Academic Journal of Applied Mathematical Sciences
ISSN(e): 2415-2188, ISSN(p): 2415-5225
Vol. 7, Issue. 4, pp: 192-197, 2021
URL: https://arpgweb.com/journal/journal/17
DOI: https://doi.org/10.32861/ajams.74.192.197

Academic Research Publishing Group

Original Research **Open Access**

Building a Predictive Model for Gynecologic Cancer Using Levels of Data Analytics

Faisal Alamri (Corresponding Author)
Departement of Statistics, Faculty of Science Jeddah university, Saudi Arabia
Email: faisal-stat@hotmail.com

Ezz H. Abdelfattah
Departement of Statistics, Faculty of Science king abdulaziz university, Saudi Arabia

Khalid Sait
Departement of obstetrics and gynecology faculty of medicine king abdulaziz university, Saudi Arabia

Nisreen M. Anfinan
Departement of obstetrics and gynecology faculty of medicine king abdulaziz university, Saudi Arabia

Hesham Sait
Departement of obstetrics and gynecology faculty of medicine king abdulaziz university, Saudi Arabia

Article History
Received: 2 July, 2021
Revised: 18 August, 2021
Accepted: 7 September, 2021
Published: 13 September, 2021

Abstract

The four levels of data analytics techniques (descriptive, diagnostic, predictive, and perspective) were used as a methodology. We also used data mining techniques to predict Gynecologic cancer before any lab test or surgical intervention. Influencing and associating between factors are used to cover hidden relationships or unknown patterns. We focused on three types of Gynecologic cancer (cervical, endometrial, and ovarian cancer). We collected an initial examination of 513 (228 benign and 285 malignant) patients from King Abdulaziz University Hospital (Saudi Arabia). Data were collected during the period of 16 years (2000-2016). After examining many models, we found that the classification trees C5 and CHAID beside the Support Vector Machine (SVM) algorithm give the highest accuracy, with the values of 87.33 %, 79.53%, and 78.36 % respectively. The sensitivity and specificity were found to be 86.18 % and 89.00 % for C5. The corresponding values for CHAID were found to be to equals to 82.20 % and 76.72 % while for support vector machine (SVM) the values were found to be 83.74 % and 77.10 %.

Keywords: Levels of data Analytics; Data mining techniques; And gynecologic cancer.

1. Introduction

Gynecologic cancer represents 15.1% from the total number of cancer incidence type in women excluding non-melanoma skin cancer. Cervical cancer, ovarian cancer and corpus cancer (endometrial cancer also referred to as corpus uterine cancer or corpus cancer) represent 95.3% of all main types of Gynecologic cancer [1]. Using the data mining techniques what is known as knowledge discovery in databases. The technique has the capability to detect hidden relationships and to reveal unknown patterns and trends [2, 3]. Also, use data analytics levels as processes have been automated into mechanical processes and algorithms that work over raw data for human consumption. The research tries to assist the specialist to predict the type of tumor. so, every patient who has a Gynecological tumor visits the clinic and records his initial data (Age, Marital Status, Medical Illness… etc.). The model will try to predict the type of tumor (benign or malignant), without the need for surgical intervention and lab tests. Through the information, we present data analytics processes and data mining techniques to describe and diagnose the disease, and then we build a predictive model to reach the foresight which tell us what will happen. Accordingly, the foresight is considering prescribing the disease to achieve optimization which tell us how can we progress after obtaining the results.

2. Gynecologic Cancer

Gynecologic cancer is a term used for all the types of malignant tumor that can occur in or on a woman's reproductive organs and genitals. Malignant tumor begins in different places within a woman's pelvis, which is the area below the stomach and in between the hip bones. This area can be divided into five main types: cervical cancer, ovarian cancer, uterine cancer (endometrial and uterine sarcoma is "rare"), vaginal cancer and vulvar cancer. All types have different signs and symptoms, different prevention strategies and different risk factors. Some risk factors, like smoking and obesity, can be changed. Others, like a person's age or family history, can't be changed [4-6].

3. Data Analytics

Data analytics can be divided into four types. The first type is descriptive analytics, which are the analytics that describe the past and existing materials. The statistics such as frequency tables, percentage, measures of location (means) and scale (dispersion), etc. fall into this category. The second type is diagnostic analytics, which is the next step to the descriptive analytics that examines data or information to find out the causes of the event. It is characterized by techniques such as correlations, and causation, it provides a very good understanding of a limited piece of our problem. One of the techniques that we use is the Apriori technique for association. Through this, we check the factor that might cause our target (dependent variable). After describing and diagnosing the data, the next step is predictive analytics, which can be defined as the ability to make predictions about unknown future parameters based on past data. It uses many techniques from data mining, such as regression techniques (like linear and logistic regression), classification trees (like C5, C&R, and CHAID tree), and artificial intelligence techniques like Neural Networks and Support vector machines. The last step is prescriptive analytics which is the area of data or analytics dedicated to finding the best course of action for a given situation. In this part, the suggestions and recommendations that emerged from the previous categories of analytics [7, 8].

4. Research Plan

Starting from data collection, the data analytics techniques are applied until reaching the conclusion. This is done by describing the data and organizing it so that it becomes prepared for the next step, which is the diagnosis of the data. In this step, and through the association model, we searching for hidden patterns and factors that may have an impact or a link to the target factor. After that, we build and test many models to reach the best models that describe the phenomenon. Differentiation between models is done by considering the accuracy of the model and the sensitivity and specificity tests. Thus, we obtain models capable of predicting the type of tumor. Through Prescriptive analytics, we explain the results and try to interpret them and suggest what may contribute to the ability to deal with Gynecologic cancer in the future.

5. Descriptive Analytics

Data were collected from one of the Gynecological cancer clinics at King Abdulaziz University Hospital. We have initial routine data of 513 patients, with 79 fields: [Age, Nationality, Body Mass Index (BMI) Parity, Miscarriage, Marital Status (Married before or not) Medical Illness (Contain 36 types of illness), Previous Surgery (Contain 36 type of surgery] and Heart block (HB), Tumor (Malignant or Benign).

Table-1. Numerical Filds

Factor	Min	Max	Mean	SD	Skewness
AGE	13.0	95.0	51.9	11.7	0.468
Parity	0	13	4.4	3.4	0.418
Miscarriage	0	8	0.5	1.1	2.761
BMI	14.6	77	31.1	7.6	1.090
HB	6.8	18.7	11.4	1.7	-0.006

In Table 1, we note that the skewness coefficients are positive for Age, Parity, Miscarriage and BMI, which means that most value of these variables are less than the average. For example, most of the patients' age are less than 51.9. While the skewness coefficients for the HB are negative, which means that most value of HB are more than the mean value of 11.4.

Table-2. Categorical Filds

		Malignant	Benign	Count
Nationalities	Saudi	108 (54.54%)	90 (45.46%)	198 (38.6%)
	Non Saudi	177 (56.2%)	138 (43.8%)	315 (61.4%)
Marital status	Previously Married	260 (54%)	221 (46%)	481 (93.76%)
	Not Married before	25 (86.2%)	4 (13.8%)	29 (5.65%)
	Unregistered	0	3 (100%)	3 (0.59%)
Gynecologic	Tumor (Target)	285 (55.6)	228 (44.4)	513

In Table 2 The fact that there are many more patients with a malignant tumor than benign tumor may seem shocking, but this is due to the fact that many of them were moved from other centres after it was confirmed that they had a malignant tumor. They transferred for examination by a gynaecologic oncologist. It is remarkable to note that about 86.2% of the unmarried patients are having a malignant tumor while only 54% of women whom had been married before are having a malignant tumor.

Table-3. The most frequent diseases and surgeries

	Most frequent	Malignant	Benign	Count
Medical Illness	Hypertensive	98 (55.7%)	78 (44.3%)	176 (34.3%)
	Diabetes Mellitus	81 (66.9%)	40 (33.1%)	121 (23.6%)
	Bronchial Asthma	14 (46.7%)	16 (53.3%)	30 (5.8%)
Previous Surgery	Dilation and Curettage	24 (45.3%)	29 (54.7%)	53 (10.3%)
	Caesarean delivery	16 (43.2%)	21 (56.8%)	37 (7.2%)
	Laparoscopic Cholecystectomy	7 (38.9%)	11 (61.1%)	18 (3.5%)

Through Table 3, Hypertension, diabetes, and bronchial asthma were the most common diseases they had according to their medical illness history while the most frequent previous surgeries that patients underwent were dilation and curettage, caesarean delivery and laparoscopic Cholecystectomy. Furthermore, approximately 70% of patients that **had diabetes** have a malignant tumor and 34.3% of all patients had hypertension.

6. Diagnostic Analytics

We use the **Apriori** algorithm as the association model, apriori is a classic algorithm for learning association rules. As it is common in association rule mining, given a set of item sets the algorithm attempts to find subsets that are common to at least a minimum number K of the item sets. Apriori uses a bottom up approach, where frequent subsets are extended one item at a time, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found (3).

The support of an association pattern is the percentage of task relevant data transactions for which the pattern is true. The formula of support:

$$\text{Support } (\mathbf{A} \Longrightarrow \mathbf{B}) = \text{Tuples containing both A and B / Totals of tuples} \quad (1)$$

Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern. The formula of Confidence:

$$\text{Confidence } (\mathbf{A} \Longrightarrow \mathbf{B}) = \text{Tuples containing both A and B / Tuples Containing A [9]} \quad (2)$$

Table-4. Apriori techique result

Consequent	Antecedent	Support %	Confidence %
Tumor	Diabetes Mellitus	23.59	66.94
	Diabetes Mellitus and previously Married	23.00	66.10
	Diabetes Mellitus and Nationality = Saudi	11.70	65.00
	Diabetes Mellitus and Hypertensive	15.01	64.94
	Diabetes Mellitus and Nationality = Saudi and Previously Married	11.50	64.41
	Diabetes Mellitus and Hypertensive and Previously Married	14.42	63.51
	Nationality = Yemeni	15.20	61.54
	Nationality = Yemeni and Previously Married	14.42	60.81

Apriori technique indicates that diabetes mellitus, marital status, and nationality are three main factors that could affect the tumor type. The level of confidence is not strong. However, the importance of these variables may increase if they appear repeatedly as an influencing factor in the predictive models.

7. Predictive Analytics

To explore the type of Gynecologic tumor, we need to analyze this amount of data and extract useful information from it. Furthermore, we need to perform; data mining that involves other processes such as data cleaning, data integration, data transformation and data presentation [10]. Using a predictive model to summarize the data automatically and extract the essence of information stored [3] to find the influencing factors. Here, we will concentrate on specific **data mining techniques**, namely the classification tree C5, the classification tree CHAID and the support vector machine (SVM), that we find to have the highest accuracies among too many different data mining techniques.

7.1. The Classification Tree C5

C5.0 algorithm is a successor of C4.5 algorithms (developed by Quinlan [11]) gives a binary tree or multi branches tree. Uses Information Gain (Entropy) as its splitting criteria.C5.0 pruning technique adopts the Binomial Confidence Limit method. C5 algorithm is an extension of C4.5. In C4.5, all the errors were taken equally. There was no segregation of the errors based on their importance or significance. A clear improvement in C5 over C4.5 is that it treats all the errors. In mathematical terms that can be calculated by the following formula [12, 13]:

$$\text{Information Entropy: } E(S) = \sum_{i=1}^{c} -p_i \log_2(p_i) \quad (3)$$

p_i: The proportion of the number of classes

$$\text{Conditional Information Entropy: } E(S|A) = -\sum_{i=1}^{v} p_i' \sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \quad (4)$$

The information gain ratio can be calculated as follows:

$$Gain\ (A) = E(S) - E(S|A) \quad (5)$$

7.2. The Classification Tree: CHAID

CHAID as a methodological approach appears in the literature under various names, including: decision tree, automatic interaction detection, classification and regression tree, artificial neural network and genetic algorithm. CHAID algorithm was the first introduced by Kass [14]. Statistical Tests CHAID used to determine the next best split: Chi-square for categorical dependent variable and F-Test continuous dependent variable [15].

Chi-square test formula:

$$\chi^2 = \sum \frac{(y - \acute{y})^2}{\acute{y}} \quad (6)$$

y: Observed value
$\acute{y}$: Expected value
C: Degrees of freedom

F-Test test formula:

$$F = \frac{BSS/(g-1)}{WSS/(n-g)} \sim F_{(g-1).(n-g)} \quad (7)$$

g is the number of groups
N is the total number of observations
BSS is the between-group sum of squares
WSS is the within-group sum of squares

7.3. Support Vector Machine (SVM)

Cortes, *et al.* [16], proposed a support vector network as a new learning machine for two-group classification problems. The key idea is that input vectors are nonlinearly mapped to a high dimensional feature space, where a linear hyperplane can be constructed that gives maximum separation of the groups. Using dot products, the solution becomes computationally feasible.

$$hypothesis\ function: \begin{cases} +1 & if\ w^T \cdot x + b \geq 1 \\ -1 & if\ w^T \cdot x + b \leq -1 \end{cases} \quad (8)$$

where w is a weight vector, x is input vector and b is scaling constant (bias). For the dual problem, we have that:

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i . x_j \quad (9)$$

$$\alpha_i \geq 0 . i = 1 \ldots m . \sum_{i=1}^{m} \alpha_i y_i = 0$$

Where α is the Lagrange multiplier and $K(x_i . x_j)$ as kernel function. So we use the polynomial kernel: $K(x_i . x_j) = (x_i . x_j + b)^d$, Where d is a degree of freedom [16].

8. Results

Table-5. Predictive Models

Model	Accuracy	Sensitivity	Specificity
C5	87.33 %	86.18 %	89.00 %
CHAID	79.73 %	82.20 %	76.72 %
SVM	78.36 %	83.74 %	77.10 %

Influencing factors of C5, CHAID and SVM (Noting that: I. refer to Illness and P.S. refers to Previous Surgery)

Table-6. Influencing Factors

CHAID		SVM		C5	
Nodes	**Importance**	**Nodes**	**Importance**	**Nodes**	**Importance**
I.DM (Diabetes Mellitus)	0.23	Miscarriage	0.12	Nationality	0.08
AGE	0.21	HB (Heart Block)	0.09	BMI	0.08
Miscarriage	0.16	Nationality	0.06	I.DM (Diabetes Mellitus)	0.06
Marital status	0.10	BMI	0.04	Miscarriage	0.03
P.S.Hernia Repair	0.05	P.S.Laparoscopic Cholecystectomy	0.03	I.Hypothyroidism	0.03
P.S.Myomectomy	0.04	P.S.C/S (Cesarean Section)	0.03	AGE	0.03
Parity	0.04	Marital status	0.02	Marital status	0.02
P.S.Laparoscopic Cholecystectomy	0.03	AGE	0.02	I.DUB (Dysfunctional Uterine Bleeding)	0.02
HB (Heart Block)	0.02	P.S.Myomectomy	0.02	P.S.Laparotomy	0.02
I.AF(Atrial fibrellation)	0.02	I.Dyslipidemia	0.02	P.S.Appendectomy	0.02

9. Conclusions and Recommendations (Prescriptive Analytics)

We have tested many models, whether they are statistical, artificial intelligence, or decision trees models. We focus on the three best models in terms of accuracy to present the sensitivity, specificity, influencing factors and association factors. C5 has the highest accuracy, sensitivity and specificity.

All the three models were able to identify the disease well with sensitivity between 86-82 %. There is a variation in the factors affecting each model (CHAID, C5 and SVM). For example, C5 sees nationality as the largest influencing factor, and SVM, ranked nationality the third influence factor, while CHAID, does not consider nationality to be one of the influence factors. For Association factors, the confidence level is not strong. We see diabetes significantly in common with some of the other factors which are nationality, marital status and hypertensive disease.

Artificial intelligence models and decision trees were able to process the data very well and the results of the association models were not strong, but they give an indication of the association of diabetes with many other factors. So adding other factors in the future may lead us to clarify the effect of diabetes on the tumor better than the current factors.

We previously mentioned the fact that each gynecologic cancer is unique, with different signs, symptoms and different risk factors. In the case of our research, the model deals with all indicators recorded for patients means it contains all three types of Gynecological tumors. This leads us to work on studying each type of Gynecological cancer separately as future works to improve our results

Getting an expectation of the type of tumor through initial routine diagnosis does not prevent laboratory testing or surgical intervention, but it helps the specialist to reduce the procedures followed to reach the correct diagnosis, and this contributes to speeding up decision-making.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. *Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA Cancer J. Clin, in press.

[2] Data mining concept by Doug Alexander. Available: http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/

[3] Zaïane and Osmar, R., 1999. "Introduction to Data Mining, CMPUT690 Principles of Knowledge Discovery in Databases Chapter."

[4] https://www.cancer.gov/.

[5] https://www.cancer.org/.

[6] https://www.cdc.gov/.

[7] Abdelfattah, E. H., 2020. "Reclassifying inferential statistics into diagnostic and predictive statistics with an application on gynecologic cancer." *Biometrics and Biostatistics International Journal*, vol. 9, pp. 146-150.

[8] Ahmed, Mohiuddin, and Al-Sakib, K. P., 2018. *Data analytics: Concepts, techniques, and applications*. CRC Press.

[9] Kusiak, A. *Association rules-the apriori algorithm. Handout of intelligent system lab*. USA: The University of IOWA.

[10] https://www.tutorialspoint.com/data_mining/dm_overview.htm.

[11] Quinlan, J. R., 1992. *C4.5 programs for machine learning*. San Mateo: CA: Morgan Kaufmann.

[12] Muhammad, A., 2018. "Decision tree algorithms C4.5 and C5.0 in data maining: A review." *International Journal of Database Theory and Application*, vol. 11, pp. 1-8.

[13] Pang, S.-l. and Ji-zhang, G., 2009. "C5. 0 classification algorithm and application on individual credit evaluation of banks." *Systems Engineering-Theory and Practice*, vol. 29, pp. 94-104.

[14] Kass, G. V., 1980. "An exploratory technique for investigatin large quantities of categorical data." *Applied Statistics*, vol. 29, pp. 119-127.

[15] Díaz-Pérez, Flora, M. A., and Bethencourt-Cejas, M. A., 2016. "CHAID algorithm as an appropriate analytical method for tourism market segmentation." *Journal of Destination Marketing & Management*, vol. 5, pp. 275-282.

[16] Cortes, Corinna, and Vladimir, V., 1995. "Support-vector networks." *Machine learning*, vol. 20, pp. 273-297.