EPH - International Journal of Applied Science
ISSN (Online): 2208-2182
Volume 02 Issue 02-September-2016

DOI: https://doi.org/10.53555/eijas.v2i3.139

ANALYZING THE BREAST CANCER DATA IN BMC USING THE TECHNIQUES OF ASSOCIATION RULES AND LINEAR DISCRIMINANT

Rami Salah Gebril[1*], Hanan Mohammed Ali[2]
*[*1, 2]University of Benghazi, Faculty of Science, Department of Statistics*

Corresponding author:-
E-mail: *rsmg4r@gmail.com*

Abstract:-

Decision making in considered to be one of the main goals of Applied Statistical Analysis or Data Mining, especially in the field of medicine when decisions about criticalmedical procedures concerning patients are to be made regarding clinical diagnosis, surgical operations and treatments. In this study, a medical data related to breast cancer patients, are being analyzed using two statistical techniques. The data represent patients records in Benghazi Medical Center (BMC) admitted in the period (2003 – 2005). The data matrix consists of 16 variables corresponding to a total of 263 patients (cases).

As a primary step of the analysis, the Exploratory Data Analysis (EDA) powerful tools are being used to explore the distribution and behavior of the data variables to provide a better understanding of patient status. The technique of Association Rules is used in this study as an alternative approach, (to the classical Correlation Analysis), to analyze the multiple causal relationships between data variables, due to their specific nature.The second technique is the Fisher Linear Discriminant Analysis (LDA) which is applied as a supervised method to classify data responses with 2-class and k-class nature based on data predictors. As a general goal of this study, it is attempted to provide a decision making statistical support to help medics in the phase of prognosis of new breast cancer patient cases, based on the previous knowledge (disease behavior).

Keywords:- *Exploratory Data Analysis, Data Mining, Association Rules, Linear Discriminant Analysis.*

1. INTRODUCTION

One of the "modern" Statistical Analysis, (known as Data Mining), various merits that it limits human subjectivity in decision making processes in general, and it handle large numbers of variables with increasing speed, thanks to the growing power of computers.

Decision making, on the other hand, is considered to be an objective of data mining and statistical analysis, since that it is expected from statistical techniques to do more than simply provide models or descriptive results, [1]. The approach of decision making is not completely new, and is already established in many fields such as in medicine, where some important decisions regarding clinical diagnosis, surgical operations, and treatments have been developed on the basis of statistical analysis results. Even though the biological mechanism of the disease is little understood because of its complexity, as in the case of some cancers.

In the field of medicine, the data mining approach is used frequently and heavily in both diagnostic and predictive applications, [2]. In the area of disease diagnoses, (which includes the identification of patient groups suitable for specific treatment protocols, where each group includes all the patients who react in the same way), studies on detecting prescription anomalies using correlations or associations between medicines is widely applied.In Predictive applications, (which includes tracing the variables responsible for death or survival in certain diseases, such as heart attacks, cancer, etc.), the statistical analysisis occasionally used in order to explore the most appropriate treatment to match the pathology and the patients.

In this study, we attempt to provide a supportive decision making tool to help medics in the phase of prognosis of new breast cancer patient cases, based on previous knowledge of disease behavior.

2. Exploring the Study Data

2.1 Description of the Breast Cancer Data

The data under study are registered from patient records in Benghazi Medical Center in the city of Benghazi in Libya during the years 2003, 2004, and 2005. The data matrix consists of 16 variables corresponding to a total of 263 patients (cases). These variables are defined in Table (2.1).

Table 2.1: Description of variables of Breast cancer Data

	Variable Name	Variable Description
1	Age	Patient age at admission in years.
2	Age-group	Patient age within pre-specified age groups; (39 or less), (40 to 49), (50 to 69), or (70 or more).
3	Surg.	The type of surgery carried out; (No, SMAC, or BCS). Where NO indicates that no surgery was done, SMAC refers to Surgical Mastectomy with Axillary Clearance, (i.e. removing the whole breast), and BCS refers to Breast Conserving Surgery, (removing only the tumor).
4	Stage	The stage of breast cancer; (I, II, III, IV, or V).
5	ER.	Estrogen Receptor blood analysis, giving one three results; plus, minus, or NA (Not Available).
6	PR.	Progesterone Receptor blood analysis, giving one three results; plus, minus, or NA (Not Available).

Cont. Table 2.1: Description of variables of Breast cancer Data

	VariableName	Variable Description
7	HER.	Hereditary factor, giving one three results; plus, minus, or NA (Not Available).
8	Chem.	The type of chemical therapy executed; (NONE, FAC (Fluorouracil), FEC (Fluorouracil Encomycin), Taxane, Anthracyclin then Taxane, CMF (Cyclophosphamide Methotrexate Fluorouracil), or other).
9	Rad.	A binary variable indicating that the patient received a radiation therapy or not; (NO or received).
10	Horm.	The type of hormonal therapy executed; (NO, Nolvadex, Nolvadex+Zoladex, Nolvadex+AL, or AL).
11	TumReturn	A binary variable that indicates return of the tumor to the patient; (NO or Yes).
12	TumReturn Time	The time (in months) passed until the return of the tumor.
13	TumReturn Int.	The time interval passed until the return of the tumor; (NO, during 1Y, during 3Y, during 5Y, during 10Y, or after 10Y).
14	TumReturn Place	The place of second tumor; (NO, Bone, Brain, Liver, Local, Local and Organ, Lung, multiple Organ, other Breast, other Breast and Organ, or other).
15	2nd Treat.	The type of second treatment given to patient; (NONE, Hormonal therapy, Single Chemotherapy, Poly Chemotherapy, or other).
16	5 Years Period	A binary variable indicating that the patient is alive or dead during a period of 5 years.

2.2 Exploratory Data Analysis (EDA) of the study Data

In this section, the results of a brief Exploratory Data Analysis are presented to gain an overall look of the variables distributions and behavior before proceeding to advanced statistical analysis, [3]. Most meaningful and representative descriptive results will be discussed here throughout the following points;

Concerning the age of breast cancer patients, it was noticed from the stem and leaf plot in Figure (2.1) that the patient ages in the sample ranges from 26 to 90 years old with mean equal to 48 years and approximately 13 years variation among them. The population mean of patient age is estimated between 47 to 50 years.

stem°leaf (leaf unit=10.00000, e.g., 6°5 = 65.00000)	Class n	Percentiles
1°	0	
2°	0	
2° 667777789	9	
3° 000000111222223333333444	24	
3° 555555556666666677777888888888999999999	39	25%
4° 000000000000001222222222222222333333333344444	45	
4° 555555555555566777777777788888999	35	median
5° 000000000000000111112222344444	31	
5° 5555556666777888888899	22	75%
6° 000000000011122222333344444	26	
6° 55555555555789	14	
7° 00000122	8	
7° 5678	4	
8° 0004	4	
8° 5	1	
9° 0	1	
9°	0	
min = 26.00000 max = 90.00000 Total N:	263	

Figure 2.1: The stem and leaf plot of patients ages.

Most patients, are between 39 to 58 years old, but this is not necessary an indication that the disease is starting at this age period because most cases are classified in advanced stages of the disease due to late diagnosis. Patients ages are not distributed normally, (according to Kolmogorov-Smirnov Test: d=0.097, with P-value < 0.05), and they are skewed to the right. In general, these results indicates that most women in Benghazi city do not get early diagnosis for breast cancer, and hence the disease is usually discovered at advanced stages.

Moving to categorizing study variables by disease stage, we obtained the following results:

(a) The variable **Surg.** Categorized by **Stage**: Figure (2.2) shows that the surgery of type SMAC is the most preferred one at stages II, III, and V, while stage I did not appear in the plot due to late diagnosis of the disease as mentioned earlier.

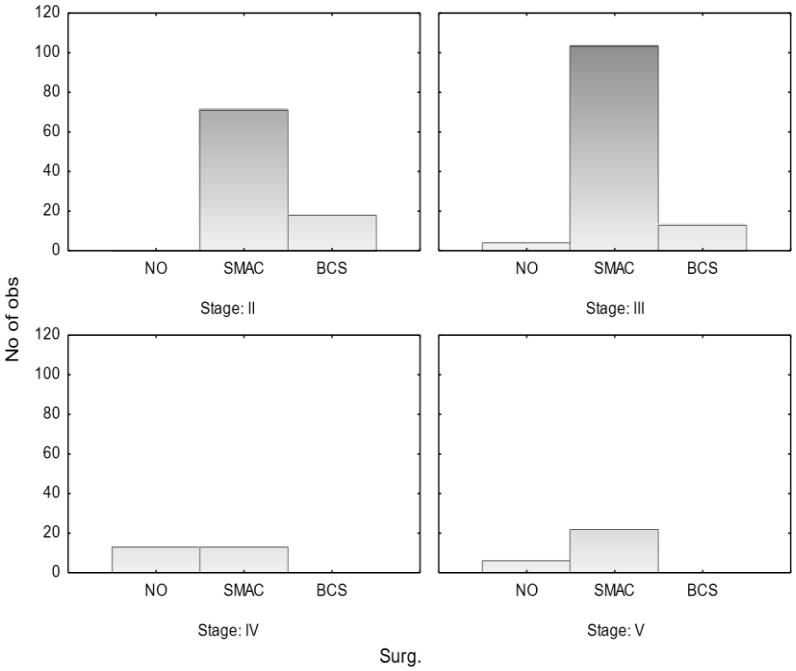

Figure 2.2: The histogram of variable Surg. Categorized by Stage.

(b) The variable **Rad.** Categorized by **Stage**: From Figure (2.3), it can be seen that the radiotherapy is mostly applied at the third stage of the disease.

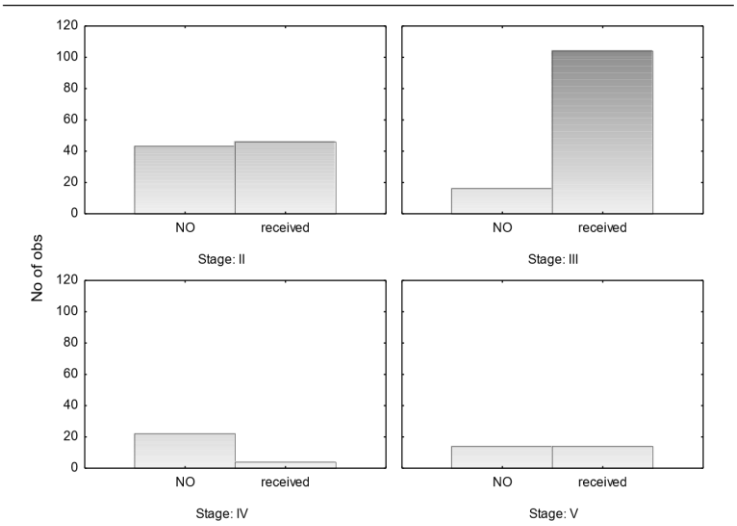

Figure 2.3: The histogram of variable Rad. Categorized by Stage.

(c) The variable **Horm.** Categorized by **Stage**: It can be noticed from Figure (2.4) that the hormonal therapy of type Nolvadex is mostly used at all stages, especially at stage II and III of the disease.

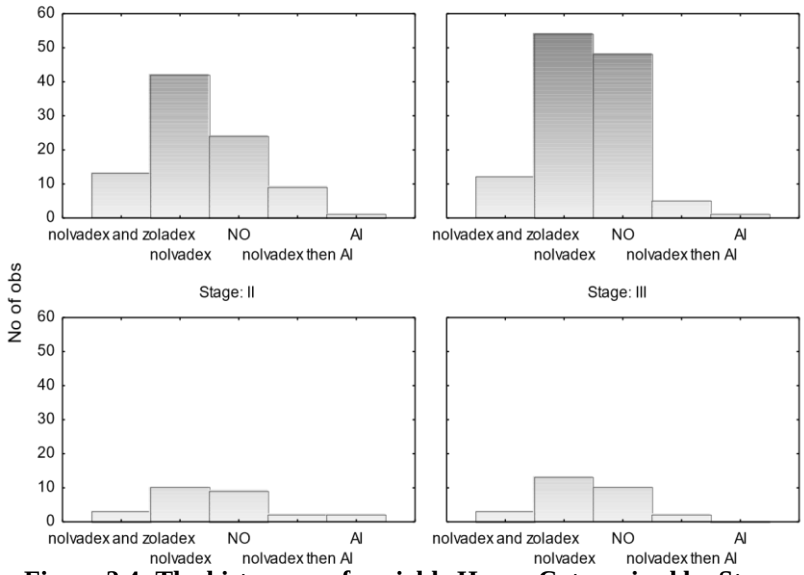

Figure 2.4: The histogram of variable Horm. Categorized by Stage.

(d) The variable **5 Years Period** Categorized by **Stage**: From Figure (2.5) it can be seen that patients at stages II and III are more likely staying alive within the period of five years while patients at advanced stages are exposed to death slightly more.

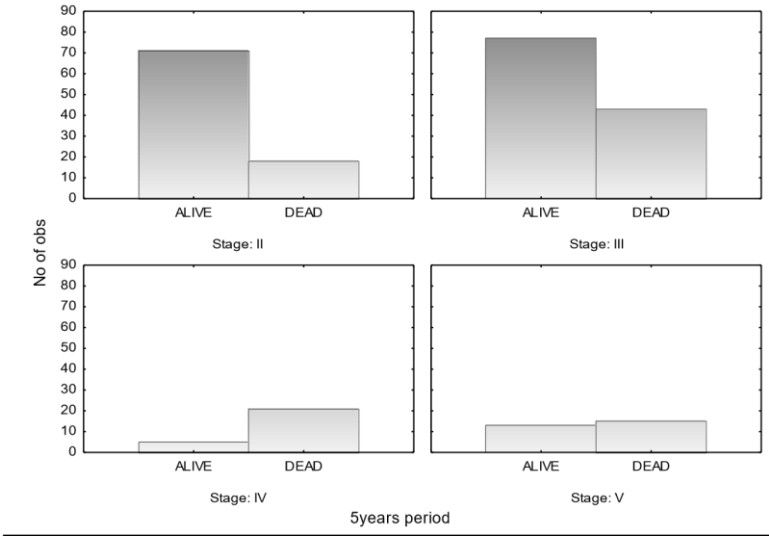

Figure 2.5: The histogram of variable 5 Years Period Categorized by Stage.

In Figure (2.6), the distribution of several variables are investigated using the histograms and the following conclusions areobserved; • The SMAC type of surgery is the most used one (79%).

- No patients were registered at stage I in the study sample, and most patients were in stages II and III (79%).
- Most used Chemotherapy was FAC (46%) followed by FEC (25%).
- 64% of the patients received Radiotherapy.
- Most patients either received Nolvadex hormonal therapy (45%) or received no hormonal therapy (35%). • For those patients that suffered from tumor return, 22% of the tumor return places were Local and Bone.
- 65% of patients did not receive 2nd treatment.

In addition, the correlations among data variables were calculated and tested, and the following significant correlations were noticed;

1. Patient age with radiotherapy, hormonal therapy, and 2nd treatment.
2. Disease stage with radiotherapy, tumor return, tumor return place, and 2nd treatment.
3. Tumor return with tumor return place, 2nd treatment, and the 5 years period of survival.
4. Tumor return place with 2nd treatment and the 5 years period of survival.
5. 2nd treatment and the 5 years period of survival.

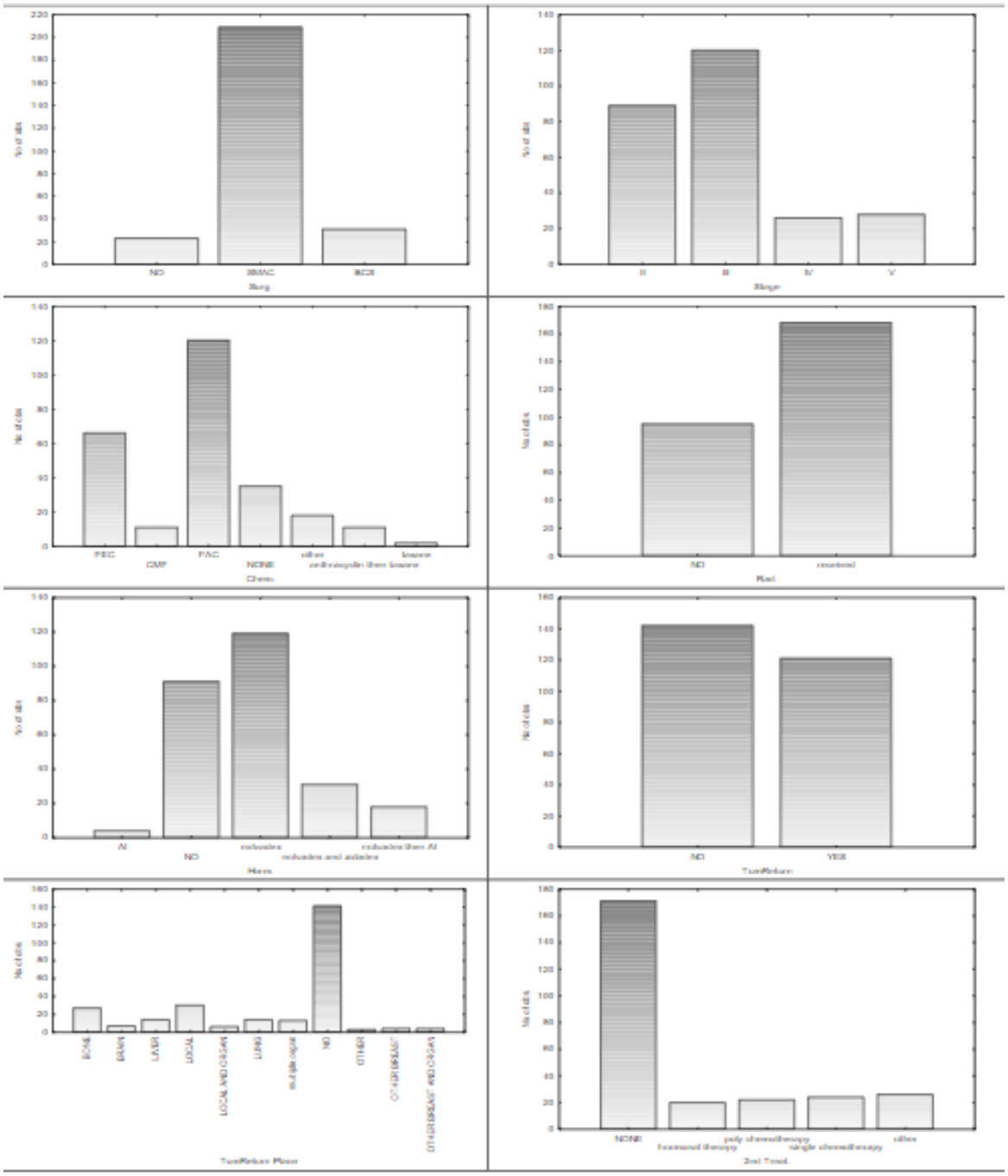

Figure 2.6: The histogram plot for the variables; Surg., Stage, Chem., Rad., Horm., TumReturn, TumReturn Place, and 2nd Treat.

Methodology of Association Rules and Linear Discriminant Analysis

The Association Rules method is adopted in this study due to its simple and powerful logic which reveals all the hidden correlations existing among the relatively large number of variables in the study data. The "ordinary" linear and non-linear correlation coefficients are usually suitable in calculating relationships between two or more variables directly, while as the association rules can reflect the relationships between several groups of variables, containing one or more variables in each group

The Discriminant Analysis, on the other hand, is a classical supervised classification method, and it is used in our study to classify several categorical responses depending on data predictors.

3.1 The Concept of Association Rules

The original definition of Association Rules Mining can be stated as follows,[4]; Define $I=\{i1,i2,...,im\}$to be a set on m items, and let $T=\{t1,t2,...,tn\}$be a set of n transactions. Each transaction in T has a unique transaction ID and contains a subset of the items in I. Now having any two subsets of items in I; A and B, an association rule can then be defined as an implication of the form:

$$A \rightarrow B \text{ where } A,B \subseteq I \text{ and } A \cap B \subseteq \emptyset$$

The item-set A is called antecedent or left-hand-side (LHS) and the item-set B consequent or right-hand-side (RHS). The item-sets A and B can consist of single values (quantitative variable), words (categorical variable), or conjunctions of values, or words. The probability of an item-set, A for example, will be calculated by;

$$P(A)=\frac{\sum_{i=1}^{n} c(A,t_i)}{n},\ where\ c(A,T)=\begin{cases}1\ if\ A\subseteq T\\0\ otherwise\end{cases}$$

In other words, you can use the association rules model to find rules of the kind:

$$\textbf{If}\ A\ \textbf{Then}\ (\text{likely})\ B$$

Or more generally:

$$\textbf{If}\ Body\ \textbf{Then}\ Head$$

The support value of A with respect to the transaction set T is defined as the proportion of transactions in the data which contains the item-set A, and it is denoted by *sup(A)*, and a high value means that the association rule involve a great part of data. Thus, we can write;

$$supp(A\rightarrow B)=P(A\cup B)$$

This conditional probability (that an observation (transaction) that contains a code or text value A also contains a code or text value B) is called the Confidence Value. In general, the confidence value denotes the conditional probability of the Head of the association rule, given the Body of the association rule. That is;

$$Conf(A\rightarrow B)=P(B\backslash A)=\frac{P(A\cap B)}{P(A)}$$

In addition, the support value for each pair of codes or text values, and a Correlation value based on the support values will be computed. The correlation value for a pair of codes or text values {A, B} is computed as

$$Corr(A,B)=\frac{supp(A\cap B)}{\sqrt{supp(A).supp(B)}}$$

3.2 The Linear Discriminant Function.

The theoretical model which the Fisher Linear Discriminant function, (or the Linear Discriminant Analysis (LDA) as it commonly known), can be defined briefly as follows, [5];Let $X1,X2,...,Xp$be a set of p predictors, and Y be a two-level categorical variable corresponding to no observations, then Fisher's model will try to find a linear function Y of the predictors $X1,X2,...,Xp$;

$$Y=const.+\alpha_1X_1+\alpha_2X_2+\cdots+\alpha_pX_p$$

such that the ratio of the between-group variance of Y to its within-group variance is maximized, that is the coefficients $\alpha j,j=1,...,p$ are chosen so that the $\text{ratio}\,V=\frac{\alpha' B\alpha}{\alpha' S\alpha}$is maximised. Where B is the covariance matrix of group means and S is the pooled within-group covariance matrix;

$$B=\sum_{i=1}^{2} n_i(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})'$$

$$S=\frac{1}{n-2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\bar{\mathbf{x}}_j)(x_{ij}-\bar{\mathbf{x}}_j)'$$

Where xij is the jth individual in the ith group, $\bar{\mathbf{x}}j$is the mean vector of the jth group, and $\bar{\mathbf{x}}i$is the overall mean vector. The number of observations in each group is n1 and n2, with n= n1+ n2.The Fisher model have the following assumptions;

- The X's are multivariate normally distributed.
- $Var(X_j)$, $j=1,\ldots,P$ are homogeneous.
- $E(X_j)$ and $Var(X_j)$, $j=1,\ldots,P$ are uncorrelated.

Now, we need an assessment method to assess the classification model, this can be done using Confusion or Classification Matrix which is given in Table (3.1);

Table 3.1: The Classification Matrix in two-level response LDA

		Predicted class		
		$\hat{\pi}_1$	$\hat{\pi}_2$	
True class	π_1	n_{1C}	n_{1m}	n_1
	π_2	n_{2m}	n_{2C}	n_2
		n_1	n_2	$n = n_1 + n_2$

- n_{1C} = number of $\hat{\pi}_1$ obs.'s *correctly* classified as π_1.
- n_{2C} = number of $\hat{\pi}_2$ obs.'s *correctly* classified as π_2.
- n_{1m} = number of $\hat{\pi}_1$ obs.'s *misclassified* classified as π_2.
- n_{2m} = number of $\hat{\pi}_2$ obs.'s *misclassified* classified as π_1.

Then the Apparent Error Rate (APER), which will measure the percentage of misclassification of the data using the fitted discriminant model, can be calculated as;

$$APER = \frac{n_{1m} + n_{2m}}{n}$$

The hypothesis Ho: Discriminant model is insignificant, can be tested using the statistic;

$$Wilk's\ Lambda = \frac{|WSS|}{|TSS|}$$

Where WSS and TSS are the within and total sum of squares of data. So the value of Wilks' Lambda will range from 0 (indicating perfect discrimination) to 1 (indicating no discrimination)

4. Applying the Association Rules and LDA on the Study Data

4.1 Results of applying the Association Rules

The results in this section will consist of a table representing the (*If* Body *Then* Head) relation rule with the corresponding Support, Confidence, and Correlation percentages. An Association Rule Network (plot) will be associated with each result table to enhance the conclusions about the importance of each relationship (association). Only the first two results will be presented here due to enormous number or complex relation existing among variables, and these foundings will be discussed in the last section of this paper.

From Table (4.1), it can be noticed, with high confidence and correlation values, that breast cancer patients who had a surgery of type "SMAC" did not show a heritage factor of the disease in the family, and at the same time didn't receive a radiotherapy.

Table 4.1: Association Rules results, with *Age-group* as a response.

Body	Implies	Head	Supp. (%)	Conf. (%)	Corr. (%)
Surg. == SMAC	==>	HER == NA	53.33	68.90	75.51
Surg. == SMAC	==>	Rad == received	52.59	67.94	75.78
ER == plus	==>	PR == plus	40.00	90.00	91.15
PR == plus	==>	ER == plus	40.00	92.31	91.15
HER == NA	==>	Surg. == SMAC	53.33	82.76	75.51
Rad == received	==>	Surg. == SMAC	52.59	84.52	75.78

The association rules can be visually investigated also as seen in Figure (4.1). From that graph, the importance of the associations between the variables *Surg.*, *Her*, and *Rad.* are observed throughout the large sizes and dark colors of their circles. These sizes and colors corresponding to the above three variables indicates the strong causal relation between the absence of heritage disease factor and receiving radiotherapy at one hand, and having the "SMAC" surgery at the other hand.

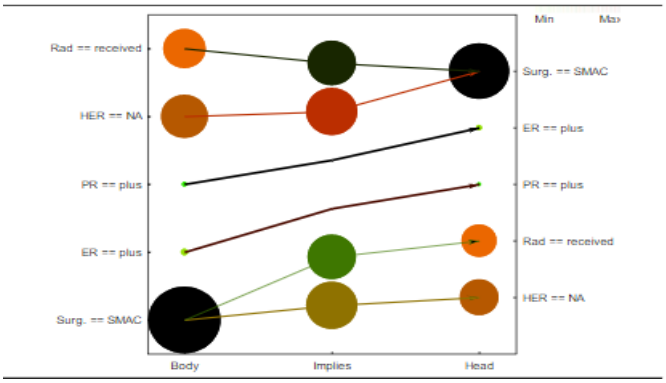

Figure 4.1: The Association Rule network, with *Age-group* as a response.

When using the variable *5 Years Per.* as a response variable, it can be seen, from Table (4.2) and Figure (4.2), that when a surgery of type "SMAC" is applied, a radiotherapy is taken with no trace of heritage factor, then this implies that the patient will survive for a 5 years period with confidence values; 67%, 63%, and 67% respectively.

Table 4.2: Association Rules results, with *5 Years Per.* as a response.

Body	Implies	Head	Supp. (%)	Conf. (%)	Corr. (%)
Surg. == SMAC	==>	HER == NA	53.33	68.90	75.51
Surg. == SMAC	==>	Rad == received	52.59	67.94	75.78
Surg. == SMAC	==>	ALIVE	51.48	66.51	74.63
ER == plus	==>	PR == plus	40.00	90.00	91.15
PR == plus	==>	ER == plus	40.00	92.31	91.15
HER == NA	==>	Surg. == SMAC	53.33	82.76	75.51
HER == NA	==>	ALIVE	40.74	63.22	64.72
Rad == received	==>	Surg. == SMAC	52.59	84.52	75.78
Rad == received	==>	ALIVE	41.85	67.26	67.67
ALIVE	==>	Surg. == SMAC	51.48	83.73	74.63
ALIVE	==>	HER == NA	40.74	66.27	64.72
ALIVE	==>	Rad == received	41.85	68.07	67.67

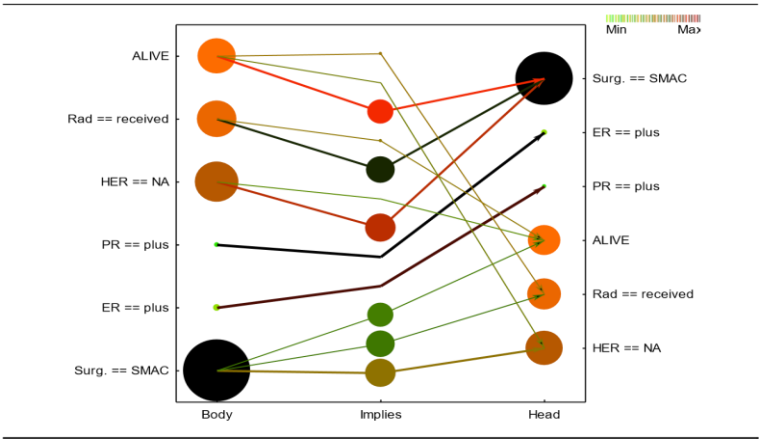

Figure 4.2: The Association Rule network, with *5 Years Per.* as a response.

4.2 Results of applying the Linear Discriminant Analysis

All possible linear discriminant functions were fitted by choosing one of the following variables as a dependent (response); *Surg.*, *Chem.*, *Horm.*, *TumReturn*, and *5Years Per.* And at the same time, several combinations of predictors were entered in models, these combinations were of 2, 3, 4, up to 10 predictors. And hence, we got a large number of models to investigate, ($C_2{}^{10} + \cdots + C_{10}{}^{10} = 1013$ linear discriminant models). Some of these models had a binary dependent variable, (that is 2-class response), and the other had a multicategory response, (k-class response).

The results of fitting all the possible models will not be all shown in this section because of the large number of related outcomes, and only the highly important results will be discussed.

(a) Variable *Surg.* as a response:

In Table (4.3), it can be seen that the predictors; *Stage*, *Rad.* and *Her.* All have a significant effect on the classification of the type of surgery performed (*Surg.*), with an overall significance of the model.

Table 4.3: Summary results of fitting LD model of *Surg.* on *Stage*, *Rad.* and *Her.*

	Overall Wilks' Lambda: 0.70, P-value < 0.05	
Predictor	Wilks' Lambda	P-value
Stage	0.83	0.00
Rad.	0.76	0.00
Her.	0.75	0.00

The Classification Function of this model is shown in Table (4.4), and this function can be used for prediction purposes.

Table 4.4: Classification function of LD model of *Surg.* on *Stage*, *Rad.* and *Her.*

Variable	Surg.: NO P=0.09	Surg.: SMAC P=0.80	Surg.: BCS P=0.11
Stage	2.92	1.25	0.53
Rad.	1.28	3.92	4.02
Her.	4.04	3.32	2.31
Const.	-9.61	-4.67	-4.91

The predictive discriminant functions can be written as follows;
$Surg.(NO) = -9.61 + 2.92\ Stage + 1.28\ Rad. + 4.04\ Her.$
$Surg.(SMAC) = -4.67 + 1.25\ Stage + 3.92\ Rad. + 3.32\ Her.$
$Surg.(BCS) = -4.91 + 0.53\ Stage + 4.02\ Rad. + 2.31\ Her.$
These predictive functions can be used to predict future decisions about the type of surgery to be carried out on cases of patients suffering from breast cancer. This is done by entering the values of the predictors *Stage*, *Rad.* and *Her.* And selecting the class with the highest score.
For example, a patient with the following state; *Stage*=II, *Rad.* =received and *Her.* =minus will achieve the following scores:

$$Surg.(NO) = -4.29, \qquad Surg.(SMAC) = 0.25, \qquad Surg.(BCS) = 0.11$$

And hence the statistical point of view will encourage the application of a surgery of type "SMAC", which is removing the whole breast.

(b) Variable *Horm.* As a response:
In Table (4.5), the overall model was significant and the predictors *Age* , *HER.*, and *ER* were found to have important effect on the type of hormonal therapy, (NO, Nolvadex, Nolvadex+Zoladex, Nolvadex+AL, or AL).

Table 4.5: Summary results of fitting LD model of *Horm*. On *Age, HER.* And *ER.*

Predictor	Overall Wilks' Lambda: 0.67, P-value < 0.05	
	Wilks' Lambda	P-value
Age	0.77	0.00
HER.	0.70	0.04
ER	0.87	0.00

Table (4.6) shows the classification function of the model. The hormonal therapy of type "Nolvadex" have the largest probability of occurrence among the other types of hormonal therapy.

Table 4.6: Classification function of LD model of *Horm*. On *Age, HER.* And *ER*.

Var.	Horm.: NO P=0.34	Horm.: Nolvadex P=0.45	Horm.: Nolvadex+Zoladex P=0.11	Horm.: Nolvadex+AL P=0.07	Horm.: AL P=0.02
Age	0.30	0.34	0.250	0.36	0.35
HER.	1.70	2.21	2.301	2.66	2.57
ER	1.36	0.21	-0.616	-0.65	-1.21
Const.	-10.18	-11.37	-8.374	-14.45	-15.10

5. Summary and Conclusions
A summary of the results obtained in this study is demonstrated here, starting with the results of breast cancer data exploration, passing by the investigation of multiple relationships between data variables, and concluding with the results of categorical variable classifications (modeling). The results are listed in the following points, where the first section will represent the results of applying the EDA, and the second section will involve the results of data modeling;

5.1 Conclusions about Breast Cancer Data Exploration
- The breast cancer patient's ages in the sample ranged from 26 to 90 years old with estimated population mean between 47 to 50 years, indicating that the disease is discovered at high ages. But this is not necessary an indication that the disease is starting at this age period because most cases are classified at advanced stages of the disease perhaps due to late diagnosis.
- The "SMAC" type of surgeries, (which means removing the whole breast), was noticed to be mostly performed for all age groups.
- Most breast cancer patients in the sample were diagnosed to be at stages II and III, and this observation fortifies the opinion about the late discovery of the disease since no early diagnosis is done for most women.
- Chemotherapy of type "FAC" and "FEC", (Fluorouracil, and Fluorouracil Encomycin respectively), are mostly applied for patients in all age groups except senior women, (70 years or above).
- Most patients in the age group 39 and under did not receive a hormonal therapy, while most patients at other age groups have received a hormonal therapy of type "Nolvadex".
- No significant difference was found about the return of the tumor in all age groups except maybe at the age group 50 to 69, where the percentage of women who didn't suffer from a tumor again is greater.
- Surgery of type "SMAC" was the mostly applied in most tumor return place categories especially in Bone, Liver, Local (Breast), and Lung. And in addition, patients who received a "SMAC" surgery as a first treatment did not suffer from tumor return in other places.
- Some patients who received radiotherapy at the first treatment suffered from tumor return in Bone and Beast.
- The hormonal therapy of type "Nolvadex" was mostly used at all stages, especially at stage II and III of the disease.
- Patients at stages II and III of cancer were more likely to stay alive within the period of five years while patients at advanced stages are exposed to death slightly more.

5.2 Summary of Breast Cancer Data Modeling

The most interesting results concerning the application of Association Rules and LDA are summarized in the following points;

- The breast cancer patients who had a surgery of type "SMAC" did not show a heritage factor of the disease in the family, and at the same time didn't receive a radiotherapy.
- The surgery of type "SMAC", implies receiving a radiotherapy, not receiving a second treatment, and staying alive for a period of five years at least with no presence of a heritage factor with good confidence values, and with a moderate chance of cancer return.
- The predictors; patient stage, radiotherapy and heritage factor all have a significant effect on the classification of the type of surgery to be performed. The Classification Function of the LD model can be used for future prediction purposes that is it can help medics diagnosing the next procedure to be taken.
- Patient age and radiotherapy were found to have important effect on the classification of chemotherapy type that involves 7-class discrimination, (NONE, FAC, FEC, Anthracyclin then Taxane, CMF, Taxane, or other). The chemotherapy of type "FAC" (Fluorouracil) had the largest probability of occurrence among the other types of therapy.
- Patient age, heritage factor, and blood test ER were found to have important effect on the type of hormonal therapy, (NO, Nolvadex, Nolvadex+Zoladex, Nolvadex+AL, or AL). The hormonal therapy of type "Nolvadex" had the largest probability of occurrence among the other types of hormonal therapy.
- Radiotherapy and having a 2nd treatment were found to have important effect on the return of the tumor, and the probability of tumor return to the patient is lower than that of none returning having the effect of receiving a radiotherapy and occasionally getting a second treatment.
- The predictors; cancer stage, type of surgery applied, hormonal therapy, the return of tumor and 2nd treatment were found to have important effect on the classification of survival of the patient for a period of 5 years. The probability of a patient to stay alive for a period of five years was found to be higher having the effect of cancer stage, type of surgery applied, and hormonal therapy, the return of tumor and 2nd treatment.

6. References

[1].**Tufféry, S., (2011)**: Data Mining and Statistics for Decision Making, John Wiley and Sons Ltd., UK.

[2].**Nisbet, R., Elder, J.,** and **Miner, G., (2009)**: Handbook of Statistical Analysis and Data Mining Applications, Elsevier Inc., USA.

[3].**Giudici, P., (2003)**: Applied Data Mining Statistical Methods for Business and Industry, John Wiley and Sons Ltd., UK.

[4].**Agrawal, R.**, **Imielinski, T.**, and **Swami, A.**, **(1993)**: Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD Conference, Washington, DC.

[5].**Johnson, R.** and**Wichern, D., (2007)**: Applied Multivariate Statistical Analysis, Pearson Education Inc., USA.

[6].**Everitt, B., (2005)**: An R and S-Plus Companion to Multivariate Analysis, Springer-Verlag London Limited, UK.