International Journal of Engineering & Technology, 7 (4.19) (2018) 806-811

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Limbless Medical Data Analyzing using CRISP Model a Case Study of UK Limbless Patients

Zahraa Shams Alden*[1], Ayad Hameed Mousa[2]

[1]*Faculty of Tourism Science, University of Kerbala, Iraq*
[2]*Collage of r Science, University of Kerbala, Iraq*
**Corresponding Author Email: zahraa.noor@uokerbala.edu.iq*

Abstract

Data mining, usually known as knowledge elicitation in the field of computer science databases, is the procedure to find out an important relationship, useful patterns in a huge amount of raw data. Besides, many sectors have adapted and used data mining in their applications such as healthcare and industry sector. In the healthcare sector, data mining can help in determining the probability of particular health cases in medical issues which the related variables pre-known as well as predicting future events. The availability of medical data for data mining usually exist in a raw data format, therefore, it needs for making ready and exploration to be willing to use. In the context of this paper, an analyzing of medical data was introduced to support prosthetics service centers to analyze find out the significant information from limbless medical cases, besides, in providing a comprehensive understanding of amputation and its types as well as the level of amputation. To ensure extract meaningful information from the intended data sets as well as to follow a systematic approach, the CRISP-DM model was adopted. The findings show the important and meaningful of the analyzing data using data mining modes.

Keywords: *Data Mining, Limbless Statistics, Data Mining Models, CRISP Model.*

1. Introduction

Data mining is the process of discovering interesting and useful patterns and relationships in large volumes of data. It's becoming significant to many applications which applied in different sectors, like industry, medicine, and markets[1, 2].
Technically, data mining technology is about gathering data from heterogeneous and homogeneous databases, cleaning and cleansing the intended data of any missing values, and finally applying some statistical queries to discover meaningful information from the intended data. There is a limitation in using data mining in the healthcare sector, these limitations emerged because of no enough methodologies which used to elicit meaningful information from medical data [3-6].
In the healthcare sector, medical data mining plays an important role by providing plenty of possibilities. In the same aspect, data mining serves a lot of medical cases by analyzing data from multi-sources whether being (heterogeneous or homogeneous) and extract meaningful information that can support the healthcare sector, and this will lead to increase the performance as well as the cost reducing. [1, 7].

1.1. State of Art

Due to the using case study in this paper, the Authors decide to review some of relevant cases. Besides, it is well established from a variety of case studies that in medical data mining are demonstrate a variety of data mining methods or approaches used to support the healthcare sector and its practitioners. this support includes disease risk prediction or cost estimation to a certain medical case. Table is highlighted some of related case studies:

Table 1: Medical Data Mining Case Studies

	Case Study	Description
1	HCC	In the context of this case study, an examination was conducted of the medical data for various patients with the hepatitis C virus (HCV) related chronic liver disease (CLD); 135 HCC, 116 cirrhotic selected patients without HCC and 64 selected patients with chronic hepatitis C, technically, the decision tree technique was used as HCC prediction's method [8].
2	CVD	In the context of this case study, a new algorithm (REMIND) was introducing (*reliable extraction and meaningful inference from non-structured data*) to filling the data gap in medical IT systems. Technically, this study has created REMIND platform by uses domain knowledge algorithms in order to support medical problems solving[9].
3	ESPHD	In the context of this case study, data warehouses for Heart Attack Prediction were built, a massive amount of heart disease medical data sets were analyzed. In that regard, a clustering algorithm was applied in discovering important patterns in terms of heart disease prediction[10].
4	Osteoporosis	In the context of this ongoing research study, an analyzing of medical data for patients who have suffered from osteoporosis disease. In that regard, this study helps GPs in early determining the intended disease[11].
5	Application of Data Mining	In the context of this ongoing research study, a predictive analysis of diabetic treatment was applied. In that regard, a regression-based was used as a data mining technique.[12].

In line with the above situation, it can be indicated that most authors in this research area have used classification technique (decision tree and prediction classification) as a method to extract meaningful information from a medical data set. In this regard, the classification algorithm was applied as a data mining technique.

2. Data Mining Pre-Processing

As discussed above, the major purpose of data mining is to extract and find out the meaningful information as well as useful patterns to solve the intended problem. Generally, the medical data sets which is collected from various data sources is usually "not clean" and must be processed and later input to data mining. Hence, that collected data should preprocessing, and the preprocessing steps are outlined:

1- Data preparations: Incudes data cleaning.
2- Data transformation: includes transforming a set of variables to a different format.
3- Data reduction: remove unnecessary variables in making decision.
4- Data exploration: incudes comprehend the relationship among the different the selected dataset variables.

3. Data Mining Frameworks

In line with the data mining projects, data mining process is very complex, thus, it needs to do several efforts need to coordinate these efforts to collect and manipulate medical data and interduce it to data mining. besides, in reviewing of data mining literature, many frameworks have existed which can be used as a blueprint for this study, to guide in mining the intended medical data. The most frequent frameworks have used are CRISP and SEMMA. in this paper, the outlining for SEMMA was given while CRISP was extensively highlighted.

4. SEMMA Data Mining Framework

SEMMA shorten for "*Sample, Explore, Modify, Model and Assess*". It introduced and adopted by the SAS Institute, developed a platform based on the intended methodology known as "*SAS Enterprise Miner Workstation"*. Refer to[13] for more detail.

5. CRISP DM Model

CRISP was created to support applying data mining technology in industry sector. CRISP provides a systematic, architecture guideline in using data mining technique and knowledge elicitation from various dataset. In the context of this paper, to elicit meaningful information from the selected dataset 'Limbless Statistics datasets". In addition of the CRISP-DM model was adapted[14]. Figure 1 visualize the CRISP model.

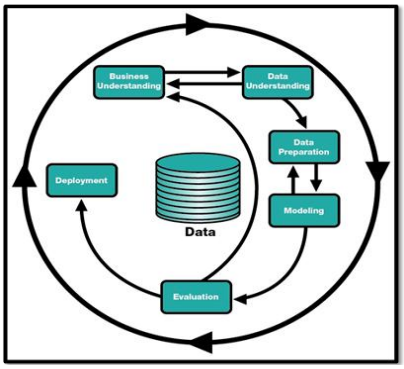


Fig. 1: Data Mining CRISP Model

As indicated from the Figure 1, the CRISP Framework composed of six stages start with business understanding, data understanding, data preparation, modeling, development, and finally, evaluation. in rest of this paper, the process of analyzing the intended data based on CRISP framework was highlighted.

6. Selection Dataset Variables

The Limbless Medical Dataset are including all intended patient's data who have lost limbs for various reasons, these reasons include accident, congenital limb absence present at birth, and relevant illness. In this regard, 39 attributes were determined; each of these attributes plays a significant role in the integration of intended dataset, and these data reached to 3259 records. The Limbless Medical Dataset consists of ten (10) classifications attributes that divided the selected dataset: three essential classifications that characterize some physical characteristics of the limbless patients, while the rest classifications that characterize further physical characteristics and relevant clinical information of the limbless patients[15].

The main classification elements are "side, site, and gender, and secondary classification attributes are all the other attributes – the level of amputation, the cause of amputation, date birth, date of amputation, date of referral, ethnicity background, center, and case category". Table 2 illustrate the frequency of the selected patients who lost their limb(s), the selected dataset was classified into the level of amputation with the general cause of amputation by using Chi-Square statistics measurement.

Table 2: Dataset Classification

General Cause	Congenital	Lower	Upper	Total
Congenital	7	71	35	113
Dysvascularity	0	1982	21	2003
Infection	0	238	8	246
Neoplasia	0	44	4	48
Neurological	0	56	1	57
No Data	0	467	45	512
Trauma	0	233	47	161
Total	7	3091	161	2359

"It can be concluded from the Table 2, there are serial variables which have ability to help and support an impact on data mining technology :like (1) general cause; (2) a specific level of amputation; and (3) limb loss for congenital causes. an interesting pattern can be concluded based on these variables; for instance, what is the most common cause of losing the upper limb or what is the most significant cause of losing the limb". The intended patients probably frequently visit doctors, and their relevant information can extract from "patient count per year" variable. in the context of this medical dataset, missing values can happen in several variables, for instance, the error was happened during the transition or some patients not giving all their information or the cause of amputation being unknown or not registered. Descriptive analyzing for the variables of the dataset would be obtainable to help comprehend the dataset.

7. Applying CRISP Data Mining Model

As discussed above, the data mining process will be based on the CRISP model, therefore, the CRISP steps should be followed. Thus, in data understanding step, limbless statistics dataset composed 3259 records of patients as well as 39 attributes spread over the period (2007-2012). besides, in order to increase the consistency of the selected data, some changed were applied. for instance, the difference between amputation and birth as well as days, week, and years numbers, and finally, the difference

between amputation and referral was conducted to support the data mining procedure. The next stage in crisp model is data preparation, in this stage limbless statistics dataset required to be analyzed in order to detect if there is a noise in the intended data or also determine if any missing values exist. Some data mining algorithms are used to clean, cleans the dataset and prepare these data to be as input to the next stage of crisp data miming model. In the same aspect, each data mining algorithm has its own way of data analyzing process.

In line with the above situation, in order to explore the whole data properties of the intended dataset, visualize the number of observations exists, how many variables can be used, as well as variables types and other relevant features such as file size and hostname created SAS Enterprise Guide was used. Figure 2 visualize them.

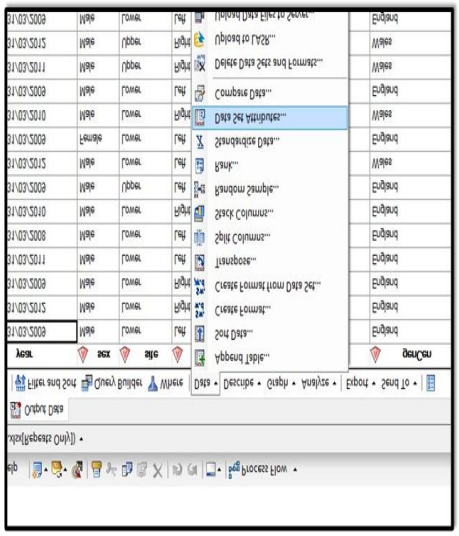
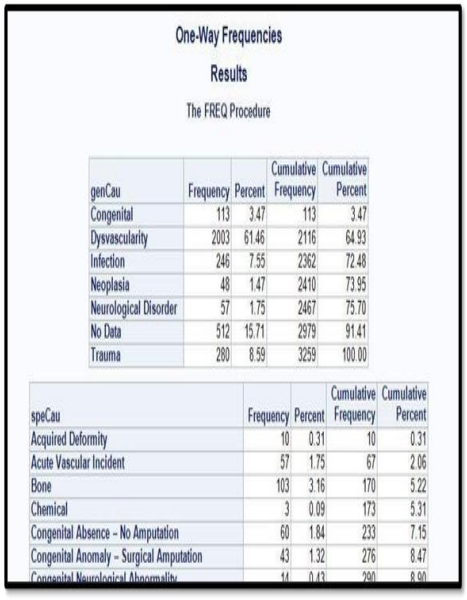


Fig. 2: Use SAS Enterprise Guide

In the same aspect, there is another feature of SAS Guide is that it has the ability to provide an analytic description for all attributes of the dataset, such as statistics summary, distribution analysis and One-Way Frequencies.

The benefit of using this tool is to describe the dataset and its variables. This data mining algorithm has the ability to compute the frequency of values of all attributes in the dataset, along with percent, cumulative frequency, and cumulative percent. Figure 3 visualizes One-Way Frequencies.



One-Way Frequencies

Results

The FREQ Procedure

genCau	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Congenital	113	3.47	113	3.47
Dysvascularity	2003	61.46	2116	64.93
Infection	246	7.55	2362	72.48
Neoplasia	48	1.47	2410	73.95
Neurological Disorder	57	1.75	2467	75.70
No Data	512	15.71	2979	91.41
Trauma	280	8.59	3259	100.00

speCau	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Acquired Deformity	10	0.31	10	0.31
Acute Vascular Incident	57	1.75	67	2.06
Bone	103	3.16	170	5.22
Chemical	3	0.09	173	5.31
Congenital Absence – No Amputation	60	1.84	233	7.15
Congenital Anomaly – Surgical Amputation	43	1.32	276	8.47
Congenital Neurological Abnormality	14	0.43	290	8.90

Fig. 3: One-Way Frequency Outcomes

As can be concluded for the Figure 3, the main cause of limb-losing is "dysvascularity" ("loss of blood pressure to limbs") because it includes the highest frequency of the cases which were recorded in the selected period 2007-2012; besides, it visualizes the proportion of samples with dysvascularity, which is total: 61.64%.

On the other aspect, the one-way frequencies table visualize the values of the existing dataset. Obviously, there is no missing value appeared in this analysis because already removed them.

Furthermore, the missing data of the selected dataset can be included during to computing of a specific variable value by "enabling it in the statistics option of the one-way frequency procedure window". For instance, suppose the selection of a cause attribute to be analyzed, and after enabling the option to show the frequency of missing values of this variable as shown in Figure 5) noticed that missing values in this data set are specified by "no data", while figure 5 illustrates the findings of the intended data mining procedure.

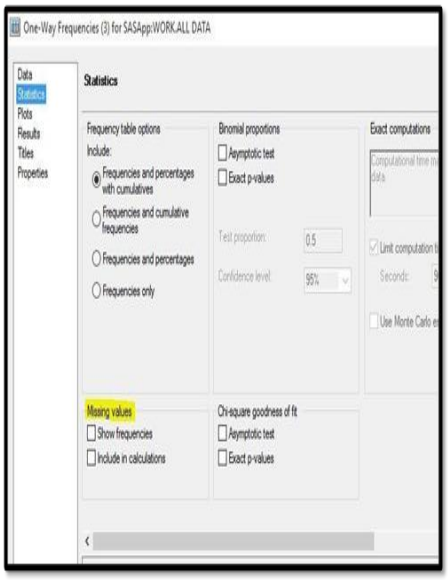
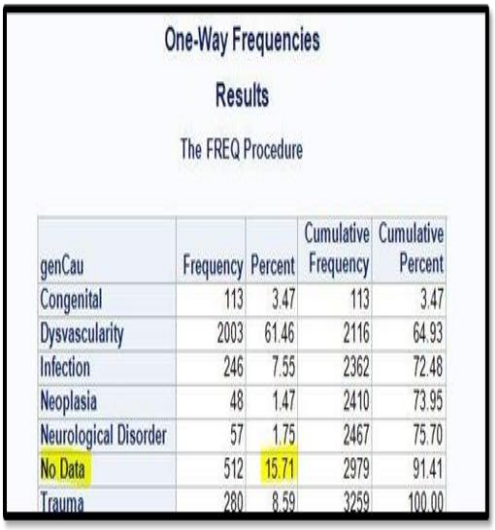


Fig. 4: One-Way Statistics Option Mining Values



One-Way Frequencies

Results

The FREQ Procedure

genCau	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Congenital	113	3.47	113	3.47
Dysvascularity	2003	61.46	2116	64.93
Infection	246	7.55	2362	72.48
Neoplasia	48	1.47	2410	73.95
Neurological Disorder	57	1.75	2467	75.70
No Data	512	15.71	2979	91.41
Trauma	280	8.59	3259	100.00

Fig. 5: One-Way Frequency Procedure Results

As discussed above, Figure 6 informs us there is 15% data in the selected dataset are either unknown or missing. In the context of this study, the missing values may be: some of the patients may do not prefer to share their information with other, or the error during the documented of the dataset may cause a missing value. While it can be concluded from Figure 6 that the distribution of main cause

attribute, because it visualizes the frequencies of missing value" no data".

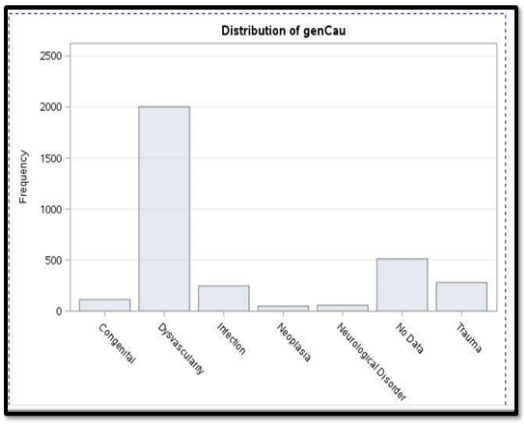


Fig. 6: Distribution of General Variables

In line with the above situations, the dysvascularity cause has the largest percentage, it is approximately which is more than 2,000 frequencies of the intended patients, while the missing data appears more than 500 frequencies of the intended patients in the number of the selected dataset values.

SAS Enterprise Guide provides another significant characteristic is distribution analysis, it is considered as descriptive statistics tool which displays (1) the missing values for the particular attribute in the selected dataset; (2) determine the number of extreme values of the selected dataset and the number of extreme observations of the selected dataset which are " the top values and low values"; and also (3) it can visualize the distribution of the selected dataset's attribute's values by means of variant charts and plots. In this selected dataset, a lot of missing data were detected for dissimilar variables and later these were represented by "no data".

In the same aspect, rows filtration tool was used in order to show the frequencies of missing values as well as to sort-filter the selected dataset values. In the context of this study, a "no data" value has been detected, and the data-rows number has been displayed which contain such values. Besides, the general cause be indicated by the variables that have missing values "no data" while the specific cause can be represented by "general ethnicity, specific ethnicity, data of amputation and date of referral". Furthermore, the outcome report that there are 1,832 patient's records that have missing values either in one or more of these variables in the selected dataset, as can clearly visualize in Figure 7.

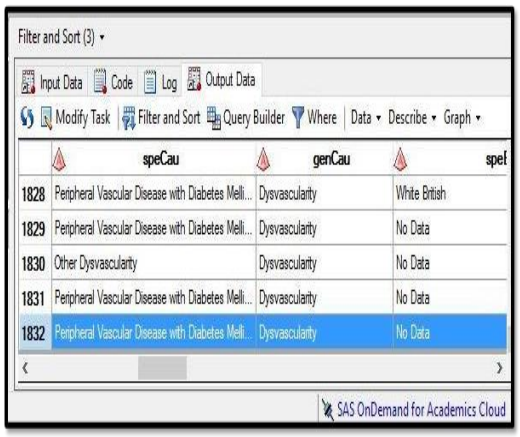


Fig. 7: Filter Procedure to Display Missing Values

Another advantage provided by SAS Enterprise Mine is used to help to discover the correlation between the dataset's variables and the target variable; this can be useful in integrating the hidden information. On the same aspect, the "state explore" mission which provided by SAS was used; it provides many descriptive statistics can help to analyze the intended data' for example CHI Square statistics which used to measure the correlation between dataset's variables and target variables. Chi-square is computed for categorical variables, therefore, by enabling the option of measure the interval variables also, SAS Enterprise Miner will help to distribute the variables into the default five bins. This enables Chi-square statistics to be computed for the binned variables.

The correlation type that used in this statistics measure is known as " Pearson's correlation by default", as well as SAS provides data mining analyzers the ability to use another correlation method "Spearman".

Figure 8 clearly shows the findings of that process when the general cause variable is the target, the particular cause variable in a strong correlation with it, since it has the highest "Chi-square" value.

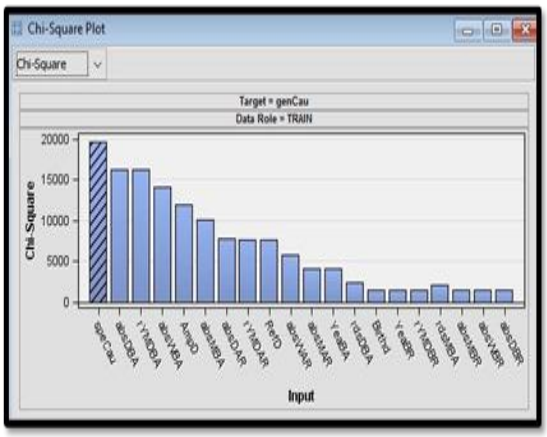


Fig. 8: Correlation between the Variables and Target Variable

On the other hand, the process of Chi-square provides another statistics measurement which calculates the amount of worth to all variables to the target variable. As clearly illustrated in Figure 9, the particular cause variable has the highest worth; as well as, the variable 'sex' has no impact on the target.

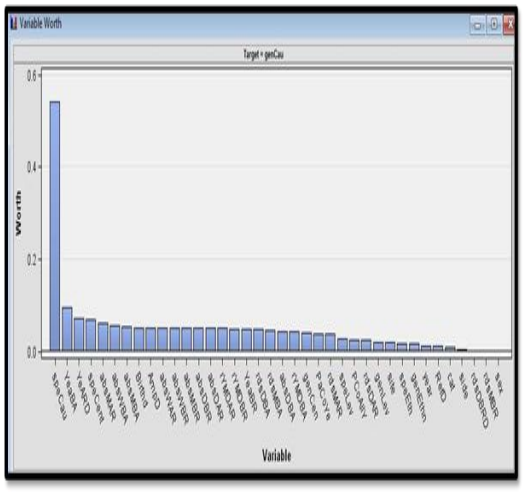


Fig. 9: Variable Worth of the Dataset

In order to explore the relationship or the correlation between these two variables, SAS Enterprise Guide also helps the data mining analyzer to access to different visualization data in the dataset, such as a bar chart. The particular cause variable was chosen to be represented by the bars and was grouped by the "general cause" variable (as clearly visualized in Figure 10), to see the frequency of each general cause value in correlation to the particular cause value.

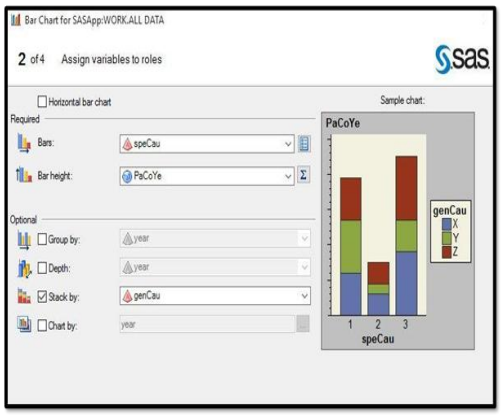


Fig. 10: Bar Chart Wizard for Cause Variable

As clearly indicated from figure 10, the bar height was specified by the number of patients per year and that it was briefed based on the number of general cause values to the particular cause variable by computed the number of patients per year.

Besides, figure 11 illustrate that intended patients who suffering from dysvascularity and use them as a general cause and peripheral vascular disease and use diabetes mellitus as the specific cause were the highest in number; next to that came selected patients who suffering from dysvascularity as the general cause and peripheral vascular disease ("no diabetes mellitus") as the particular cause.

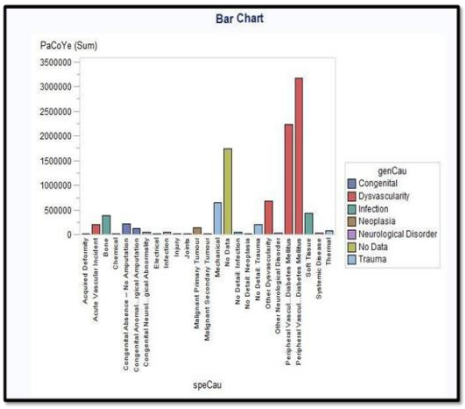


Fig. 11: Results of Bar Chart for Cause Variable

"As clearly illustrated in Figure 12, the most selected patients were of white ethnicity, white British to be exact, and that the next highest frequency recorded was missing values, which means that either the patients preferred to keep the ethnicity of their record or that a transition error had occurred".

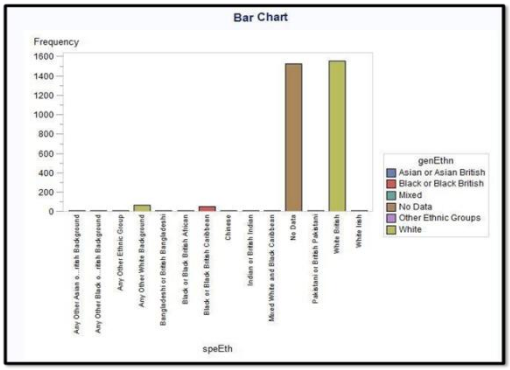


Fig. 12: Bar Chart of Specific Ethnicity Variable

On the other aspect, another exploration can apply to the selected dataset is to analyses the level of amputation. The bar chart was used for plotting the variables which relevant to the level values, that are generally level and particular level variables.

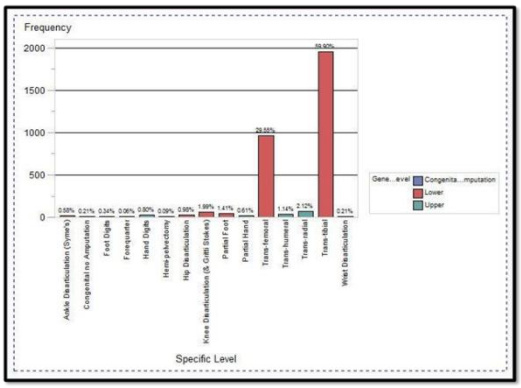


Fig. 13: Frequencies of Specific Level Bar Chart

It can be concluded from Figure 13 that the most patients who lost their limbs at a particular transtibial level ("an amputation of the lower leg between the ankle and knee"), while the second highest number of the selected patients is 'trans-femoral' level ("amputation of the lower leg between the knee and the hip").

Figure 14 classify the number of patients who had lost their limbs for variant causes based on male and female which recorded in the many medical centers in the selected dataset.

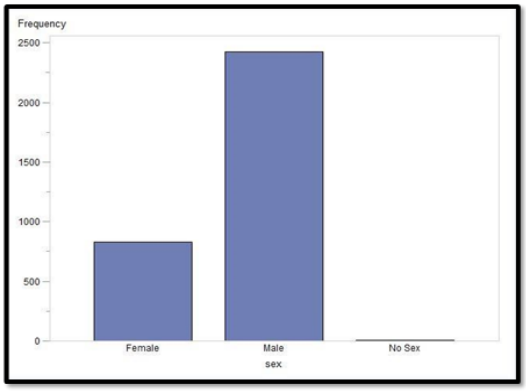


Fig.14: Bar Chart of Gender Variable

8. Conclusion

The main aim of this study was to analyses, extract the meaningful information and convey them to the knowledge, and also find out the significant relationships and correlations between the attributes of the selected dataset "Limbless Statistics". Besides, it hoped to prove that data mining techniques can be helping and support to use as analyzing tool medical dataset and can lead to discovering hidden knowledge.

During the applying the data mining techniques to the selected dataset, interesting patterns and knowledge were discovered and presented as a set of rules. And later was produced using a decision tree as a classification technique. Different statistical measurements, such as chi-square and one-way frequencies, were used to analyze the data set; and other statistics, such as average square error and misclassification rate, were conducted and used in order to measure the accuracy and goodness of the model.

This research is considered to be the first study to extract and analyze and discover the knowledge of the Limbless Statistics dataset and the process of data mining technique was successful at finding good-quality models to deal with medical dataset it can be used in prediction the level of amputation of a limb based on the cause of amputation, for example, and to estimate the number of years between amputation and referral date. As the data mining model and algorithms were applied to the intended healthcare

dataset, it was a successful attempt to prove the plays significantly role of these techniques in the healthcare industry. Because it deal with a huge amount of data which produced by IT healthcare applications of the different medical centers which provide medical dataset to extract, analyze the meaningful information to support and help to improve and prediction and discovering the knowledge as well as improve the quality of healthcare services.
SAS Enterprise Miner consider as one of extremely useful for preparing the selected dataset to be an input for data mining procedure and discovering the relationship and correlation between the attributes of the selected dataset.

References

[1] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, et al., "Data mining in healthcare and biomedicine: a survey of the literature," Journal of medical systems, vol. 36, pp. 2431-2448, 2012.

[2] A. H. Mousa, "DATA VIRTUALIZATION DESIGN MODEL FOR NEAR REA TIME DECISION MAKING IN BUSINESS INTELLIGENCE ENVIRONMENT," PhD PhD Thesis, Computer Science Department, Universiti Utara Malaysia, Malaysia, 2017.

[3] R. Canlas, "Data mining in healthcare: Current applications and issues," School of Information Systems & Management, Carnegie Mellon University, Australia, 2009.

[4] A. H. Mousa, N. Shiratuddin, and M. S. A. Bakar, "RGMDV: An approach to requirements gathering and the management of data virtualization projects," in INNOVATION AND ANALYTICS CONFERENCE AND EXHIBITION (IACE 2015): Proceedings of the 2nd Innovation and Analytics Conference & Exhibition, 2015, p. 030024.

[5] A. H. Mousa, N. Shiratuddin, and M. S. A. Bakar, "Process Oriented Data Virtualization Design Model for Business Processes Evaluation (PODVDM) Research in Progress," Jurnal Teknologi, vol. 72, 2015.

[6] A. H. Mousa and N. Shiratuddin, "Data Warehouse and Data Virtualization Comparative Study," in Developments of E-Systems Engineering (DeSE), 2015 International Conference on, 2015, pp. 369-372.

[7] D. Crockett and B. Eliason, "What is data mining in healthcare?," HealthCatalyst,[Online]. Available: https://www. health catalyst. com/data-mining-in-healthcare.[Accessed 30 November 2015], 2014.

[8] D. A. E. H. Omran, A. H. Awad, M. A. El, R. Mabrouk, A. F. Soliman, and A. O. A. Aziz, "Application of Data Mining Techniques to Explore Predictors of HCC in Egyptian Patients with HCV-related Chronic Liver," Asian Pacific Journal of Cancer Prevention, vol. 16, pp. 381-385, 2015.

[9] R. B. Rao, S. Krishnan, and R. S. Niculescu, "Data mining for improved cardiac care," ACM SIGKDD Explorations Newsletter, vol. 8, pp. 3-10, 2006.

[10] S. B. Patil and Y. Kumaraswamy, "Extraction of significant patterns from heart disease warehouses for heart attack prediction," IJCSNS, vol. 9, pp. 228-235, 2009.

[11] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, and W. E. Hammond, "Medical data mining: knowledge discovery in a clinical data warehouse," in Proceedings of the AMIA annual fall symposium, 1997, p. 101.

[12] A. A. Aljumah, M. G. Ahamad, and M. K. Siddiqui, "Application of data mining: Diabetes health care in young and old patients," Journal of King Saud University-Computer and Information Sciences, vol. 25, pp. 127-136, 2013.

[13] D. L. Olson and D. Delen, Advanced data mining techniques: Springer Science & Business Media, 2008.

[14] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000, pp. 29-39.

[15] M.-L. Antonie, O. R. Zaiane, and A. Coman, "Application of data mining techniques for medical image classification," in Proceedings of the Second International Conference on Multimedia Data Mining, 2001, pp. 94-101.