Generación de conocimiento en la calidad de vida para pacientes terminales oncológicos con minería de datos.

José Sotelo Román, Dra. Doricela Gutiérrez Cruz, Dr. Israel Gutiérrez González, Dr. Ricardo Rico Molina, Dr. Ricardo Segura Martínez, Dra. Rebeca Gutiérrez Cruz.

[a] Centro Universitario UAEM Nezahualcóyotl, jsotelor.uaem@gmail.com, dgtutierrezcr@uaemex.mx, Estado de México, Méx.
[b] Hospital Regional 1° de Octubre, rebekgut@hotmail.com, Ciudad de México, Méx.

Resumen

En la presente investigación se muestra un análisis descriptivo respecto a la incidencia en síntomas de pacientes con diversos tipos cáncer y las posibles correlaciones entre variables como sexo, edad, tipo de cáncer y síntoma presentado. Además de ser evaluados con los estándares de cuidados paliativos enfocados al diagnóstico oncológico dejando en una escala los grados de intensidad en la clasificación del síntoma, donde la sintomatología presentada otorga mayor precisión en la evaluación de los datos siendo la ansiedad y nauseas los síntomas con más presencia dentro del análisis.
Los datos fueron proporcionados por el Hospital Regional 1° de Octubre en un periodo de estudio que abarca los años 2010-2021 de un total de 10,925 datos que integran un total de 576 expedientes correspondientes a pacientes con este tipo de diagnóstico. Tales datos fueron tratados con el modelo de minería de datos KDD realizando todas las fases que lo conforman. Para su clasificación se manejó el software Weka donde fueron utilizados diversos algoritmos para detallar un análisis de los datos.

Palabras clave— Cáncer, KDD, Minería de datos, Síntoma, WEKA.

Abstract

The present research shows a descriptive analysis regarding the incidence of symptoms of patients with various types of cancer and the possible correlations between variables such as sex, age, type of cancer and symptom presented. In addition to being evaluated with the standards of palliative care focused on the oncological diagnosis, leaving on a scale the degrees of intensity in the classification of the symptom, where the symptomatology presented gives greater precision in the evaluation of the data, being anxiety and nausea the symptoms with more presence within the analysis.
The datawere provided by the Regional Hospital 1° de October in a study period that covers the years 2010-2021 of a total of 10,925 data that integrate a total of 576 files corresponding to patients with this type of diagnosis. Such data were treated with the KDD mining model performing all the phases that make it up. For its classification, the Weka software was used, where various algorithms were used to detail an analysis of the data.

Keywords— Cancer, KDD, Data mining, Symptom, WEKA.

1. INTRODUCCIÓN

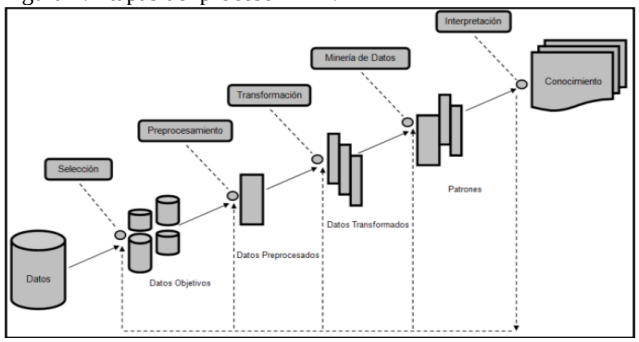
El cáncer es un problema de salud pública que está presente en todo el mundo [1]. En México, el cáncer ocupa una de las primeras causas de muerte [13] siendo diversos los diagnósticos que tienen presencia.
En 2020 hubo un total de 683,823 defunciones siendo 97, 323 defunciones a causa de cáncer en México. Los tipos de cáncer que se presentan en el análisis son cáncer de próstata, mama, colon, pulmonar y de recto. Un año antes, 2019 respectivamente se registraron un total de 747,784 defunciones de las cuales 88,683 defunciones fueron a causa de estos padecimientos. La distribución por sexo indica que hay más decesos de mujeres con un total de 51% mientras que los hombres ocupan el 49% restante, datos extraídos de INEGI. [14] El cáncer puede desarrollarse en cualquier parte del cuerpo. Se origina cuando las células crecen sin control y sobrepasan en número a las células normales, dificultando su funcionamiento. Por lo anterior su detección oportuna esto hace que al cuerpo le resulte difícil de funcionar de la manera que debería hacerlo. [16]. Conforme avanza el proceso de los pacientes con este tipo de diagnóstico se presentan los cuidados paliativos que no son lo mismo que los cuidados para enfermos terminales, estos tipos de cuidados paliativos son otro tipo de cuidados para enfermos terminales. Cuyo objetivo es mantener al paciente lo más cómodo posible cuando no se espera que el tratamiento cure el cáncer o bien tratar de prolongar la vida del paciente [17]. También se puede mejorar la calidad de vida del paciente con cuidados paliativos y apoyo psicosocial [18]. Ante los avances de la ciencia los tratamientos para atender estos diagnósticos han sido de gran ayuda para las personas con alguna de estas afecciones ya que el paciente presenta alteraciones en las diversas etapas del tratamiento y los cuidados paliativos actúan siendo responsables de dar una calidad de vida mejor, estando directamente relacionados con los tratamientos de estos diagnósticos.
Con base en ello se pueden identificar ciertos patrones y/o relaciones que ayuden a concretar la investigación y poder obtener un resultado eficiente utilizando los recursos que estas herramientas ofrecen. [19]. El enlace clínico que tiene relación entre el sector médico y la minería de datos es la ayuda en el control estadístico de los diagnósticos de patologías que permiten a su vez crear e implementar programas en las diversas instancias medicas para la detección y atención oportuna de los diagnósticos presentados. Existen factores de riesgo que complican la estabilidad de la salud del paciente como lo son: la mala alimentación, la inactividad física, el consumo de tabaco y de alcohol, la presencia de algunas enfermedades no transmisibles y/o hereditarias, sin embargo, no se ha comprobado que estos factores sean determinantes para provocar un cáncer.
Cada tipo de cáncer requiere un tratamiento y protocolo especifico, que puede ir desde una cirugía, radioterapia, quimioterapia, inmunoterapia, etc. [22] Por eso es esencial un diagnóstico correcto para poder dar el tratamiento adecuado, para brindar al paciente una mejor calidad de vida. Con la información obtenida se pretende manipularla de

mejor manera con una herramienta de la minería de datos que es el procesamiento de los datos para encontrar patrones de comportamiento que sean de utilidad para la toma de decisiones; se relaciona de manera estrecha con la estadística al usar técnicas de muestreo y visualización de datos y depuración [24]. La información obtenida es valiosa para el proceso de la toma de decisiones, esto si se le aplica los métodos y herramientas optimas. Una de las más grandes ventajas de la Minería de Datos (MD) está su facilidad de uso y la aplicabilidad de un conocimiento adecuado de los distintos tipos de algoritmos empleados con lo cual se facilitan muchas relaciones con otras herramientas (bases de datos, hojas de cálculo, etc.). Los términos MD y KDD son a menudo confundidos como sinónimos. Otros pasos en el proceso KDD [Fig. 1], son la preparación de los datos, la selección y limpieza de estos, la incorporación de conocimiento previo, y la propia interpretación de los resultados de minería. La presente investigación tiene como objetivo analizar con ayuda de la minería de datos ver la relación entre variables como diagnóstico, sexo y síntomas de afección [20]**.**

2. MATERIALES Y MÉTODOS

En el banco de datos propuesto se analizaron un total de 10,925 datos los cuales integran 576 expedientes de pacientes con diagnostico oncológico dentro del periodo de estudio 2010-2021, estos datos proporcionados por la unidad de Cuidados Paliativos del Hospital Regional 1° de Octubre de la Ciudad de México. La investigación y previo análisis se desarrollaron dentro del enfoque cuantitativo de tipo descriptivo, utilizando como mayor referencia de proceso al conocido como Knowledge Discovery in Databases (KDD) [Figura 1] para el tratamiento del banco de datos utilizado y este proceso se abarco lo siguiente: planteamiento del problema, entendimiento del problema, búsqueda de datos, selección de datos y preprocesamiento de datos, limpieza de datos y transformación de datos, análisis de datos y pruebas unitarias.[12] En todo el desarrollo del proceso se mantuvo control y supervisión de la integridad del banco de datos, asegurando un control total de la seguridad de los datos.

Figura 1. Etapas del proceso KDD.



Fuente: “Timarán, Hernández, Caicedo, Hidalgo & Alvarado, 2016”.

El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido reconocido por muchos investigadores como un tópico de investigación clave en los sistemas de bases de datos, y por muchas compañías industriales como una importante área y una oportunidad para obtener mayores ganancias [20].

2.1. PLANTEAMIENTO DEL PROBLEMA

Se dispone a desarrollar un análisis descriptivo, respecto a la asociación de variables y su incidencia en síntomas de pacientes con diversos diagnósticos oncológicos, utilizando la minería de datos.

2.2. ENTENDIMIENTO DEL PROBLEMA

Una vez realizado el análisis de la problemática y el planteamiento de esta con la recolección de datos, la aplicación de la metodología y la selección de herramientas adecuadas se podrá generar la solución por medio de la minería de datos.

2.3. BÚSQUEDA DE DATOS

El banco de datos con el que se realizara el presente trabajo proviene de expedientes clínicos de la unidad de cuidados paliativos, por lo que únicamente se utiliza información general sin hacer uso de datos personales.

2.4. SELECCIÓN DE DATOS Y PREPROCESAMIENTO DE DATOS

Esta se obtuvo mediante la selección de las principales variables que integraran el banco final de ya que la base de datos es amplia y con variables inútiles, se optó por determinar las óptimas para la conformación del banco de datos, dejando fuera variables que indicaban cuestiones medicas no necesarias. Así como comenzar con la revisión de la integridad de datos y descartar aquellos que están incompletos para que no afecten el resultante.

2.5. LIMPIEZA DE DATOS Y TRANSFORMACIÓN DE DATOS

En un inicio el banco de datos utilizado constaba de un total de 576 expedientes, con 21 variables iniciales (diagnóstico, sexo, edad, año, promedio dolor, promedio debilidad-cansancio-fatiga, promedio nauseas, promedio depresión, promedio ansiedad, promedio somnolencia, promedio falta de apetito, promedio malestar, promedio falta de aire, promedio insomnio, promedio estreñimiento, promedio alucinaciones, promedio hinchazón, promedio fiebre y promedio hemorragia-sangrado, fluidos, sangrado) de las cuales 19 se tomaron mediante la selección de las variables que integrarán el banco final (diagnóstico, sexo, edad, año, promedio dolor, promedio debilidad-cansancio-fatiga, promedio nauseas, promedio depresión, promedio ansiedad, promedio somnolencia, promedio falta de apetito, promedio malestar, promedio falta de aire, promedio insomnio, promedio estreñimiento, promedio alucinaciones, promedio hinchazón, promedio fiebre y promedio hemorragia-sangrado), las variables eliminadas no aportaban información útil para la integración del banco final ya que presentaban datos nulos por lo anterior en la tabla [Tabla 1] se visualiza la descripción con nombre, variable y definición final.

Tabla 1. Variables Principales

Nombre variable	Definición	Descripción
Dx	Diagnostico	Indica el diagnóstico del paciente
Sexo	Sexo	Indica el sexo del paciente
Edad	Edad	Indica la edad del paciente
Año	Año	Indica el año de registro del paciente
ProDol	Promedio dolor	Indica el promedio de dolor del paciente
ProDCF	Promedio Debilidad-Cansancio-Fatiga	Indica el promedio de Debilidad-Cansancio-Fatiga del paciente
ProNau	Promedio Nauseas	Indica el promedio de nauseas del paciente
ProDep	Promedio Depresión	Indica el promedio de depresión del paciente
ProAnsi	Promedio Ansiedad	Indica el promedio de ansiedad del paciente
ProS	Promedio Somnolencia	Indica el promedio de somnolencia del paciente
ProFa	Promedio Falta apetito	Indica el promedio de falta de apetito del paciente
ProM	Promedio Malestar	Indica el promedio de malestar del paciente
ProFaire	Promedio Falta de aire	Indica el promedio de falta de aire del paciente
ProI	Promedio Insomnio	Indica el promedio de insomnio del paciente
ProE	Promedio Estreñimiento	Indica el promedio de estreñimiento del paciente
ProAl	Promedio Alucinaciones	Indica el promedio de alucinaciones del paciente
ProH	Promedio Hinchazón	Indica el promedio de hinchazón del paciente
ProFi	Promedio Fiebre	Indica el promedio de fiebre del paciente
ProHS	Promedio Hemorragia-Sangrado	Indica el promedio de hemorragia-sangrado del paciente

Fuente: Expedientes Clínicos.

En la tabla [Tabla 2] se muestra la adaptación con descripción de los diagnósticos oncológicos presentados, así como los grados de afectación y su respectiva definición.

Tabla 2. Descripción de los atributos referentes a los tipos de datos presentes en el banco de datos.

Acotación Variable	Nombre Variable
Cacol	Cáncer colón
Cama	Cáncer mama
Capo	Cáncer próstata
Capu	Cáncer pulmón
Carec	Cáncer recto
MG	Moderado grave
LM	Leve moderado
MI	Muy intense

Fuente: Expedientes Clínicos.

2.6. ANÁLISIS DE DATOS Y PRUEBAS UNITARIAS

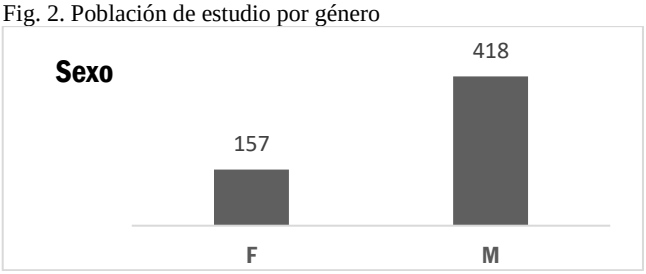
Con lo realizado en las fases anteriores se procede a tratar el banco de datos todo esto dentro del software especializado WEKA [25] dejando de lado un poco las demás variables primarias que ayudaron a ponderar otras cuestiones dentro del proceso.

Los datos analizados dentro el software al estar agrupados en categorías fueron tratados por los algoritmos de ZeroR, algoritmo tomado como punto de prueba, (ZeroR es el método de clasificación más simple que existe y depende solo en el target ignorando todos los predictores. El clasificador ZeroR predice sobre la clase o categoría principal (majority category). El algoritmo ZeroR tiene gran poder predictivo, así como el método PART algoritmo (El algoritmo PART en Weka realiza casi las mismas operaciones que el algoritmo J48, con la diferencia que este algoritmo no genera árboles de decisión, sino es un algoritmo para la obtención de reglas de un árbol de decisión, pero recibe algunos parámetros similares que el J48), además del algoritmo OneR que está basado en el algoritmo ID3, donde la meta principal consiste en adquirir las reglas de clasificación directamente desde el conjunto de datos de entrenamiento. Es por tanto el algoritmo de clasificación seleccionado ya que es simple y efectivo, de uso frecuente en el aprendizaje de máquinas, utiliza un único atributo para la clasificación, el cual es el de menor porciento de error y se obtiene un conjunto de reglas. Los algoritmos con la mayor efectividad devuelta en la clasificación del banco de datos y dentro de la rama de árboles de decisión fue el J48 (El algoritmo J48 de WEKA es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos más utilizado. Se trata de un refinamiento del modelo generado con OneR. Supone una mejora moderada en las prestaciones y podrá conseguir una probabilidad de acierto ligeramente superior al del anterior clasificador.) que pertenece al algoritmo C4.5, junto a ello el algoritmo PART y OneR otorgando la misma efectividad en la clasificación de los datos.

3. RESULTADOS

De acuerdo con las diversas pruebas obtenidas en las diferentes evaluaciones hechas con la ayuda de la minería de datos y las herramientas de análisis de datos en específico las tareas de clasificación, se devuelve el análisis del banco de datos proporcionado y la evaluación de las variables principales arrojando un total de 576 expedientes clínicos de los cuales 157 expedientes corresponden al sexo Femenino (F) con un porcentaje del (27.30%) y los 418 corresponden al sexo Masculino (M) con un porcentaje del (72.70%) [Figura 2].

Fig. 2. Población de estudio por género

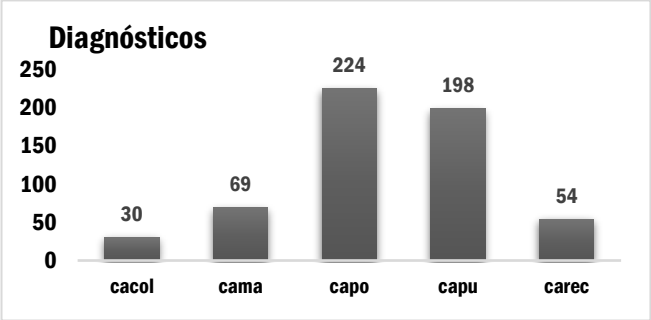


Fuente: elaboración propia a partir de Expedientes clínicos

Considerando los registros del banco de datos, en el periodo de estudio (2010-2021) se muestra la cantidad total de diagnósticos por tipo de cáncer dejando ver que el cáncer de próstata es el que más afluencia muestra con un total de (224 registros) seguido del cáncer de pulmón con un total de (198 registros), los siguientes muestran una menor incidencia tal como el cáncer de mama con un total de (69 registros), el cáncer de recto con un total de (54 registros) y por último el cáncer de colon con un total de (30 registros) [Figura 3]

dejando interpretar que los diagnósticos corresponden a los datos analizados.

Figura 3. Rangos de diagnósticos con respecto a la población de estudio



Fuente: elaboración propia a partir de Expedientes clínicos

Al utilizar los diferentes algoritmos para el análisis de las variables se visualiza la incidencia en cuanto a los registros de la variable sexo que está marcada hacia el sexo masculino con un 72.70% contra un 27.30% restante que corresponde al género femenino [Figura 4], donde se muestra los totales devueltos tras la evaluación del algoritmo.

Figura 4. Clasificación de pacientes por sexo.

```
Correctly Classified Instances         418               72.6957 %
Incorrectly Classified Instances       157               27.3043 %
Kappa statistic                          0
Mean absolute error                      0.3974
Root mean squared error                  0.4455
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              575

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.000    0.000    ?          0.000    ?          ?        0.491     0.269     F
               1.000    1.000    0.727      1.000    0.842      ?        0.491     0.723     M
Weighted Avg.  0.727    0.727    ?          0.727    ?          ?        0.491     0.599

=== Confusion Matrix ===

   a   b   <-- classified as
   0 157 |   a = F
   0 418 |   b = M
```

Fuente: elaboración propia a partir de Expedientes clínicos

En cuanto a la parte de la clasificación de los diagnósticos [Figura 5] se muestra que los tipos de cáncer que conforman el banco de datos con mayor incidencia para el periodo reportado fueron: cáncer de próstata 21.5%, pulmón 11.78%, de mama 5.1%, de recto 31.5% y de colon 6.05%, ejecutados mediante el algoritmo J48 de WEKA, cuya eficacia de este algoritmo muestra un 89.04% de instancias correctamente clasificadas, en tanto que la tasa de error es equivalente a un 10.96% de tal manera ya que al ser clasificados en ambos sexos el algoritmo lo divide según las variables proporcionadas.

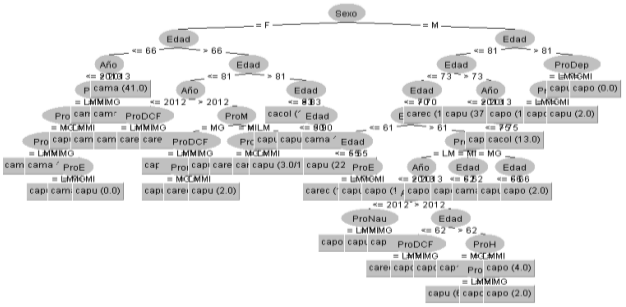
Figura 5. Clasificación de variable diagnósticos mediante el algoritmo J48 en WEKA

```
Correctly Classified Instances         512               89.0435 %
Incorrectly Classified Instances        63               10.9565 %
Kappa statistic                          0.8445
Mean absolute error                      0.0533
Root mean squared error                  0.1948
Relative absolute error                 18.9189 %
Root relative squared error             51.9124 %
Total Number of Instances              575
```

Fuente: elaboración propia a partir de Expedientes clínicos

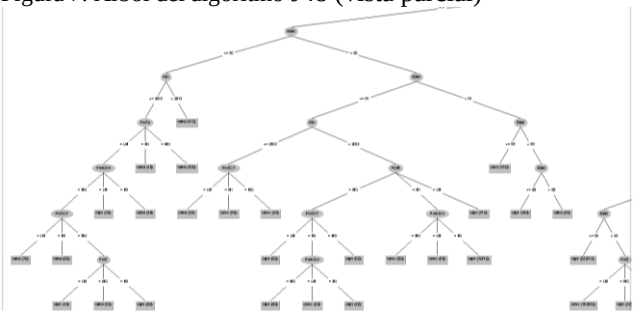
En las Figuras 6 y 7 se muestra el árbol que se divide en dos sexos, con el algoritmo clasificador J48, el cual es una implementación del algoritmo C.4.5. Las variables más importantes que se consideran son: sexo y diagnóstico.

Figura 6. Árbol del algoritmo J48 (vista general)



Fuente: elaboración propia a partir de Expedientes clínicos

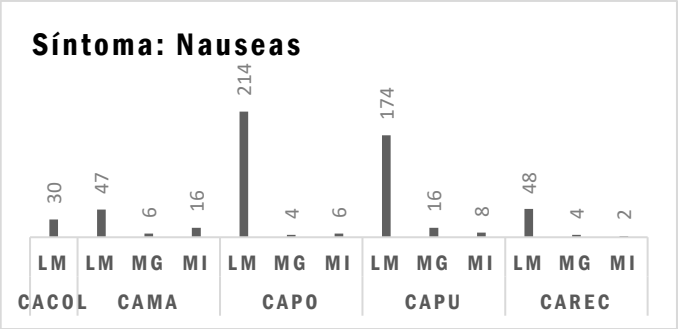
Figura 7. Árbol del algoritmo J48 (vista parcial)



Fuente: elaboración propia a partir de Expedientes clínicos

Dentro de la evaluación de los síntomas presentados se destaca la afluencia de la ansiedad y depresión de los pacientes estando directamente relacionada con los diagnósticos como lo muestran las figuras 8 y 9 siendo las náuseas el síntoma analizado y el algoritmo utiliza fue el PART, siendo el cáncer de próstata el que más afluencia tiene y el grado Leve Moderado el que más se presenta con 214 registros.

Figura 8. Gráfica síntoma: Náuseas relación con Diagnóstico.



Fuente: elaboración propia a partir de Expedientes clínicos

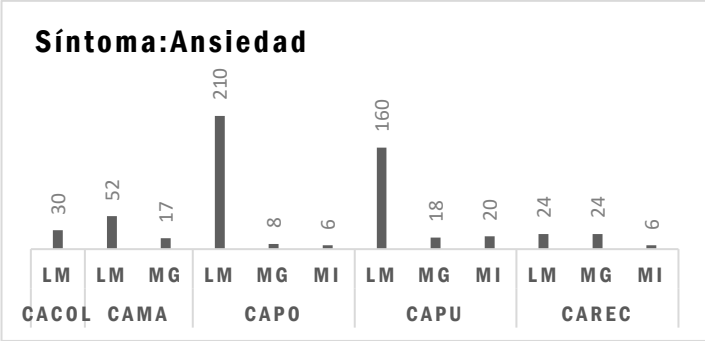
Figura 9. Análisis Síntoma Náuseas con algoritmo PART.

```
Correctly Classified Instances         525               91.3043 %
Incorrectly Classified Instances        50                8.6957 %
Kappa statistic                          0.5149
Mean absolute error                      0.0603
Root mean squared error                  0.2193
Relative absolute error                 45.0279 %
Root relative squared error             85.3238 %
Total Number of Instances              575

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.967    0.484    0.943      0.967    0.955      0.537    0.917     0.986     LM
               0.594    0.017    0.679      0.594    0.633      0.615    0.967     0.780     MI
               0.333    0.020    0.476      0.333    0.392      0.371    0.904     0.439     MG
Weighted Avg.  0.913    0.434    0.904      0.913    0.908      0.532    0.919     0.946

=== Confusion Matrix ===

   a   b   c   <-- classified as
 496   7  10 |   a = LM
  12  19   1 |   b = MI
  18   2  10 |   c = MG
```

Fuente: elaboración propia a partir de Expedientes clínicos

Mientras que en las figuras 10 y 11 se muestra el síntoma ansiedad utilizando el algoritmo JRip, siendo el cáncer de próstata el que más afluencia tiene y el grado Leve Moderado el que más se presenta con 210 registros.

Figura 10. Gráfica síntoma: Ansiedad relación Diagnóstico.



Fuente: elaboración propia a partir de Expedientes clínicos

Figura 11. Análisis Síntoma Ansiedad con algoritmo JRip.

```
Correctly Classified Instances         531               92.3478 %
Incorrectly Classified Instances        44                7.6522 %
Kappa statistic                          0.7176
Mean absolute error                      0.0717
Root mean squared error                  0.2145
Relative absolute error                 35.8139 %
Root relative squared error             68.0439 %
Total Number of Instances              575

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.687    0.020    0.821      0.687    0.748      0.722    0.879     0.755     MG
               0.985    0.283    0.944      0.985    0.964      0.774    0.861     0.948     LM
               0.500    0.011    0.727      0.500    0.593      0.584    0.818     0.525     MI
Weighted Avg.  0.923    0.237    0.917      0.923    0.918      0.758    0.861     0.902

=== Confusion Matrix ===

   a   b   c   <-- classified as
  46  15   6 |   a = MG
   7 469   0 |   b = LM
```

Fuente: elaboración propia a partir de Expedientes clínicos

4. DISCUSIÓN

Los resultados obtenidos durante el desarrollo de la investigación devuelven puntos importantes a destacar, entre ellos la relación que confirma la información recabada con lo analizado esto conforme las diferentes herramientas de la minería de datos y técnicas utilizadas, destacar el incremento en la incidencia de los diagnósticos, la intervención de los cuidados paliativos en el cuidado y atención de los pacientes para tratar de brindar una calidad de vida más llevadera ante toda la sintomatología que la enfermedad presenta, ya que estos cuidados al ser especializados para el tratamiento de estos diagnósticos son bastante útiles e importantes para mejorar la calidad de vida en los pacientes, con ello poder contribuir en el posible planteamiento de programas que ayuden a promover una solución a la desatención y falta de apoyo en los cuidados paliativos de este problema. Gracias a la ayuda de las diversas herramientas de la minería de datos en las diversas fases se visualiza de mejor manera el comportamiento de este problema que aqueja a la población con este diagnóstico. Considerando la estimación y asociación se conocieron los principales factores, en este caso manejados como variables en las cuales se puede denotar la discriminación entre géneros y diagnósticos como los más relevantes para el desarrollo de la investigación coincidiendo con los resultados presentados por la Cancer Journal en los diversos números aplicando las técnicas de discriminación en el análisis [1].

Con base en ello las diversas etapas dentro del análisis confirman lo propuesto por la Sociedad Americana [11] contra el cáncer, que año con año recaban la información de dichos diagnósticos en gran volumen, la relación que presentan los datos con las diversas relaciones entre variables, más la intervención de los cuidados paliativos abre la posibilidad de auxiliar con más herramientas tecnológicas las investigaciones como lo interpreta Gutiérrez [19], siendo la escala de Edmonton el auxiliar en la interpretación de los datos. En las etapas anteriores se muestra una coincidencia determinada con lo propuesto, además de poder visualizar de manera óptima los datos del banco de datos tratado.

Por lo tanto, se puede confirmar de manera eficiente que el análisis, tratamiento, procesamiento y la interpretación de los datos fue exitoso, tal como lo muestran las diversas figuras y algoritmos propuestos que son ilustrados en los resultados.

5. CONCLUSIONES

Conforme a los resultados obtenidos de los 576 expedientes, se pudo observar en las representaciones gráficas que los diagnósticos con más incidencia en cuanto a sexo los hombres se ven más afectados siendo el cáncer de próstata el que más incidencia presenta. Todo esto confirmado con los estudios publicados por la Sociedad Americana contra el Cáncer donde indica que este tipo de diagnósticos se han ido incrementando cada año.

Dentro de las muestras arrojadas por el análisis cabe aclarar que los síntomas presentados son los que más destacaron para el desarrollo de los resultados.

Por lo anterior cabe destacar que el análisis y procesamiento de los grandes volúmenes de datos y su respectiva interpretación con técnicas de minería de datos aplicados en un caso real resultan importantes al momento de encontrar patrones que sean de utilidad para el descubrimiento del conocimiento en cualquier área de estudio. Así mismo, la recopilación, el preprocesamiento y la visualización de los datos que hemos manejado nos permiten obtener con más eficiencia mejores resultados.

6. REFERENCIAS

[1] CA: A Cancer Journal for Clinicians. (2010) [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.20073

[2] CA: A Cancer Journal for Clinicians. (2011) [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.20121

[3] CA: A Cancer Journal for Clinicians. (2012) [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21262

[4] CA: A Cancer Journal for Clinicians. (2013) [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21166

[5] CA: A Cancer Journal for Clinicians. (2014) [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21208

[6] CA: A Cancer Journal for Clinicians. (2016) [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/journal/15424863

[7] CA: A Cancer Journal for Clinicians. (2017) "Cancer Statistics, 2017" [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21387

[8] CA: A Cancer Journal for Clinicians. (2018) "Cancer Statistics, 2018" [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21442

[9] CA: A Cancer Journal for Clinicians. (2019) "Cancer Statistics, 2019" [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21551

[10] CA: A Cancer Journal for Clinicians. (2020) "Cancer Statistics, 2020" [Fecha de consulta 5 de Septiembre de 2022]. Disponible en: https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21590

[11] CA: A Cancer Journal for Clinicians. (2021) "Cancer Statistics, 2021" [Fecha de consulta 5 de Septiembre de 2022]. https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21654

[12] Büchner, A. G., Anand, S. S., Mulvenna, M. D. & Hughes, J. G. (2002). Discovering Internet Marketing Intelligence through Web Log Mining. *Mining the Internet for Marketing IntelligenCe*, *26749*(1). http://uir.ulster.ac.uk/7920/1/Discovering_Internet_Marketing_Intelligence_through_Web_Log_Mining.pdf

[13] Gaceta Mexicana de Oncología. (2018)" Mortalidad por Cáncer en México: actualización 2015" [Fecha de consulta 8 de Septiembre de 2022]. Disponible en: https://www.researchgate.net/publication/324522811_Mortalidad_por_Cancer_en_Mexico_actualizacion_2015

[14] Instituto Nacional de Estadística y geografía (INEGI). (2020) "ESTADÍSTICAS A PROPÓSITO DEL DÍA MUNDIAL CONTRA EL CÁNCER". [Fecha de consulta 8 de Septiembre de 2022]. Disponible en: https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2021/cancer2021_Nal.pdf

[15] Cuidados Paliativos, Guías para el manejo clínico (2020). [Fecha de consulta 8 de Septiembre de 2022]. Disponible en: https://www.paho.org/hq/dmdocuments/2012/PAHO-Guias-Manejo-Clinico-2002-Spa.pdf

[16] Juntos contra el cáncer. (2020) "Panorama del cáncer en México". [Fecha de consulta 8 de Septiembre de 2022]. https://juntoscontraelcancer.mx/panorama-del-cancer-en-mexico/

[17] Synthon (2021) "Cáncer en México y el mundo". [Fecha de consulta 20 de Diciembre de 2021]. Disponible en: https://www.synthon.com/mx/nuestro-negocio/oncologia/el-cancer-en-mexico-y-el-mundo

[18] Ascencio Huertas, L. (2015). Adaptación en español de la escala de actitudes ante cuidados paliativos: confiabilidad y análisis factorial. Psicooncología, 12(2-3), 367-381.

[19] Gutierrez, P. L., Gutierrez, C. D. (2020) "Descubrimiento de conocimiento en incidencia de tipo de cáncer para pacientes terminales mediante minería de datos". *Ciencia Ergo-Sum.* (28) (1-10). https://doi.org/10.30878/ces.v28n1a5

[20] Timarán Pereira, R. (2009). La minería de datos en el descubrimiento de perfiles de deserción estudiantil en la Universidad De Nariño. *Universidad Y Salud*, *1*(11). Recuperado a partir de https://revistas.udenar.edu.co/index.php/usalud/article/view/218

[21] Torres Sánchez, L., Rojas Martínez, R., Escamilla Núñez, C. y Lazcano-Ponce, E. (2014). Tendencias en la mortalidad por cáncer en México de 1980 a 2011. Salud Pública Mex, 56(5), 473-491. doi: http://dx.doi.org/10.21149/spm.v56i5.7373

[22] Gómez Dantés H, Lamadrid-Figueroa H, Cahuana-Hurtado L, Silverman-Retana O, Montero P, González-Robledo MC, Fitzmaurice C, Pain A, Allen C, Dicker DJ, Hamavid H, López A, Murray C, Naghavi M, Lozano R (2016). The burden of cancer in Mexico, 1990-2013. Recuperado a partir de: https://www.medigraphic.com/pdfs/salpubmex/sal-2016/sal162e.pdf

[23] Sierra, M., Cueva, P., Bravo, L., & Forman, D. (2016). Stomach cancer burden in Central and South America. Cancer Epidemiology, 44(Suppl 1), S62-S73. doi: http://dx.doi.org/10.1016/j.canep.2016.03.008

[24] H. Witten, Ian& Frank, Eibe. (2005). Data Mining. Editorial Elsevier. ISBN: 0-12-088407-0, United States of America. https://www.wi.hswismar.de/~cleve/vorl/projects/dm/ss13/HierarClustern/Literatur/WittenFrank-DM-3rd.pdf

[25] Software WEKA