

Sistema de apoyo a la detección de tumoración maligna en el tejido mamario.

Trabajo Terminal No. 2020 - A069

Alumno: Ortiz Rivas Julio César

Director: Ocampo Botello Fabiola

Turno para la presentación del TT: VESPERTINO

e-mail: julio_cesar502@hotmail.com

Resumen – Este Trabajo Terminal tiene como propósito desarrollar un sistema de información para determinar si una muestra de tejido mamario presenta tumoración maligna o benigna, teniendo como referencia características de muestras de tejidos las cuales ya se han sido catalogadas con presencia de tumoración maligna o benigna, según sea el caso.

Palabras clave – Minería de datos, ingeniería de software, aprendizaje automático, análisis de datos, sistema predictivo, modelo predictivo, tumoración maligna, cáncer de mama.

1. Introducción

El cáncer es una de las principales patologías que afectan a la población a nivel mundial. La introducción de estrategias de detección precoz y mejoras en la terapia del cáncer han permitido en países desarrollados disminuir su incidencia y mejorar la sobrevivencia de los pacientes [1].

El cáncer de mama es la neoplasia (formación de tumores) más frecuente en el mundo y es la causa con mayor mortalidad entre las mujeres, estimándose en el año 2016 la muerte de 521, 907 de estas, lo cual representa el 14.7% de muertes en la población femenina [2].

El aumento global en la frecuencia del cáncer de mama ha sido más preocupante en las naciones con economías en desarrollo como la de México, donde las carencias en infraestructura, culturales y en la comunicación fallan para difundir la gravedad del problema y los mecanismos para su detección temprana, así como su tratamiento adecuado [3].

Actualmente, según el Instituto Nacional de Cancerología (INCan), se diagnostican 191 mil casos al año, de los cuales fallecen 84 mil [4], representando casi el 44% de los casos diagnosticados. Estas cifras se vuelen alarmantes puesto que en su mayoría las defunciones se dan por un diagnóstico tardío, lo cual hace importante su detección en la etapa inicial, para con ello tratar de disminuir la cifra de defunciones.

Una cuestión que es importante mencionar es que en la etapa inicial del cáncer no se presentan síntomas, por lo que las personas no consideran importante o relevante el chequeo constante hasta que se presenta alguna molestia. En el momento que aparecen los síntomas es porque el cáncer ya está avanzado y lo que sucede es que el tumor es tan grande que puede llegar a comprimir los nervios, lo cual ocasiona dolor u obstrucción.

De esta manera, teniendo un historial de características claves para determinar si un tumor es maligno o benigno, la minería de datos permite desarrollar un modelo probabilístico en el cual se determinan las características que hacen que un tumor se considere maligno o benigno y estos a su vez sirvan de referencia para el análisis de nuevas muestras.

La minería de datos es frecuentemente definida como la acción de encontrar información escondida en una base de datos [5]. Esto es, que no solo se ven o se extraen los datos como tal, si no que se les da un sentido o un significado, generando un conocimiento que tal vez no se tenía sobre estos.

Para ello existen muchos y diferentes algoritmos que desarrollan distintas tareas. Dichos algoritmos examinan los datos y determinan un modelo que es muy cercano a las características de los datos que se están examinando.

Los algoritmos de minería de datos pueden caracterizarse por las siguientes tres cualidades:

- **Modelo.** El propósito del algoritmo es que los datos sean ajustables o compatibles para el modelo.
- **Preferencia.** Algún criterio debe ser usado para elegir el modelo que más se asimile a la estructura de los datos.
- **Búsqueda.** Todos los algoritmos requieren de alguna técnica para realizar una búsqueda en los datos.

Existen dos tipos de modelos que pueden ser creados, los cuales pueden ser predictivo, donde el modelo hace una predicción de ciertos valores utilizando resultados ya conocidos, o descriptivo, donde el modelo identifica patrones y relaciones en los datos.

Teniendo en cuenta los dos tipos de modelos, el que se desarrollará para este caso de estudio en particular será un modelo predictivo, ya que se cuenta con un conjunto de datos que contienen las características de muestras de tejido mamario donde se determinó si fue encontrada tumoración maligna o benigna.

1.1 Estado del arte.

En la *Figura 1* se encuentran descritos sistemas o estrategias similares que se han desarrollado.

SISTEMA	CARACTERÍSTICAS	PRECIO EN EL MERCADO
Minería de datos como soporte en el diagnóstico y tratamiento del cáncer de mama [6].	<ul style="list-style-type: none"> Tesis. Aplica el proceso de minería de datos en repositorios relacionados con el cáncer de mama con la finalidad de descubrir conocimiento útil que apoye al proceso de diagnóstico y tratamiento del cáncer de mama en sus diferentes etapas. 	No se comercializó.
Análisis comparativo entre: «el análisis exploratorio de datos» y los modelos de «árboles de decisión» y «k-medias» en el diagnóstico de la malignidad en algunos exámenes de cáncer de mama [7].	<ul style="list-style-type: none"> Artículo. Busca determinar la malignidad de una masa detectada en el seno de un paciente, a partir de los atributos registrados de la masa. 	No se comercializó.
Diagnostico y pronostico de cáncer de mama mediante programación lineal [8].	<ul style="list-style-type: none"> Artículo. Se usan técnicas de aprendizaje automático basadas en programación lineal para aumentar la exactitud del diagnostico y pronostico del cáncer de mama. 	No se comercializó.

Figura 1. Proyectos similares.

2. Objetivo

Desarrollar un sistema de información para predecir la probable presencia de tumoración maligna mediante el análisis de datos de muestras de tejido mamario considerando un modelo predictivo generado a partir de la aplicación de técnicas de minería de datos a registros históricos de muestras que presentan este padecimiento.

3. Justificación

Gracias a la medicina preventiva se ha logrado aumentar notablemente el promedio de vida de una población, así como la calidad de esta. Esto se ve en la implementación de los programas de vacunación y la aplicación de los conocimientos adquiridos al conocer acerca de la historia natural de las enfermedades, tanto su prevención como su tratamiento.

Por otra parte, cuando se trata de padecimientos malignos el esfuerzo realizado para que sean evitados no ha alcanzado buenos resultados en la inmensa mayoría de los cánceres. Como consecuencia de esto la única esperanza es la detección temprana de los tumores malignos, ya que se ha visto que cuando algunos de ellos son detectados en su etapa inicial esto permite aplicar los métodos curativos actuales a los pacientes antes de que dicha enfermedad avance y se extienda, puesto que en esta etapa se logra, en el mejor de los casos, mitigar dichos padecimientos [3].

Tomando como referencia múltiples encuestas y estadísticas realizadas por la OMS (Organización Mundial de la Salud) y la Secretaría de Salud en México, se determina que no es posible prevenir el cáncer de mama, sin embargo, dichos estudios parecen demostrar que el riesgo de padecer cáncer de mama se puede reducir considerablemente.

En México, se realizan de forma anual más de 5.6 millones de exploraciones anuales para la prevención y detección del cáncer de mama. Una de las estrategias para la detección temprana es realizar la autoexploración y la exploración clínica a mujeres, pero esto no es suficiente ya que sólo se cuentan con 284 mastógrafos y se atienden apenas 18 estudios diarios [9].

El sistema a desarrollar en este trabajo está pensado como herramienta de apoyo a las instituciones de salud para ayudar a la detección de neoplasias mamarias, esto con el fin de que las autoridades sanitarias pertinentes determinen los tratamientos adecuados para combatir el padecimiento.

4. Productos o resultados esperados

Al finalizar el trabajo se pretende tener un sistema de información predictivo que determinará si en una muestra de tejido mamario se encuentra o no tumoración maligna, basado en el modelo predictivo resultante del proceso de minería, el cual se explicará en el punto 5.

Se ingresarán las características encontradas en el tejido, las cuales se seleccionarán en el proceso de análisis. Después, estos datos serán ingresados al modelo predictivo, el cual será entrenado con los datos arrojados por el análisis. Por último, el algoritmo determinará la presencia de tumoración maligna o benigna en la muestra tomada.

La explicación anterior, la cual se refiere al funcionamiento de la aplicación, se encuentra interpretada en la *Figura 2*.

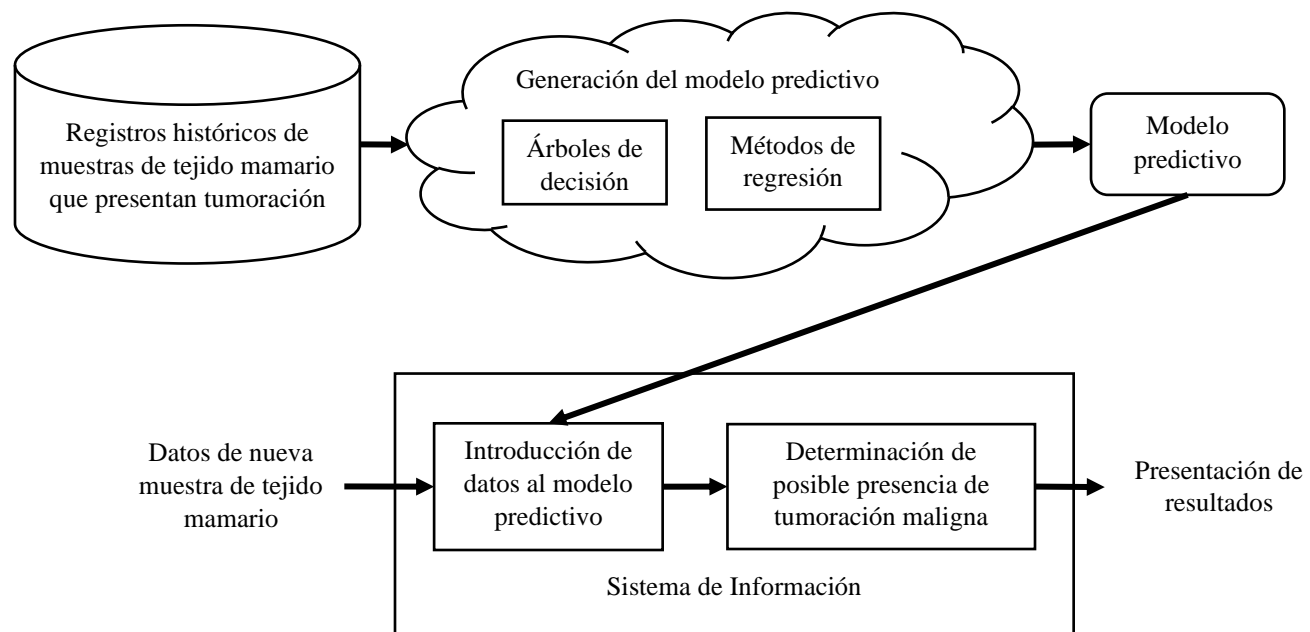


Figura 2. Arquitectura del sistema.

Durante el desarrollo del trabajo se hará la documentación respectiva a este, donde se describirán los aspectos técnicos del sistema y el proceso de análisis, así como un manual de usuario.

5. Metodología

A continuación, se definen las metodologías que se optaron como las más convenientes para el análisis de los datos y el desarrollo del software.

5.1 Análisis de datos

El análisis de los datos es una parte crucial en el desarrollo del trabajo, ya que a partir de este se determinarán los aspectos claves en las características de las células del tejido mamario que determinan si un tumor es maligno o benigno. El método que se eligió es uno

de los más utilizados en la extracción de información desde una base de datos: Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases, KDD*) (Figura 3). Este método es muy eficiente ya que no solo extrae la información, sino que desarrolla conocimiento a partir de esta y para ello consta de las siguientes etapas:

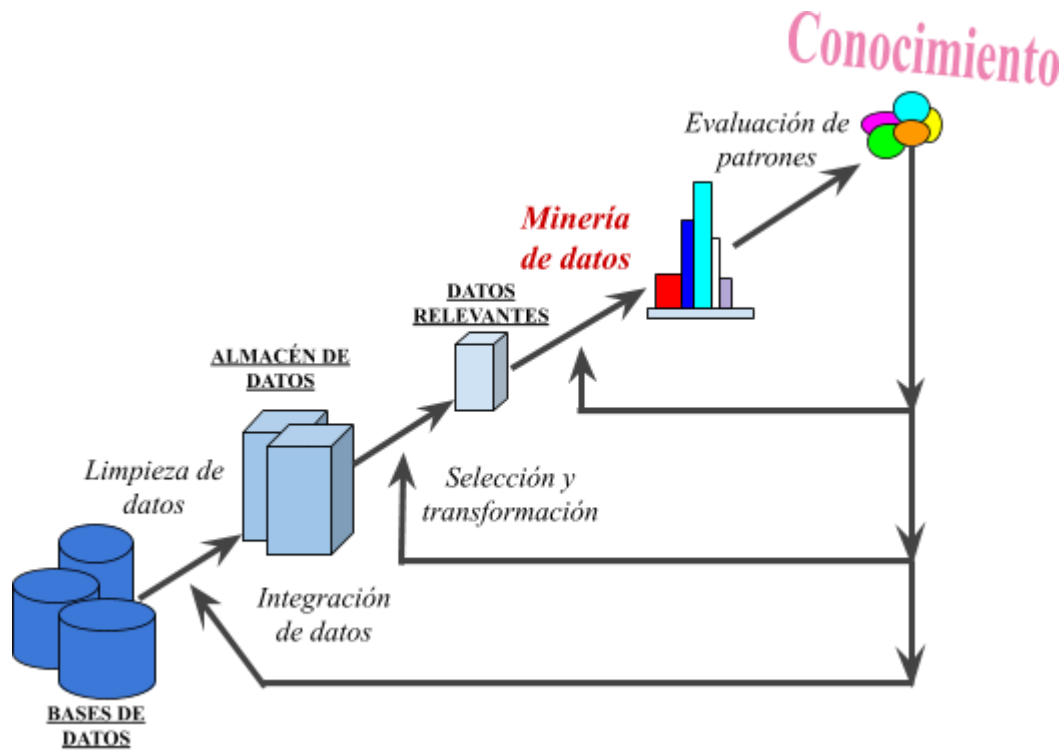


Figura 3. Descubrimiento de Conocimiento en Bases de Datos. Figura basada y traducida de [10].

- **Limpieza de datos:** Es la fase en la que los datos sucios son eliminados o corregidos de la colección de datos.
- **Integración de los datos:** En esta se combinan datos de diferentes fuentes en un solo conjunto de datos.
- **Selección de datos:** Se eligen los datos relevantes para el análisis.
- **Transformación de datos:** Los datos seleccionados son transformados de manera que se puedan adaptar al proceso de minería.
- **Minería de datos:** Es la etapa crucial del proceso, es aquí donde se aplican técnicas ingeniosas para determinar patrones interesantes potencialmente útiles.
- **Evaluación de patrones:** Estrictamente los patrones que representan conocimiento son identificados basándose en medidas dadas.
- **Representación del conocimiento:** Es la etapa final en la cual el conocimiento descubierto es presentado al usuario. Esta etapa esencial utiliza técnicas de visualización para ayudar a los usuarios a entender e interpretar los resultados de la minería de datos [10].

5.1.1 Algoritmos de minería de datos

En este apartado se describe aproximación metodológica que se aplicará en las técnicas de minería de datos como parte del método de Descubrimiento de Conocimiento en Bases de Datos.

Considerando la naturaleza de los datos que se tienen, para este proyecto se contemplan dos técnicas de minería de datos:

1. **Árboles de decisión.** Estos permiten crear un modelo de comportamiento de los datos de entrada, mediante la identificación de las relaciones que guardan entre ellos. Se genera un modelo con la finalidad de aplicarlo para identificar este tipo de relaciones en nuevas muestras de datos.
Los algoritmos que producen los árboles de decisión son llamados inductores porque se comportan como discípulos que aprenden un conjunto de datos de entrada y forman un modelo que generaliza la relación entre los atributos de entrada y el atributo fuente [11].

2. **Regresión.** En la minería de datos se consideran como métodos de predicción. El objetivo de un modelo de regresión es tratar de explicar la relación que existe entre una variante dependiente (variable respuesta) y un conjunto de variables independientes (variables explicativas) [12]. Este método se considera utilizar para tratar de encontrar una ecuación que represente el comportamiento entre las variables que se analizan.

5.2 Desarrollo de software

La metodología elegida para el desarrollo del software ha sido el Modelo Incremental (*Figura 4*). Este modelo resulta de gran ayuda ya que permite asociar el análisis de los datos como parte del desarrollo del software, puesto que este sería considerado un incremento y el producto entregado en este será el modelo de predicción.

Una vez terminada la fase de análisis se determinarán los requerimientos del sistema y posteriormente se pasará a su diseño, codificación y pruebas. Teniendo así otros dos incrementos para la parte de desarrollo del software.

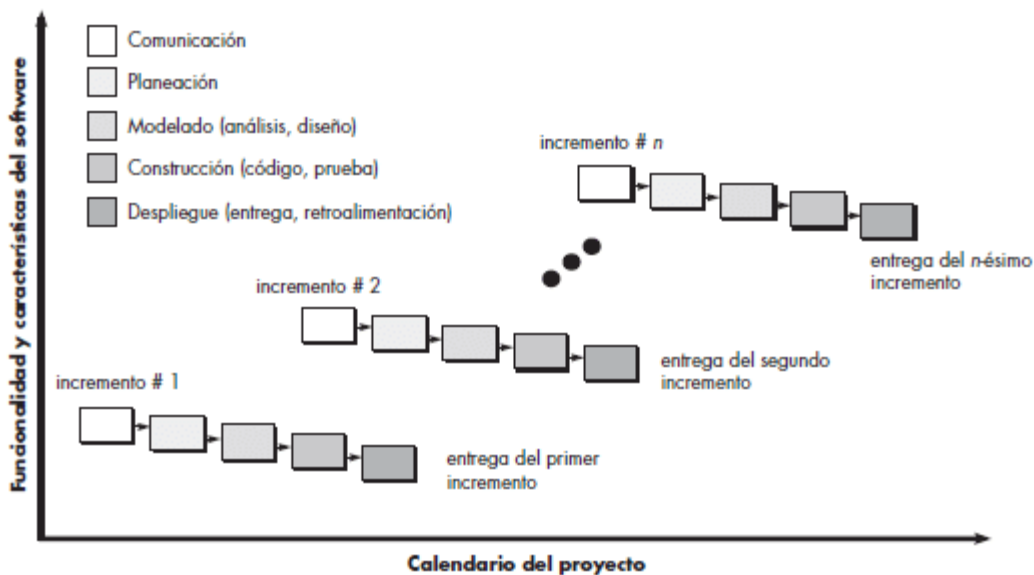


Figura 4. Modelo Incremental. Figura tomada de [13].

6. Cronograma

Las actividades para la realización de este trabajo terminal se plantean en un plazo de 10 meses de trabajo, considerando los semestres que integran el calendario lectivo publicado por el Instituto Politécnico Nacional. Al final de este documento se presenta el esquema de trabajo a realizar para este proyecto.

7. Referencias

- [1] N. C. Sánchez, «Conociendo y comprendiendo la célula cancerosa: Fisiopatología del cáncer,» *Revista Médica Clínica Las Condes*, vol. 24, nº 4, pp. 553-562, 2013.
- [2] Centro Nacional de Equidad de Género y Salud Reproductiva, «gob.mx,» 2 diciembre 2016. [En línea]. Available: <https://www.gob.mx/salud%7Ccnegrs/acciones-y-programas/informacion-estadistica-cancer-de-mama>. [Último acceso: 4 3 2020].
- [3] M. D. C. Lara Tamburrino y Á. Olmedo Zorrilla, «Detección temprana y diagnóstico del cáncer mamario,» *Revista de la Facultad de Medicina (México)*, vol. 54, nº 1, pp. 4-17, 2011.

- [4] Infobae, «infobae.com,» 4 febrero 2020. [En línea]. Available: <https://www.infobae.com/america/mexico/2020/02/04/dia-mundial-contra-el-cancer-2020-aumento-20-mortandad-en-mexico-desde-el-ano-2000/>. [Último acceso: 4 3 2020].
- [5] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.
- [6] C. C. L. Portillo, Minería de datos como soporte en el diagnóstico y tratamiento del cáncer de mama, Ensenada, Baja California, México: Centro de Investigación Científica y de Educación Superior de Ensenada, 2013.
- [7] C. C. Sánchez Zuleta, L. M. Giraldo Marín, C. C. Piedrahita Escobar, I. Bonet, C. Lochmüller, M. S. Tabares Betancur y A. Peña, «Análisis comparativo entre: «el análisis exploratorio de datos» y los modelos de «árboles de decisión» y «k-means» en el diagnóstico de la malignidad en algunos exámenes de cáncer de mama. Un estudio de caso,» *Espacios*, vol. 39, nº 28, p. 21, 2018.
- [8] O. L. Mangasarian, W. N. Street y W. H. Wolberg, «Breast Cancer Diagnosis and Prognosis Via Linear Programming,» *Operations Research*, vol. 43, nº 4, pp. 548-725, 1995.
- [9] Redacción Digital El Herald de México, «heraldodemexico.com,» 4 octubre 2019. [En línea]. Available: <https://heraldodemexico.com.mx/pais/cuantos-casos-de-cancer-de-mama-son-atendidos-en-mexico-anualmente/>. [Último acceso: 4 marzo 2020].
- [10] H. Sahu, S. Shrma y S. Gondhalakar, «A Brief Overview on Data Mining Survey,» *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, vol. 1, nº 3, 2011.
- [11] O. Maimon y L. Rokach, The Data Mining and Knowledge Discovery Handbook, Springer, 2010.
- [12] M. C. Carollo Limeres, Regresión lineal simple. Apuntes del Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, 2012.
- [13] R. S. Pressman, Ingeniería de Software: Un enfoque práctico, Ciudad de México: McGraw Hill, 2010.

8. Alumnos y directores

Ortiz Rivas Julio César. - Alumno de la carrera de Ing. en Sistemas Computacionales en ESCOM, Especialidad Sistemas, Boleta: 2017631191, Tel. 6691541466, email julio_cesar502@hotmail.com.

CARÁCTER: Confidencial
FUNDAMENTO LEGAL: Artículo 11 Fracc. V y Artículos 108, 113 y 117 de la Ley Federal de Transparencia y Acceso a la Información Pública.
PARTES CONFIDENCIALES: Número de boleta y teléfono.

Firma: _____





Ocampo Botello Fabiola. - Ingeniero en Sistemas Computacionales por el Instituto Tecnológico de Ciudad Madero, Maestría en Ciencias de la Computación por el CINVESTAV-IPN y Doctorado en Educación Internacional por la Universidad Autónoma de Tamaulipas (UAT). Profesora de la Escuela Superior de Cómputo (ESCOM) desde 1995, Tel.: 57296000 Ext.: 52082, email: focampob@ipn.mx.


Firma: _____

Título del TT: Sistema de apoyo a la detección de tumoración maligna en tejido mamario.



Actividad	SEP1	SEP2	OCT1	OCT2	NOV1	NOV2	DIC	ENE	FEB1	FEB2	MAR1	MAR2	ABR1	ABR2	MAY	JUN
Análisis de los datos históricos de tejidos ya catalogados para corregir datos sucios.																
Seleccionar los datos que serán de utilidad para el proceso predictivo.																
Evaluar técnicas de minería de datos más adecuadas al conjunto de datos.																
Selección de la técnica para realizar la predicción y preparar los datos para adaptarlos a la técnica de minería seleccionada.																
Evaluar los patrones generados por el algoritmo predictivo.																
Entrega del Incremento #1: Modelo predictivo.																
Análisis de requerimientos del sistema. Al final de esta fase se determinará el incremento #2.																
Evaluación de TT I.																
Modelado de requerimientos.																
Generación de código.																
Pruebas del sistema.																
Entrega del Incremento #2. Este se determinará en el análisis de requerimientos del sistema.																
Análisis y modelado de los nuevos requerimientos identificados.																
Generación de código.																
Pruebas del sistema.																
Entrega del Incremento #3: Sistema terminado.																
Generación del manual de usuario.																
Generación del reporte técnico.																
Evaluación de TT II.																


Por motivo de la pandemia del COVID-19 se adjunta un acuse de recibido por parte de la directora de este TT (Fabiola Ocampo Botello) a manera de firma.






Julio Ortiz
focampob@ipn.mx

Ayer




ProtocoloTT_2020-A069_Reestructurado.pdf
PDF - 389 KB




Buenas tardes profesora Fabiola

Por este medio le hago llegar el documento con la reestructuración del protocolo asignado con el número 2020-A069 "Sistema de apoyo a la detección de tumoración maligna en el tejido mamario".


Agradecería si pudiera responder que lo ha recibido, a manera de acuse de recibo, ya que debo enviar esa evidencia a la CATT.

Sin más por el momento, agradezco su atención. Quedo en espera de su respuesta.

Que tenga una excelente tarde.



Fabiola Ocampo Botello
Julio Ortiz

2:34 p. m.


Hola Julio

Buen día.

He revisado y hemos comentado las observaciones que hicieron los sinodales, atendimos las que a nuestro juicio es posible expresar en este momento y quedaron plasmadas en esta nueva versión.

Recibe un cordial saludo,

Fabiola Ocampo Botello