ISSN 1870-4069

Oncological Analysis Using Data Mining

Doricela Gutiérrez Cruz[1], Ricardo Rico Molina[1], Liliana Rodríguez Páez[1], Karla Vilchis Hernández[1], Bernardo Soto Rivera[2], Yaroslaf Albarrán Fernández[1], Alma Gutierrez Cruz[2], Mónica Ortega[1]

[1] Unidad Académica Profesional Nezahualcóyolt UAEM,
Mexico

[2] Hospital Regional 1º de octubre,
Mexico

gutierrezcruzdo@yahoo.com.mx, rico_molina@hotmail.com

Abstract. Data mining is a technique that involves the application of specific algorithms, which generate a list of patterns based on large volumes of information that is useful for decision making in wide fields of application, as detection of patterns in diseases. This paper proposes to carry out a cancer analysis by means of data mining techniques, applied on a sample of 3365 cases, using association algorithms as Apriori and J48 classification algorithm. The data were obtained from clinical reports of patients from the Palliative Care Unit of the Hospital Regional 1o de Octubre. Once the data mining process was completed, it was possible to identify some of the most relevant types of cancer, age as characteristic factor in the emergence of diseases, as well as the sex affected.

Keyword: Data mining, cancer, apriori, J48 classification algorithms.

1. Introduction

The increasing volume and variety of information computerized in digital databases has grown in the last decade [3], a considerable part of this information is historical, it represents transactions or situations that have occurred and it may be useful for understanding future information [14, 15,16].

Technological development both in the computer and data transmission fields, promotes a better handling and storage of information; as is the case of Data Mining (DM), which can be defined as the consistent use of specific algorithms that generate a list of patterns based on preprocessed data that is useful for decision-making. [17, 10, 13]. That list of patterns is closely related to statistics, using sampling techniques and data visualization. Research and development to analyze large volumes of data became increasingly necessary, and they can be done based on files. Even if the advantages increase when there are large volumes of data [4], to discover knowledge from this huge volume of data is a challenge in itself. DM is an attempt to make sense of the explosion of information that currently can be stored [5]. The phase of data mining is the representation of the type of model obtained. It focuses on searching, which will have one or more forms of representation, depending on the type of model obtained [6]. The data analysis can provide real knowledge to help in decision making [7].

Cancer is one of the main causes of death in the world. In 2012 about 14 million new cases and 8.2 millions of chronic and degenerative deaths related to this condition were reported; 30% of those deaths were due to dietary and behavioral risk factors. On the other hand, over 60% of total annual new cases in the world occur in Africa, Asia, and Central and South America, these regions represent 70% of cancer deaths in the world [1].

The historical behavior of cancer mortality has been on an upward trend, and international registers allow the problem that cancer represents to be visualized, however existing information in Mexico is limited, therefore it is difficult to have access to it to determine the real impact of cancer on health more accurately [2].

It should be noted that the methodology DM is part of a process called "Knowledge Discovery in Databases" (KDD), which indicates the necessary steps to reduce risks in the search of knowledge models when applying DM techniques. For example, in the KDD process the data require a substantial preprocessing to be modeled (cleaning and preparation) [9, 10, 11].

In the clinical setting the DM is helpful for the identification and diagnosis of diseases. Likewise, it has significance for the discovery of possible interrelationships between diverse diseases [8]. In medical applications, where it is not possible to ignore the importance of the temporal component, data mining techniques have acquired major significance [11]. In this context, the aim of this paper is to characterize and classify the types of cancer and their impact on the population using data mining techniques, in the expectation of finding underlying relationships that cannot be identified by a classic statistical treatment.

2. Case Study

The Hospital Regional 1o de Octubre (HR1O) was inaugurated on December 5th, 1974 and the population attended is diverse: urban, suburban, rural and marginalized.

In accordance with the above, regarding our area of influence, the population figures are as follows: covered population 2172,49, registered population 1086,124, user population 651,674. Approximately 50% of the beneficiaries are workers and pensioners, and nearly a quarter are children under the age of 18.

Since each year 10 million people suffer from pain caused by a disease, and cancer represents 5.5 million of them, palliative care practices were implemented in the Hospital eight years ago; they are focused on controlling pain and attending the psychosocial aspects of the patient and the family members who support him during the process.

3. Experimental Development

To carry out this study, the general process KDD was applied in order to discover patterns and relationships in data that can be used to make valid predictions [9, 17], the universe of the study were 3365 cases of cancer; while the WEKA software was used for the analysis and construction of the data mining model (http://www.cs.waikato.ac.nz/ml/weka/).

3.1 Data Selection

To conduct this study, the representative variables to address the problem: *sex*, *year*, *age*, *diagnosis* and *type of beneficiary*, were taken from the clinical reports provided by the palliative care unit of the hospital.

3.2 Pre-Processing Data

This step consisted of a data cleansing, in order to get quality patterns; that is: without outliers or null values. The data obtained from clinical reports were analyzed to identify inconsistencies using the WEKA system. This process was only performed in the data related to diagnosis, as these were either oncological or not oncological.

The description of the five most important attributes for conducting this research is shown in Table 1, while Table 2 shows the attributes of the different types of cancer addressed in this research.

Table 1. Description of the most important attributes.

Attribute	Definition	Description
S	Sex	Gender of patients
Y	Year	Year when the diagnostic is reported to the unit
Dx	Cancer	Type of cancer suffered by the patient

Table 2. Description of the attributes associated with different types of cancer.

Attribute	Definition
ca co	heart cancer
ca ce	brain cancer
ca ab	abdominal cancer
ca ceu	cervical cancer
ca a	appendix cancer
ca c v	spinal cancer
ca cvcr	spinal and rectal cancer
ca es	esophagus cancer
ca e	stomach cancer
ca h	liver cancer
ca av	cancer of ampulla of Vater
ca meo	bone marrow cancer
ca ov	ovarian cancer
ca pi	skin cancer
ca p	prostate cancer
ca pu	lung cancer
ca r	rectal cancer
ca ri	kidney cancer
ca te	testicular cancer

Attribute	Definition
ca v	bladder cancer
ca sl	cancer of lymphatic system
ca m	breast cancer
ca o	oral cancer
ca os	bone cancer
ca pa	pancreatic cancer
le	leukemia
ca mm	multiple myeloma
sa te b	soft tissue sarcoma

3.3 Analysis of Data

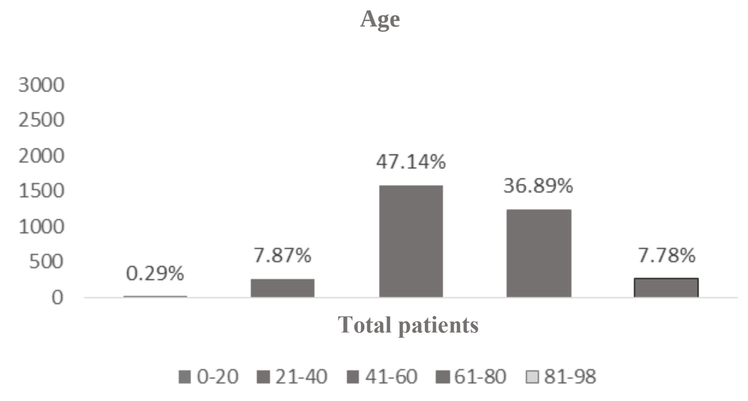
The search and finding of unsuspected patterns is carried out during the data mining stage where discovery tasks, such as classification [22, 23], *clustering* [18, 19], *sequential patterns* [20], *associations* [21], among others, are applied.

One of the differences in gender condition was observed in a study on disease burden with higher rates of morbi-mortality in men [33]. Higher mortality rates for men or women, depending on the location of the tumor, were identified in other studies carried out in work places, whereas women were the most affected in other studies [33, 34].

Age is a characteristic factor in the onset of chronic diseases [32], since at the time of the diagnosis it was evaluated as a prognostic element, and it was reported that elderly patients have lower life expectancy, even in early stages, when compared with younger patients.

The effect of age represents a change in rates associated with age, which is important given that the onset of chronic diseases is usually greater with increasing age [31, 35].

Figure 1 shows the frequency analysis of some of the variables studied. As it can be denoted, during 2009-2015, 61% of the patients reported by the palliative care unit of HR1O were women, and the remaining 33% men. Concerning the types of cancer and their incidence reported during the same period, breast cancer had an incidence of case of 19%, followed by of spinal and rectal cancer with 9%, lung and prostate cancer with 8%, and cervical and bone marrow cancer with 7%.



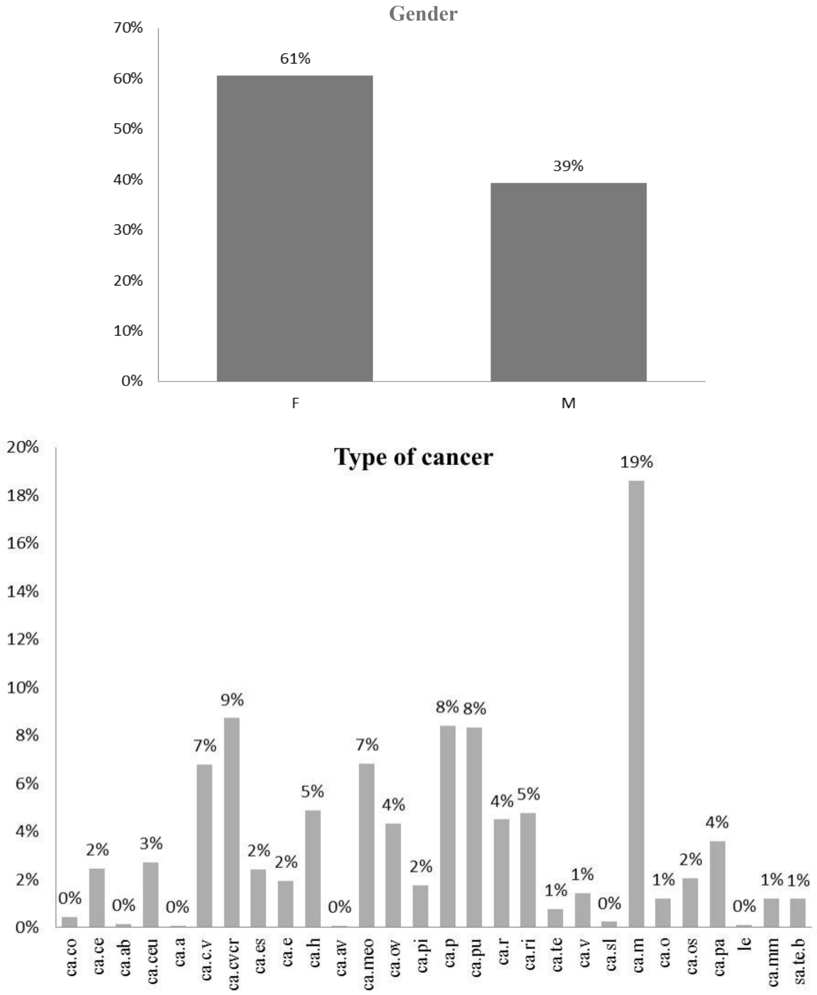


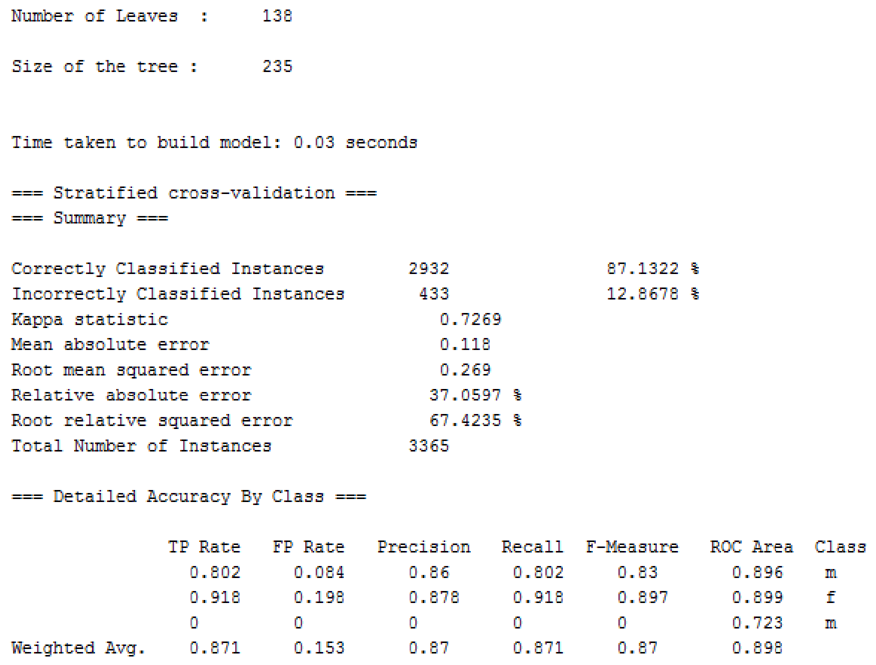
Fig. 1. Frequency analysis of the gender, age and types of cancer with highest incidence during 2009-2015 (Own elaboration, 2016).

3.4 Classification

For this activity, as mentioned, WEKA software was used. It was selected owing to the data included in the data set are mostly categorical, and because it uses decision trees. From the available algorithms, the algorithm J48 corresponding to the C4.5 algorithm was used [30]. Default specifications of WEKA and the stratified cross-validation method were used for its execution, as well.

To generate the tree an attribute as root must be selected, and a branch with each of the possible values of that attribute must be created; this process is carried out with each resulting branch. An attribute to continue dividing must be selected on each node, to do this the attribute that best separate the examples according to the class is selected.

Figure 2 shows that, by gender, most of the patients treated for some type of cancer during the reported period were women (91.8%).



```
Number of Leaves  :     138

Size of the tree :      235


Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2932               87.1322 %
Incorrectly Classified Instances       433               12.8678 %
Kappa statistic                          0.7269
Mean absolute error                      0.118
Root mean squared error                  0.269
Relative absolute error                 37.0597 %
Root relative squared error             67.4235 %
Total Number of Instances             3365

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.802     0.084     0.86      0.802     0.83       0.896    m
                 0.918     0.198     0.878     0.918     0.897      0.899    f
                 0         0         0         0         0          0.723    m
Weighted Avg.    0.871     0.153     0.87      0.871     0.87       0.898
```

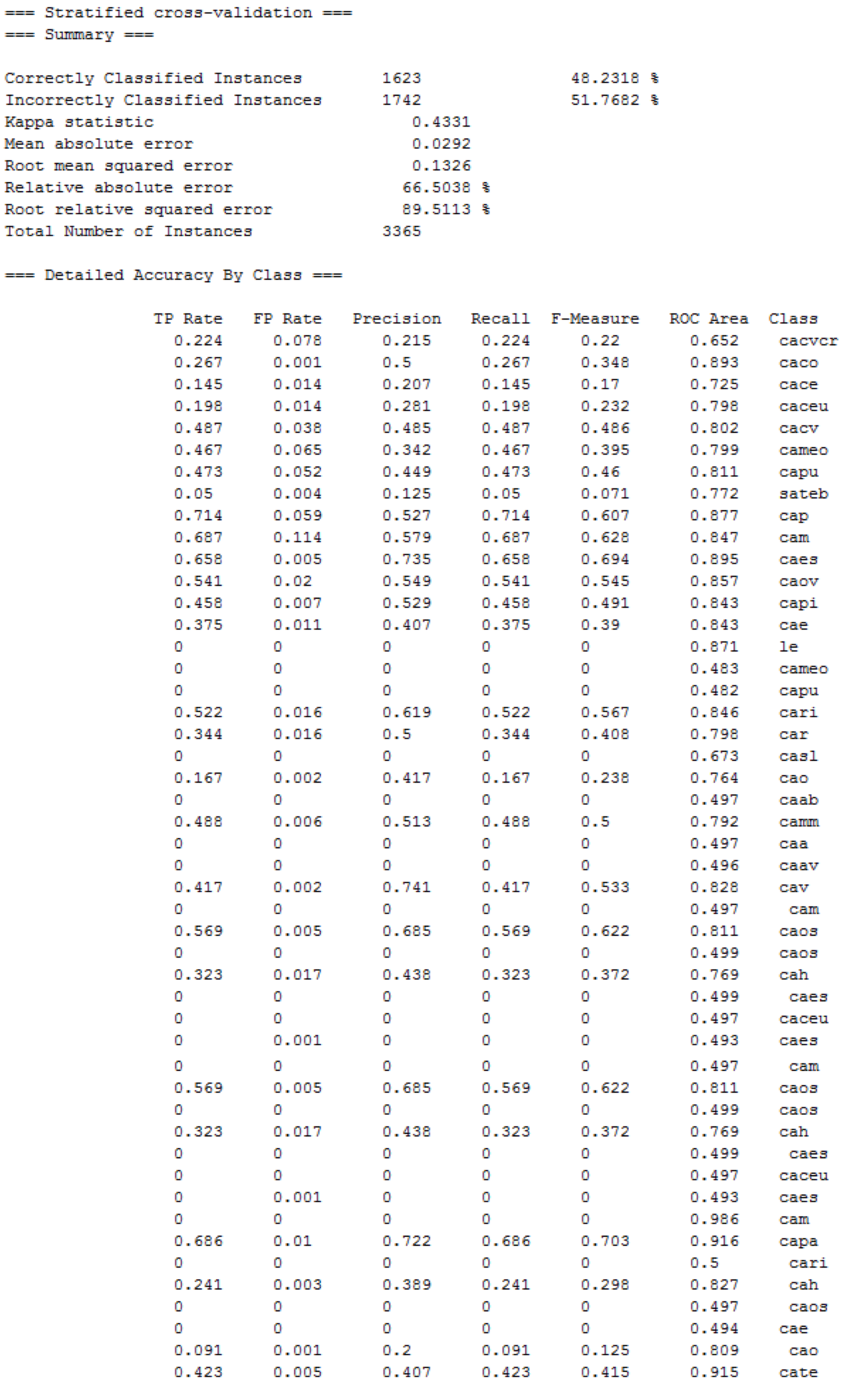
Fig. 2. Classification of patients by sex suffering from cancer during 2009-2015 (Own elaboration, 2016).

Figure 3 shows that the types of cancer with highest incidence during the reported period were: prostate cancer with 71.4%, breast cancer 68.7%, pancreatic cancer 68.6%, esophageal cancer 65.7%, whereas the type with the lowest incidence was eye cancer with 9.1%.

Subsequently, the Apriori algorithm was applied to mine data, which were used to obtain association rules between sets of a data repository [36]. It is shows the results of 3365 cases of cancer:

Apriori algorithm association rules

1. Dx=cacvcr 172 ==> Ano=2009 172   conf:(1)
2. Sexo=f Dx=cacvcr 118 ==> Ano=2009 118   conf:(1)
3. Edad=41_60 Dx=cameo 109 ==> Ano=2009 107   conf:(0.98)
4. Sexo=f Dx=cameo 116 ==> Ano=2009 113   conf:(0.97)
5. Dx=cameo 169 ==> Ano=2009 164   conf:(0.97)
6. Edad=41_60 Dh=2 221 ==> Sexo=f 213   conf:(0.96)
7. Edad=41_60 Dh=2 Ano=2009 186 ==> Sexo=f 179   conf:(0.96)
8. Dh=2 327 ==> Sexo=f 310   conf:(0.95)
9. Dh=2 Ano=2009 278 ==> Sexo=f 263   conf:(0.95)
10. Dx=cam 130 ==> Sexo=f 119   conf:(0.92)



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1623               48.2318 %
Incorrectly Classified Instances      1742               51.7682 %
Kappa statistic                          0.4331
Mean absolute error                      0.0292
Root mean squared error                  0.1326
Relative absolute error                 66.5038 %
Root relative squared error             89.5113 %
Total Number of Instances             3365

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.224     0.078      0.215     0.224     0.22       0.652    cacvcr
                 0.267     0.001      0.5       0.267     0.348      0.893    caco
                 0.145     0.014      0.207     0.145     0.17       0.725    cace
                 0.198     0.014      0.281     0.198     0.232      0.798    caceu
                 0.487     0.038      0.485     0.487     0.486      0.802    cacv
                 0.467     0.065      0.342     0.467     0.395      0.799    cameo
                 0.473     0.052      0.449     0.473     0.46       0.811    capu
                 0.05      0.004      0.125     0.05      0.071      0.772    sateb
                 0.714     0.059      0.527     0.714     0.607      0.877    cap
                 0.687     0.114      0.579     0.687     0.628      0.847    cam
                 0.658     0.005      0.735     0.658     0.694      0.895    caes
                 0.541     0.02       0.549     0.541     0.545      0.857    caov
                 0.458     0.007      0.529     0.458     0.491      0.843    capi
                 0.375     0.011      0.407     0.375     0.39       0.843    cae
                 0         0          0         0         0          0.871    le
                 0         0          0         0         0          0.483    cameo
                 0         0          0         0         0          0.482    capu
                 0.522     0.016      0.619     0.522     0.567      0.846    cari
                 0.344     0.016      0.5       0.344     0.408      0.798    car
                 0         0          0         0         0          0.673    casl
                 0.167     0.002      0.417     0.167     0.238      0.764    cao
                 0         0          0         0         0          0.497    caab
                 0.488     0.006      0.513     0.488     0.5        0.792    camm
                 0         0          0         0         0          0.497    caa
                 0         0          0         0         0          0.496    caav
                 0.417     0.002      0.741     0.417     0.533      0.828    cav
                 0         0          0         0         0          0.497     cam
                 0.569     0.005      0.685     0.569     0.622      0.811    caos
                 0         0          0         0         0          0.499    caos
                 0.323     0.017      0.438     0.323     0.372      0.769    cah
                 0         0          0         0         0          0.499     caes
                 0         0          0         0         0          0.497    caceu
                 0         0.001      0         0         0          0.493    caes
                 0         0          0         0         0          0.497     cam
                 0.569     0.005      0.685     0.569     0.622      0.811    caos
                 0         0          0         0         0          0.499    caos
                 0.323     0.017      0.438     0.323     0.372      0.769    cah
                 0         0          0         0         0          0.499     caes
                 0         0          0         0         0          0.497    caceu
                 0         0.001      0         0         0          0.493    caes
                 0         0          0         0         0          0.986    cam
                 0.686     0.01       0.722     0.686     0.703      0.916    capa
                 0         0          0         0         0          0.5       cari
                 0.241     0.003      0.389     0.241     0.298      0.827     cah
                 0         0          0         0         0          0.497     caos
                 0         0          0         0         0          0.494    cae
                 0.091     0.001      0.2       0.091     0.125      0.809     cao
                 0.423     0.005      0.407     0.423     0.415      0.915    cate
```

Fig. 3. Classification of types of cancer reported during 2009-2015 (Own elaboration, 2016).

The results obtained through algorithms of apriori classification include 10 association rules, each and every one of them with a certainty that goes from 0.92 to 1. It can be stressed that Rule 3 shows that, in 2009, there were 2 related attributes: age "41-60" years, and “cameo” diagnosis; whereas Rule 4 registers that female gender and “cameo” diagnosis were more significant during the same period; Rule 6 shows that beneficiary patients type 2 between 41 and 60 years were women; Rule 10 exposes that the type of cancer "Cam", is associated with the attribute "woman". Finally, there is an association between Rules 4 and 10, related to diagnosis and sex attributes, the latter being women.

4. Conclusions

Based on the analysis, it can be said:

1. The latest report from the World Health Organization shows that the burden of cancer is increasing at an alarming rate and underlines the need for effective prevention strategies to curb the disease. Also, that the most frequent types of cancer are different in men and women [33, 37, 38, 39]. This research confirms the latter, since from the 3365 reported patients, 91.8% of the women were the most affected by some type of cancer, compared to 80.2 % of men.
2. Currently, the emergence of new chronic diseases has impacted many people around the world, the predominant types of cancer worldwide with the highest annual number of deaths include: lung, liver, stomach, colon and breast [37]. The predominant types of cancer in the palliative care unit of HR1O were: prostate cancer 71.4%, breast cancer 68.7%, pancreatic cancer 68.6%, and esophageal cancer 65.7%.
3. Age is a factor in the onset of chronic diseases, the older the patient, the greater the incidence of diseases [32, 40], as it was shown, the most incidences of cancer are reported in patients between 40 and 60 years.

Among the existing chronic diseases, the study of cancer has become very important; since it has been a worldwide public health problem for several decades. The costs of cancer burden are even hurting economies, besides exerting pressure on health care systems.

Cancer is one of the chronic degenerative disease with highest incidence among adults [1, 2], and currently one of the leading causes of morbid-mortality among the population younger than 20 [3]. The most representative factors are the male gender and the group of 70 years or older. The most common locations of cancer in both genders were trachea, bronchus and lung, followed by prostate in man cancer and breast cancer and uterus (body and neck) in women [4].

In general, the lifestyle habits that most influence on the risk of suffering from cancer are smoking and obesity, in both genders; whereas, hormonal aspects in women exclusively [5]

Acknowledgements. The authors thank the information, monitoring and facilities granted to this research by Hospital Regional 1° de Octubre to the “Universidad

Autónoma del Estado de México" "Unidad Académica Profesional Nezahualcóyotl", and appreciate the support given by Eng. Luis Antonio Gutierrez Perez.

References

1. Organización Mundial de la Salud (OMS): Mortalidad a nivel mundial: Las 10 causas principales de defunción en el mundo. http://www who int/mediacentre/factsheets/fs310/es/index2 html (2015)
2. Rizo, P., González, A., Sánchez, F., Murguía, P.: Tendencia de la mortalidad por cáncer en México: 1990-2012. Evidencia Médica e Investigación en salud, 8(1), pp. 5–15 (2015)
3. Hernández, J., Ramírez, J., Ferri, C.: Introducción a la Minería de Datos. Departamento de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, Madrid (2004)
4. Chiotti, S. M., Cidisi, O.: Minería de Datos en Base de Datos de Servicios de Salud – UTN – FRSF. Ingar UTN- CONICET (2013)
5. Riquelme, J., Ruiz, R., Gilbert, K.: Minería de Datos: Conceptos y Tendencias. Departamento de Lenguajes y Sistemas Informáticos, Universidad de Sevilla (2006)
6. Quesada, Y., Wong, D., Rosete, A.: Minería de Datos aplicada a la Gestión Hospitalaria. En: 14 Convención Científica de Ingeniería y Arquitectura, CUJAE (2008)
7. Marcano, Y. J., Talavera, R.: Minería de Datos como soporte a la toma de decisiones empresariales. Opción, Año 23, No. 52, pp. 104–118 (2007)
8. Zamarrón, C., García, V., Calvo, U., Pichel, F., Rodríguez, J. R.: Aplicación de la Minería de Datos al estudio de las alteraciones respiratorias durante el sueño. Hospital Clínico Universitario de Santiago de Compostela, Servicio de Neumología (2006)
9. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, USA (2006)
10. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. Cambridge, MA: The MIT Press (2001)
11. Hernández, J., Ramírez, M. J., Ferri, C.: Introducción a la Minería de Datos. Madrid, Pearson Educación, S A (2004)
12. Fayyad, U. M., Grinstein, G., Wierse, A.: Information Visualization in data Mining and Knowledge discovery. Morgan Kaufmann, Harcourt Intl. (2001)
13. Frawley, W., Piatetsky-Shapiro, G., Matheus, C.: Knowledge discovery in databases: An Overview. AI magazine, Vol. 13, No. 3, pp. 57 (1992)
14. Simon, A.: Data Warehouse, data Mining and OLAP. John Wiley & Sons, USA (1997)
15. Berson, A., Smith, S. J.: Data Warehouse, Data Mining & OLAP. McGraw Hill, USA (1997)
16. White, C. J.: IBM Enterprise Analytics for the Intelligent e-business. IBM Press, USA (2001)
17. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, Vol. 39, No. 11, New York, USA, pp. 27–34 (1996)
18. Zhang, T., Ramakrishnan, R., Linny, M.: An efficient data clustering method for very large databases. In: ACM SIGMOD International Conference on Management of Data Montreal, Canada (1996)
19. Agrawal, R., Srikant, R.: Mining sequential patterns. In: The 11th International Conference on Data Engineering, ICDE, Taipei, Taiwan (1995)
20. Agrawal, R., Ghosh, S., Imielinski, T., Iyer, B., Swami, A.: An interval classifier for database mining applications. In: Proceedings VLDB Conference, Vancouver, Canada (1992)

21. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: VLDB Conference, Santiago de Chile, Chile (1994)
22. Witten, I., Frank, E.: Data mining: Practical machine learning tools and techniques with Java Implementations. Morgan Kaufman Publishers, San Francisco, CA, USA (2000)
23. Ng, R., Han, J.: Efficient and effective clustering method for spatial data mining. In: The 20th International Conference on Very Large Data Bases (VLDB 94), Santiago de Chile, Chile (1994)
24. Hernández, E., Lorente, R.: Minería de datos aplicada a la detección de Cáncer de Mama. Universidad Carlos III, Madrid, España, http://www.it.uc3m.es/jvillena/irc/practicas/08-9/14 pdf (2009)
25. Yépez, M. C.: Supervivencia de mujeres con cáncer de cuello uterino. Municipio de Pasto Revista Universidad y Salud, Vol.2, No.14, pp 7–18 (2011)
26. Mora, R.: El papel de la minería de datos en la detección y diagnóstico de cáncer. Universidad de Salamanca, Salamanca, España, http://sistemaminergescon.blogspot.com (2011)
27. Febles, J. P., González, A.: Aplicación de la minería de datos en la bioinformática. ACIMED, Vol. 10, No. 2, pp. 69–76, http://scielo.sld.cu/scielo php?script=sci_arttext&pid=S1024-94352002000200003&lng=es&nrm=iso (2002)
28. Dávila, F., Sánchez, Y.: Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas. RCIM, Vol. 4, No. 2, pp. 174–183, http://scielo.sld.cu/scielo php?script=sci_arttext&pid=S1684-18592012000200007&lng=es&nrm=iso (2012)
29. Camacho, S. S.: Método Heurístico para el Diagnóstico de Cáncer de Mama basado en Minería de Datos. Revista PGI, No. 1, pp. 97–101, http://www.revistasbolivianas.org bo/scielo php?script=sci_arttext&pid=S3333-77772014000100020&lng=es&nrm=iso (2014)
30. Basilio, A.: Aprendizaje Automático: conceptos básicos y avanzados, Aspectos prácticos utilizando el software Weka. Vol. 10, pp. 84–832 (2006)
31. Mery, C.M., Pappas, A.N., Bueno, R., Colson, Y.L., Linden, P., Sugarbaker, D.J.: Similar long-term survival of elderly patients with non-small cell lung cancer treated with lobectomy or wedge resection within the surveillance. Epidemiology and end results database, 128, pp. 237–245 (2005)
32. González, J.R., Llorca, F.J., Moreno, V.: Algunos aspectos metodológicos sobre los modelos edad-período-cohorte. Aplicación a las tendencias de mortalidad por cáncer, Nota Metodológica (2002)
33. Rodríguez, P., Fernández, J., Delgado, L., Garrote, I., Morales, J.M., Achiong, F.J.: Mortalidad por cáncer y condición de género. Rev méd electrón (2009)
34. Borràs, J.: La perspectiva del género en el cáncer: Una vision relevante necesaria. Universidad de Barcelona (2015)
35. Organización Mundial de la Salud (OMS): Salud Mundial: Retos actuales
36. La salud de los adultos en peligro: El ritmo de las mejoras disminuyen y las diferencias se acentúan. http://www.who.int/whr/2003/chapter1/es/index3.html (2003)
37. Paresh, T., Yogesh, G.: Using Apriori with WEKA for Frequent Pattern Mining. International Journal of Engineering Trends and Technology (IJETT), 12(3), pp. 121–131 (2014)
38. Stewart, B.W., Wild, C.P.: World Cancer Report 2014. International Agency for Research on Cancer (2014)
39. Menvielle, G., Luce, D., Goldberg, P., Leclerc, A.: Smoking, alcohol drinking, occupational exposures and social inequalities in hypopharyngeal and laryngeal cancer. Int J Epidemiol, 33(4), pp. 799–806 (2004)
40. Hansen, R.P., Olesen, F., Sorensen, H.T., Sokolowski, I., Sondergaard, J.: Socioeconomic patient characteristics predict delay in cancer diagnosis: a Danish cohort study. BMC Health Serv. Res., 8(49) (2008)

41. Serrano-Olvera, A., Gerson, R.: Supervivencia en relación con la edad en cáncer pulmonar de células no pequeñas. Gac. Méd. Méx., Vol.145, No.1 (2009)