patient's age, assuming they had multiple sex partners and used contraceptives, and suffered from sexually transmitted diseases. K-means clustering results show that age equal to 24 years with first sexual interaction at the lowest age of 16 caused cervical cancer.

The treatment dataset indicates the treatment attributes once the warts are detected. Here, time is a critical factor. K-means clustering results show that most likely time equal to around seven months with an age of 32 years showed signs of cancer. The minimum time of occurrence of cervical cancer stands out as three months post detection of warts. There are four cases where cancer got detected after six months. Thus the time range of 3 to 7 months is most crucial.

The accuracy of the prediction of the disease appeared to be highest with the approach, namely Decision-Tree. The probability of type 1 and 2 errors was 12.7% and 3.2%, respectively. The association rule associated with this approach is:

Schiller's test result factor is the best factor for the detection of cervical cancer.

For the Schiller's test result = 1, 64.9 % of women have cervical cancer.
For the Schiller's test result = 0, the next best predictor is age.
If age > 19 and <= 21 then, 6 % of women would suffer from cervical cancer
For the age <= 19, the next best predictor is STDs : a pelvic inflammatory disease.
For STDs : pelvic = 0.0, 0.0 % women suffer from cervical cancer
For STDs : pelvic = missing, 3.1 of women suffer from cervical cancer
For the age > 21, the next best predictor is First sexual intercourse.
For First sexual intercourse <= 14, 2.6% of women have cervical cancer.
For First sexual intercourse > 14, 0.0% of women have cervical cancer.

The accuracy of predicting the treatment of the disease appeared to be highest with approaches, namely Logistics-Regression and DT. The DT approach yielded a 7% type 1 error. The association rule associated with these two approaches are:

For the time <= 5.250, 84.6 % of women get treated.
If time > 5.250 and <= 8.000 then, 100% of women get treated
If time > 8.000 and <= 10.500 then, 83.3% of women get treated
For the time > 10.500, 29.4 % of women get treated.

DT analysis showed consistency in performance, while Random-forest performed better in a data set with a distinct dichotomy. In the case of disease data, the ratio of biopsy=1 and biopsy=0 was 94:6. Hence, the results from the RF-based Decision-Tree were not significant. Whereas in the treatment dataset, the ratio was 80:20, the results reflected higher prediction accuracy. The accuracy level of the LR on RF determined variables showed a superior result (96.2%) among all methods for predicting the treatment.

Hence, the alternate hypothesis – "integration of data mining approaches lead to better prediction" stands accepted for a dataset with dichotomy at the ratio of 80: 20 or better. The significant variables when the disease is detected include – Age, schiller – STD parameters. The critical factor for successful treatment is time. The critical values of these factors are- Age: 19 years and above,; Time: 3 to 7 months; Schiller Test = 1; STD test (of different parameters) = 1.

VII. CONCLUSION

Data mining techniques are suitable for evaluating medical records, but the outcomes depend on the nature of the dataset - that is, if the outcome is categorical, then the number of records for each category impacts accuracy. The decision tree results may be better than enhanced ensembled methods such as random forest (RF). The importance of variables is

required to identify the cause and focus on the treatment procedures. Results may vary between the original dataset and the revised one that comprises records of the essential variables. This paper showed that the accuracy level of the LR on RF determined variables showed a superior result (96.2%) among all methods for predicting the treatment.

Unsupervised methods also aid in diagnosing along with supervised learning methods. In this paper, the K-means cluster revealed significant findings - at age equal to 24 years with first sexual interaction at the lowest age of 16 caused cervical cancer. Otherwise, age greater than 19 years with Schiller test and STD results being positive causes disease. Time was found to be the critical factor for the start of treatment once the warts are noticed. Warts show malignancy during 3 to 7 months, where most cases showed the signs after five months.

REFERENCES

[1] Liao, S. H., Chu, P. H., & Hsiao, P. Y., 2012. Data mining techniques and applications–A decade review from 2000 to 2011. Expert systems with applications, 39(12), 11303-11311.
[2] Jothi, N., & Husain, W., 2015. Data mining in healthcare–a review. Procedia computer science, 72, 306-313.
[3] Koutsky, L., 1997. Epidemiology of genital human papillomavirus infection. The American journal of medicine, 102(5), 3-8.
[4] Gabbey. A. E., Jacquelyn. C. Human Papillomavirus Infection Medically reviewed by Debra Rose Wilson, PhD, MSN, RN, IBCLC, AHN-BC, CHT, 2017.
[5] Shouman, M., Turner, T., & Stocker, R., 2012, March. Using data mining techniques in heart disease diagnosis and treatment. In 2012 Japan-Egypt Conference on Electronics, Communications and Computers (pp. 173-177). IEEE.
[6] Sankaranarayanan, R., & Ferlay, J., 2006. Worldwide burden of gynaecological cancer: the size of the problem. Best practice & research Clinical obstetrics & gynaecology, 20(2), 207-225.
[7] WHO 2007 WHO/ICO Information Centre on HPV and Cervical Cancer (HPV Information Centre). Summary report on HPV and cervical cancer statistics in India 2007. [Last Assessed on 2008 May 1]. Available from: http://www.who.int/hpvcentre .
[8] Gribskov, M., McLachlan, A. D., & Eisenberg, D., 1987. Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences, 84(13), 4355-4358.
[9] https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/
[10] Singh, N., 2005. HPV and Cervical cancer-prospects for prevention through vaccination. Indian J Med Paediatr Oncol, 26(1), 20-23.
[11] Sen, T., & Das, S., 2013. An approach to pancreatic cancer detection using artificial neural network. In Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering-CEEE (pp. 56-60).
[12] Kalyankar, M. A., & Chopde, N. R., 2013. Cancer Detection: Survey. Int. Journal of Advanced Research in Computer Science and Software Engineering, 3(11), 1536-1539.
[13] Kanimozhi, V. A., & Karthikeyan, T., 2016. A Survey on Machine Learning Algorithms in Data Mining for Prediction of Heart Disease. International Journal of Advanced Research in Computer and Communication Engineering, 5(4), 2278-1021.
[14] Fatima, M., & Pasha, M., 2017. Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.
[15] Breiman, L., 2001. Random forests. Machine learning, 45(1), 5-32.
[16] Chen, X., & Ishwaran, H., 2012. Random forests for genomic data analysis. Genomics, 99(6), 323-329.
[17] Hall, G. H., & Round, A. P., 1994. Logistic regression—explanation and use. Journal of the Royal College of Physicians of London, 28(3), 242.

APPENDIX I

TABLE XI. ATTRIBUTE INFORMATION OF SECOND DATASET(WITH 25 CLUSTERS)

Final Cluster Centers																									
	Cluster																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Age	41	36	24	35	35	49	30	31	30	28	29	18	29	28	26	36	28	38	19	27	20	36	33	33	42
Number of sexual partners	3	1	1	3	3	2	5	2	3	4	3	7	3	3	3	3	3	2	2	4	2	1	3	4	3
First sexual intercourse	17	22	20	20	17	15	16	20	17	14	19	16	15	16	14	19	15	20	15	17	18	28	16	17	19
Num of pregnancies	4	4	1	2	6	6	4	2	3	4	2	1	3	3	3	3	6	2	2	3	1	1	4	0	3
Smokes	0	1	0	0	1	0	0	1	0	0	0	1	0	1	1	0	1	0	0	0	1	1	1	0	0
Smokes (years)	0	16	0	0	13	0	0	9	0	0	0	5	0	12	7	0	14	0	0	0	13	16	14	0	0
Smokes (packs/year)	0	5	0	0	3	0	0	5	0	0	0	5	0	6	1	0	2	0	0	0	7	2	1	0	0
Hormonal Contraceptives	1	0	1	0	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	1	1
Hormonal Contraceptives (years)	10	0	4	0	7	2	0	6	11	0	3	2	0	7	2	7	7	0	0	1	0	0	1	1	3
IUD	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1
IUD (years)	0	0	0	10	0	0	7	0	1	0	4	0	0	0	0	0	4	0	0	0	0	0	0	0	3
STDs	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
STDs (number)	1	3	2	2	1	1	1	2	2	1	2	2	1	1	3	2	2	1	2	2	2	3	1	1	2
STDs:condylomatosis	0	1	1	1	0	0	0	1	1	0	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1
STDs:cervicalcondylomatosis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STDs:vaginalcondylomatosis	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
STDs:vulvo-perinealcondylomatosis	0	1	1	1	0	0	0	1	1	0	1	1	0	0	1	1	1	0	1	0	0	1	0	0	1
STDs:syphilis	1	0	0	0	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
STDs:pelvic inflammatory disease	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
STDs:genital herpes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STDs:molluscumcontagiosum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STDs:AIDS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
STDs:HIV	0	0	0	0	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0	0	1	0	1	0	0
STDs:Hepatitis B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
STDs:HPV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
STDs: Number of diagnosis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
STDs: Time since first diagnosis	21	7	2	3	12	3	11	16	8	8	5	3	16	2	5	18	10	3	2	3	21	1	16	11	20
STDs: Time since last diagnosis	21	7	2	3	12	3	11	16	8	8	5	3	16	2	5	18	10	3	2	3	21	1	16	11	20
Dx:Cancer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Dx:CIN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Dx:HPV	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Dx	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Hinselmann	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Schiller	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
Citology	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Biopsy	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0