APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO DEL CÁNCER DE MAMA

Autores:

ANDRÉS CAMILO GONZÁLEZ OTERO

JUAN FRANCISCO ALMENARES ARAGÓN

UNIVERSIDAD POPULAR DEL CESAR

FACULTAD DE INGENIERÍAS Y TECNOLÓGICAS

PROGRAMA INGENIERÍA DE SISTEMAS

VALLEDUPAR – CESAR

2022

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO DEL CÁNCER DE MAMA

Autores:

ANDRÉS CAMILO GONZÁLEZ OTERO

JUAN FRANCISCO ALMENARES ARAGÓN

Trabajo de grado presentado como requisito para optar por el título de

INGENIERO DE SISTEMAS

Director:

ALVARO OÑATE BOWEN

INGENIERO DE SISTEMAS

Linea de investigacion:

TRANSFORMACIÓN DIGITAL

UNIVERSIDAD POPULAR DEL CESAR

FACULTAD DE INGENIERÍAS Y TECNOLÓGICAS

PROGRAMA INGENIERÍA DE SISTEMAS

VALLEDUPAR – CESAR

2022

NOTA DE ACEPTACIÓN:

_____

_____

_____

_____

_____

PRESIDENTE DEL JURADO

_____

JURADO

_____

JURADO

"El éxito no es un accidente. Es trabajo duro, perseverancia, aprendizaje, estudio, sacrificio y lo más importante de todo, amor por lo que estás haciendo o aprendiendo a hacer".

Pelé.

AGRADECIMIENTOS

Primeramente, damos gracias a Dios por darnos sabiduría, entendimiento y permitirnos realizar nuestro proyecto de grado, crecer a nivel personal y profesional, a nuestros padres y familiares por todo su apoyo siendo fundamentales anímicamente y emocionalmente a lo largo de nuestro periodo académico, a la Liga Contra el Cancer – Seccional Cesar por abrirnos las puertas, por la confianza brindada y su paciencia, por supuesto a nuestra universidad y a nuestros profesores, en especial a nuestro director y profesor Alvaro Oñate Bowen por compartirnos sus conocimientos con tanto profesionalismo y entrega.

Finalmente, agradecemos a quienes lean nuestro documento de tesis deseándoles éxitos y que nuestro proyecto les sea de gran ayuda y aporte a sus conocimientos.

Dios los bendiga.

RESUMEN

Este trabajo de grado se enmarca en la construcción de modelos a través de la metodología CRISD - DM para el diagnóstico del cáncer de mama en benigno y maligno integrados en un aplicativo web para la gestión de este proceso en la Liga Contra el Cáncer – Seccional Cesar. Por lo tanto, este proyecto está dividido en tres etapas las cuales son: la descripción general del proyecto, la descripción situacional donde se presenta la problemática de estudio, se planean los objetivos y definen las metodologías y aspectos de la investigación, y por ultimo, se realizo el desarrollo cientifico tecnologico donde se aplicó la metodología propuesta, se exponen el análisis de requerimientos, el diseño del sistema y socialización del aplicativo web llamado Diagnosis Breast Cancer. Para finalizar se presentan los resultados del proyecto donde se aplicaron un total de siete modelos: LogisticRegression (exactitud: 55%), KNN (exactitud: 75%), SVM (exactitud: 66%), DecisionTree (exactitud: 100%), NaiveBayes (exactitud: 65%), RandomForest (exactitud: 99%) y el GradientBoosting (exactitud: 100%), fueron elegidos solo tres modelos para la integración al aplicativo web debido a sus altos porcentajes el DecisionTree, GradientBoosting y RandomForest los cuales permiten de manera efectiva y eficaz el diagnóstico del cáncer de mama.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE GRÁFICOS

INTRODUCCIÓN

El uso de técnicas de minería de datos da lugar al descubrimiento de información y a la adquisición de conocimiento útil a partir del estudio y análisis de diversos datos, de manera automatica o semiautomatica, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto, de tal manera, que en la actualidad es un área importante de investigación y desarrollo que aporta gran valor a todo tipo de organizaciones.

En consideración, básicamente, la minería de datos surge para intentar comprender el contenido de un repositorio de datos, con este fin, hace uso de prácticas estadísticas y en algunos casos, de algoritmos próximos a la inteligencia artificial y a las redes neurales para llevar a cabo tareas como la clasificación o predicción de resultados donde los datos son la materia prima bruta. Entonces, desde que el usuario de dichos datos les atribuye algún significado especial pasan a convertirse en información valiosa que representa de alguna manera un valor agregado.

En este sentido, para este proyecto este tipo de técnicas facilita obtener a través del análisis del conjunto de datos denominado *Breast Cancer Wisconsin (Diagnostic) Data Set* y obtenido del repositorio de datos libre UCI tomar decisiones en la Liga Contra el Cancer – Seccional Cesar enfocadas a diagnosticar el cáncer de mama y a aprovechar de manera racional los resultados obtenidos .

Es precisamente la tendencia de que cada vez más las mujeres sufren de este tipo de cáncer donde la minería de datos ha venido a resolver este y otro tipo de problemas en el ámbito médico, la clasificación de variables con sus determinadas características con el fin de profundizar y de alguna manera dar solución a través de este tipo de tecnologías.

Por consiguiente, este estudio presenta el desarrollo y los resultados de todo un proceso investigativo realizado para poder diagnosticar y predecir el tipo de cáncer de mama en benigno y maligno. En efecto, se aplicará la metodología CRISP – DM que comprende etapas para el análisis del problema y el estudio de la estructuración de los datos para la construcción de modelos que constituyen el proceso principal de la minería de datos. Luego, de la aplicación de la minería de datos se desarrolla un aplicativo web que permita

integrar los modelos construidos para facilitar el despliegue y uso de los mismos en la Liga Contra el Cáncer – Seccional Cesar.

SECCIÓN I: DESCRIPCIÓN GENERAL

1.1 TÍTULO DEL PROYECTO

Aplicación de técnicas de minería de datos para la detección y el diagnóstico del cáncer de mama.

1.2 DIRECCIÓN DE EJECUCIÓN DEL PROYECTO

Liga contra el cáncer - Seccional Cesar. Cra. 19 #16-191, Valledupar – Cesar.

1.3 LAPSO DE EJECUCIÓN DEL PROYECTO

Cuatro (4) meses y ajuste a calendario de ejecución por las partes integradoras.

1.4 ORGANISMO Y SECCIÓN RESPONSABLE

Liga contra el cáncer - Seccional Cesar.

1.5 INFORMACIÓN DE CONTACTO DE LOS ESTUDIANTES

NOMBRES	APELLIDOS	CÉDULA	TELÉFONO	CORREO
Andrés Camilo	González Otero	1122414594	3012137806	acamilogonzalez@unicesar.edu.co
Juan Francisco	Almenares Aragón	1003233568	3122885907	juanfranalmenares@gmail.com

1.6 LÍNEA DE INVESTIGACIÓN: Transformación Digital

1.6.1 SUBLÍNEA DE INVESTIGACIÓN: Big Data y Analytics

1.6.2 AREA: Data Mining

1.6.3 GRUPO: GISICO

SECCIÓN II: DESCRIPCIÓN SITUACIONAL

2.1 IDENTIFICACIÓN DEL PROBLEMA

El cáncer es una enfermedad en constante evolución, que afecta a un gran número de individuos en todo el mundo. Según la Organización Mundial de la Salud mejor conocida por sus siglas en español OMS, en su reporte global sobre el cáncer, una de cada seis personas en el mundo muere de cáncer cada año, una cifra que va en aumento y que solo en 2018, el año más reciente, de los que se disponen datos, aproximadamente 9,6 millones de personas han muerto de cáncer [1]. Asimismo, el cáncer que más afecta a las mujeres en el mundo es el de mama, el cual representa el 16% de este tipo de tumor cancerígeno en mujeres [2].

En este orden de ideas, el cáncer de mama ocupa un lugar destacado en lista de mortalidad de la población femenina en Colombia y según cifras estadísticas alrededor de 2.120 mujeres pierden la batalla a causa de esta enfermedad mortal [3]. Por otro lado, Nubia Bautista, subdirectora de Enfermedades No Transmisibles, anunció gran preocupación en la salud pública por este tipo de cáncer en la población femenina que se va haciendo cada vez más notable no solo a nivel nacional, sino también a nivel mundial, por lo que es necesario y de suma importancia fortalecer los mensajes y campañas de prevención en las que se den a conocer las circunstancias o condiciones, el estilo de vida, signos y síntomas de esta enfermedad, el autocuidado y la importancia de acudir a alguna entidad prestadora del servicio de salud oportunamente y de igual manera Bautista declaró que “la mortalidad femenina por cáncer de mama aumenta cada año. Si bien hubo 2.243 muertes en 2009, hubo 3.535 en 2019, un aumento del 36,5% y en una década fueron fallecidas 22.174 mujeres por esta enfermedad entre las edades de 30 y 70 años” [4].

Ante esta aseveración, la Liga contra el cáncer - Seccional Cesar, entendiendo la necesidad de una asistencia médico completo para las mujeres que padecen esta enfermedad silenciosa, siendo geográficamente Valledupar en el norte del Cesar y principal referente en la atención médica del departamento, aunado al deseo de un gremio de médicos profesionales, radiólogos y oncólogos capaces y organizados, han querido

avanzar, facilitar y mejorar los diagnósticos para tratar, si es el caso, a tiempo a las mujeres que padezcan de este tipo cáncer en sus etapas más tempranas.
Por ello, esta institución sin ánimos de lucro, tiene como misión ser una institución prestadora del servicio de salud (IPS) para la prevención del cáncer con el propósito de minimizar la morbilidad y mortalidad por cáncer, contribuyendo a una mejor calidad de vida de la población del departamento del cesar y las regiones aledañas a través de medidas para educar, promover y tratar esta enfermedad y para rehabilitar a los pacientes y como visión ser un modelo de excelencia en cuanto a la generación del conocimiento científico para la satisfacción de todos los usuarios para la conversión de una sociedad sana.

Ahora bien, en entrevistas y reuniones sostenidas con el ingeniero de sistemas y la directora ejecutiva y de calidad de la Liga contra el cáncer - Seccional Cesar (Ver Anexo I), manifestaron que:

- A pesar de su trabajo, ha habido un aumento significativo en el número de casos de este tipo de cáncer en los últimos años, es decir, el cáncer de mama, esto debido, a que los pacientes acuden a las consultas en los centros de salud una vez ya la enfermedad se encuentra muy avanzada, ya sea por el poco conocimiento acerca de esta enfermedad como los son los primeros síntomas y malestares o por la poca conciencia que tienen las personas acerca de las posibilidades de sufrir de este tipo de cáncer.

- Por otra parte, a pesar de ser una institución importante en la región y, sobre todo, pensando en su potencial como institución prestadora de salud (IPS), que tiene como fin prevenir distintos tipos de cáncer, entre los recursos que disponen para llevar a cabo una consulta o un examen médico, estos no cuentan con algún software o infraestructura informática que les permita o les sea de apoyo a sus profesionales de la salud, realizar diagnósticos.

- Adicionalmente, otra falencia que se pudo determinar es el de no estudiar los datos obtenidos para prevenir casos futuros, ya que los datos obtenidos en cada

uno de los casos son utilizados para realizar informes solo de tipo estadísticos ante los entes de control.

Por lo anterior, se evidencia la ausencia de nuevas tecnologías que al poder detectar o diagnosticar el cáncer de mama en sus etapas iniciales e igualmente la ausencia con respecto a la realización de estudios a partir de la recolección de los datos obtenidos de sus pacientes, con las que se puedan crear canales de información para concientizar más a las mujeres e incluso a hombres (ya que esta enfermedad también puede desarrollarse en hombres) sobre esta enfermedad y agilizar los procesos que conllevan los respectivos diagnósticos.

Es por esto, que se propone desarrollar un sistema inteligente, capaz de detectar y diagnosticar el cáncer de mama a través de una serie de datos apoyados en los reportes clínicos de mujeres con esta enfermedad en la Liga de lucha contra el cáncer - Seccional Cesar y así optimizar el trabajo de los médicos. El sistema inteligente integrará varias técnicas de minería de datos para determinar cuál presenta mejor desempeño para la evaluación del modelo. En la etapa de entrenamiento del proyecto los modelos serán entrenados con un conjunto de datos que se encuentra en la plataforma UCI Machine Learning Repository.

En la etapa de evaluación de los modelos se utilizarán el conjunto de datos de las mujeres con esta enfermedad registradas en la Liga de lucha contra el cáncer - Seccional Cesar, para evaluar las bondades del modelo. En este mismo sentido, el sistema estará desarrollado en ambiente web y contará con los siguientes módulos: (a) gestión de pacientes, para llevar un registro e inspección en cada una de las personas que se realizan el examen, (b) el módulo de predicción o diagnóstico, para determinar la localización del tumor en el paciente, en donde este diagnóstico servirá de base para la aplicación de tratamientos médicos. Finalmente, (c) el módulo de informes, donde se visualizarán los datos sobre los diagnósticos realizados en la Liga contra el cáncer - Seccional Cesar.

2.1.1 FORMULACIÓN DEL PROBLEMA

Con base a los planteamientos anteriormente expuestos, se formula el interrogante central de investigación:

¿De qué manera la aplicación de modelos descriptivos y predictivos permitirían detectar y diagnosticar el cáncer de mama en la Liga contra el cáncer - Seccional Cesar?.

2.2 JUSTIFICACIÓN DEL PROBLEMA

En la actualidad, las organizaciones tienen cada vez más un mayor volumen de información siendo el activo más importante. Asimismo, el aumento en la última década en Colombia de los casos de cáncer de mama, ha motivado a buscar una manera para detectar y diagnosticar a tiempo este tipo de cáncer y poder así, de esta manera reducir la mortalidad y morbilidad consecuencia de esta, creando soluciones innovadoras y significativas para mitigar esta problemática. De ahí, la implementación de técnica de Minería de Datos es de suma importancia ya que permiten descubrir patrones e información en conjunto de datos históricos que pueden ser de gran valor y que apoyan la toma de decisiones.

Por este motivo, este proyecto se centra en la aplicación de técnicas de minería de datos para crear modelos descriptivos y predictivos que permitan descubrir información oculta en el conjunto de datos para ser examinada y diagnosticada en una etapa temprana y en el momento oportuno a las personas que padecen de cáncer de mama. Por consiguiente, la información que se obtendrá beneficiará no solo a las personas posiblemente afectadas por esta enfermedad, sino también a todas las entidades y profesionales de la salud, otorgando un mejor servicio de salud a tiempo.

En el mismo sentido, el estudio conforma un aporte teórico, ya que se recolecta información ordenada a través de teorías de expertos en la temática e investigaciones realizadas con anterioridad, ya que su aplicación y conclusiones servirán de base para el tema en desarrollo. De igual manera, este proyecto se justifica teóricamente por cuanto la misma brindará una serie de teorías, principios y conceptos que permitan el razonamiento científico de los aspectos referidos a la problemática tratada, aquellas asociadas al cáncer de mama, esto a partir de una bibliografía consultada y actualizada, centrada en

información que contribuye, además con conocimiento obtenido a partir de las experiencias de años de servicio de un personal calificado en este campo del saber.

La investigación también es relevante desde un punto de vista práctico, ya que sus conclusiones y recomendaciones se pueden utilizar para mejorar la atención médica por parte de Liga Contra el Cáncer - Seccional Cesar en Valledupar, derivando igualmente su aplicabilidad para otras organizaciones o áreas cuyas características sean similares a la que se analiza. De igual forma, dado que la problemática es muy notable, es conveniente la creación de un aplicativo web conforme a las aplicaciones de las técnicas y modelos de minería de datos a desarrollar, para que se pueda detectar y diagnosticar este cáncer conforme a los datos que se suministren en este entorno y de esta manera contribuir a la solución de dicha problemática y prevenir esta enfermedad a tiempo.

Con relación al valor metodológico del estudio, se considerará oportuno para el impulso del proyecto la metodología CRISP - DM que se aplicará entendiendo que es de gran ayuda, ya que establece las fases que todo proyecto de minería de datos debe cumplir para que se desarrolle el proyecto de manera exitosa. De igual manera, es fundamental que los hallazgos alcanzados mediante diferentes técnicas o modelos que se implementaran sean validados y verificados para obtener información de relevancia, la cual permitirá lograr el cumplimiento de los objetivos trazados. Además, podrá servir de consulta de apoyo para futuras investigaciones referidas a la misma temática de estudio.

2.3 OBJETIVOS

2.3.1 OBJETIVO GENERAL

Implementar técnicas de minería de datos para la detección y el diagnóstico temprano del cáncer de mama en la Liga Contra el Cáncer - Seccional Cesar.

2.3.2 OBJETIVOS ESPECÍFICOS

- Analizar los procesos que se realizan en el diagnóstico del cáncer de mama en la Liga Contra el Cáncer - Seccional Cesar para obtener un conocimiento preliminar de los mismos y la formulación de hipótesis.

- Diseñar modelos descriptivos y de clasificación para caracterizar y diagnosticar el tipo de cáncer de mama benigno o maligno.

- Evaluar los modelos propuestos analizando los resultados obtenidos respecto al porcentaje de precisión y tasa de errores.

- Implementar el aplicativo web con todos los módulos integrados en la Liga Contra el Cáncer - Seccional Cesar.

2.4 MARCO REFERENCIAL

A través de la historia hemos visto cómo la revolución digital ha hecho posible la adquisición del conocimiento a través del almacenamiento y el procesamiento de los datos que cada vez son más fáciles de capturar [5] y con las nuevas tecnologías como es la minería de datos el cual se conceptualiza como, el proceso de obtener información productiva y que antes era desconocida, desde grandes conjuntos de datos o bases de datos de distintos formatos [6] y otras tecnologías de avanzada como son la Inteligencia Artificial, el Aprendizaje Automático y el Aprendizaje Profundo.

De igual manera, estas tecnologías permiten extraer valor a partir de la información que recopila y gestiona las diferentes entidades indistintamente de su naturaleza y objeto social, lo que contribuye a aunar esfuerzos hacia aquellos conjuntos de datos auténticos que pueden ser de diversas índoles. En este caso, el cáncer de mama presenta una alta mortalidad de mujeres siendo de suma importancia este tipo de herramientas para los estudios en donde son de gran beneficio para el diagnóstico temprano.

2.4.1 ANTECEDENTES

Existen varios estudios en donde se aplica la minería de datos para el encuentro de conocimiento en repositorios clínicos. Algunos de ellos tratan problemas como la evaluación del riesgo de padecer alguna enfermedad, o para asistir en el tratamiento y pronóstico de otras enfermedades. En esta sección se presentan algunos trabajos

encontrados producto de la revisión de artículos científicos sobre la temática aquí mencionada y los resultados que se obtuvieron, seleccionando las siguientes investigaciones:

A nivel internacional

En primer lugar, el artículo científico titulado *BREAST CANCER RISK ASSESSMENT AND EARLY DIAGNOSIS USING PRINCIPAL COMPONENT ANALYSIS AND SUPPORT VECTOR MACHINE TECHNIQUES* [7]*,* utilizaron un enfoque híbrido entre el Análisis de Componentes Principales (siglas en inglés PCA, Principal Component Analysis) y Máquinas de Soporte Vectorial (siglas en inglés SVM, Support Vector Machines, SVMs) con la cual, crearon un modelo para evaluar el riesgo y la exactitud de un análisis diagnóstico de cáncer de mama en su etapa más temprana. Dicho modelo, obtuvo una precisión del 97.62% y con base a este resultado se concluyó que al aplicar estas técnicas ayudan potencialmente a procesar datos con respecto al cáncer de mama y clasificar a los pacientes con respecto a las categorías identificadas como probables e improbables y el tipo de cáncer malignos y benignos.

Igualmente, la investigación realizada en China que tiene por título *BREAST CANCER DETECTION AND DIAGNOSIS USING MAMMOGRAPHIC DATA: SYSTEMATIC REVIEW* [8] el propósito estuvo dirigido a una exploración de la literatura sobre el Aprendizaje Automático (siglas en inglés ML, Machine Learning) y el Aprendizaje Profundo (siglas en inglés DL, Machine Learning) para la aplicación del dictamen acerca del cáncer de mama.

Por lo tanto, este artículo investigativo se realizó la descripción de distintas técnicas de Aprendizaje Automático y el Aprendizaje Profundo teniendo en cuenta revisiones hechas en bibliotecas y repositorios de investigaciones como PubMed, Google Scholar, MEDLINE, ScienceDirect, Springer y Web of Science durante los últimos 5 años con respecto a la realización de este artículo, en las que utilizaron un dataset con datos de mamografías e implementaron métodos como el perfilado y procesamiento de mamas por medio de sistemas de Diagnósticos asistido por Computadora (siglas en inglés CAD, Computer aided diagnosis), técnicas de extracción de características convencionales como el Aprendizaje Automático, el Análisis Discriminante Lineal (siglas en inglés LDA,

Linear Discriminant Analysis), técnicas de filtrado como la prueba chi-cuadrado, técnicas de clasificación en las que son de gran utilidad para determinar las zonas o regiones normales o cancerígenas con métodos en los que se incluyen Máquinas de Soporte Vectorial (siglas en inglés SVM, Support Vector Machines, SVMs), Redes Neuronales Artificiales (siglas en inglés ANN, Artificial Neural Networks), K – vecinos más cercanos (siglas en inglés KNN, k-Nearest Neighbors), Árboles de Decisión, entre otros más y por último el Aprendizaje Profundo que mejora significativamente los resultados en comparación a los métodos Aprendizaje Automático con respecto a la clasificación de imágenes.

En conclusión, se realizó un análisis que revela la importancia del uso de estos métodos, con la gran posibilidad de analizar casos clínicos y mejorar los procesos y la precisión de los sistemas de diagnósticos asistido por computadora. De igual manera, determinan con la revisión de la literatura, que se pueden presentar algunos casos en donde el conjunto de datos presenta densidades mamarias heterogéneas haciendo difícil la detección y clasificación de los tejidos cancerígenos y, asimismo, el potencial apoyo que presentan los métodos aprendizaje profundo para mejorar los diagnósticos del cáncer de mama a comparación de los métodos aprendizaje automático.

Otro artículo relevante fue realizado por el laboratorio SI2M del Departamento de Ciencias de la Computación del Instituto Nacional de Estadística y Economía Aplicada de la ciudad de Rabat, Marruecos titulado *OPTIMIZATION OF K-NN ALGORITHM BY CLUSTERING AND RELIABILITY COEFFICIENTS: APPLICATION TO BREAST-CANCER DIAGNOSIS* [9], en donde propusieron mejorar el rendimiento del algoritmo K – vecinos más cercanos (siglas en inglés KNN, k-Nearest Neighbors) con ayuda de la técnica de agrupamiento utilizando el algoritmo k-medias (en inglés, k-means).

Por consiguiente, utilizaron un conjunto de datos sobre el cáncer de mama compuesto por imágenes digitalizadas que contienen 355 cáncer de tipo benignos y 210 malignos, y otros atributos como el radio, textura, dimensión fractal, etc., en el que metodológicamente agruparon las instancias que representan una clase, seleccionaron los atributos más significativos y midieron los coeficientes de confiabilidad, para luego aplicar los algoritmos de clasificación Máquinas Soporte de Vectores (siglas en inglés SVM, Support Vector

Machines, SVMs), Redes Neuronales Artificiales (siglas en inglés ANN, Artificial Neural Networks), Naive Bayes y K – vecinos más cercanos (siglas en inglés KNN, k-Nearest Neighbors) para comparar los resultados. Finalmente, el resultado del modelo propuesto tuvo un acierto del 94% y se llegó a la conclusión que las tareas, técnicas y algoritmos que se implementaron son de utilidad para analizar conjuntos de datos grandes porque disminuye las instancias y los atributos insignificantes.

De la misma manera, en la India se realizó la investigación titulada *STUDY OF BREAST CANCER DETECTION METHODS USING IMAGE PROCESSING WITH DATA MINING TECHNIQUES* [10]. Esta investigación, tuvo como finalidad analizar y comparar diversas técnicas para la detección del cáncer mama como los métodos de minería de datos y el procesamiento de imágenes. Por lo tanto, se propuso una metodología para la identificación del cáncer maligno y benigno en mamografías digitales compuestas por actividades como, adquisición del conjunto de datos, selección de las imágenes de entrada, pre procesamiento, extracción y selección de características, aplicación como también prueba de los modelos de clasificación. Así pues, esta investigación determinó a través de una extensa revisión literaria que el algoritmo Máquinas Soporte de Vectores, es uno de los más eficaces para la clasificación del cáncer de mama.

Otro artículo que se encontró relevante fue realizado en España titulado *MINERÍA DE DATOS APLICADA A LA DETECCIÓN DE CÁNCER DE MAMA* [11]. Se tuvo como propósito realizar un análisis de datos para poder identificar ciertos patrones y así identificar potenciales casos de cáncer de mama a través de la minería de datos.

Por tal motivo, metodológicamente se utilizó el ciclo de vida de proyectos de minería de datos Crisp – Dm y un conjunto de datos llamado Wisconsin Breast Cancer Dataset, el cual consta de 699 registros y 9 atributos sobre las observaciones de los tumores que fueron complementados con 2 atributos más, donde se clasificaban los tumores como benigno o maligno. Por lo cual, aplicaron herramientas como el conjunto de librerías Java para extraer conocimientos de una base de datos conocida como Weka, y varios técnicas de minería de datos como: el algoritmo de clasificación supervisada (siglas en inglés KNN, k-Nearest Neighbors), el perceptrón multicapa (siglas en inglés, MultiLayer Perceptron) que permitió resolver problemas que no son linealmente separables, el clasificador

probabilístico Naive Bayes, algoritmos de árbol de decisión C4, el SMO que implementa un algoritmo secuencial de optimización mínima para, entre otras cosas, entrenar una Support Vector Machines, SVMs, entre otros.

Se concluyó que aunque los atributos pueden tener más o menos influencia en cómo se clasifica un tumor como benigno o maligno, y que, en general, cuanto menor es el valor del atributo, mayor es la probabilidad de que el caso sea benigno cómo es decir, en el caso del atributo mitosis se dedujo que su remoción del estudio no afectaría significativamente los resultados. Por último, las técnicas utilizadas tienen una tasa de error del 3 al 5% al momento de extraer los resultados buscados para este estudio, siendo el clasificador de optimización mínima secuencial el que presentó la menor tasa de error mientras que con la aplicación de Naive Bayes se obtuvo un error mayor a comparación con la aplicación del algoritmo K – vecinos más cercanos, siendo el error del RFB coincidente con el error del Naive Bayes.

Por otro lado, el artículo realizado en México y titulado *ESTUDIO DE HERRAMIENTAS DE MINERÍA DE DATOS PARA LA TAREA DE CLASIFICACIÓN* [12]*,* tuvo como principal objetivo valorar la práctica en modelos de codificación construidos con algoritmos de clasificación suministrado por HMDs (herramientas de minería de datos), evaluando experimentalmente bajo las mismas condiciones algoritmos como el de clasificación Máquinas Vectorial de Soporte, árboles de decisión, RandomForest y K – vecinos más cercanos. Se desarrolló una aplicación web usando base de datos con enfermedades como diabetes, cáncer de mama cada vez más recurrentes en las mujeres.

Por consiguiente, se realizaron tareas tales como: integrar las diversas fuentes de datos médicos en relación con estas dos condiciones, luego usar técnicas de preprocesamiento o transformación de datos, luego construir los modelos de clasificación para llevar la realización de predicciones sobre la clase, es decir, determinar si el paciente tiene o no la enfermedad, de manera que se realizaron las pruebas en diferentes herramientas como Weka, RapidMiner y R, tomando los resultados experimentales.

Con esto, se llegó a la conclusión de que la herramienta que mejor clasificó los datos fue R, mientras que los algoritmos que mejor desempeño tuvieron fueron el árbol de

decisiones para pronosticar diabetes y el algoritmo Random Forest el cáncer de mama en pacientes. Igualmente, se observó que las técnicas de clasificación sumada a la aplicación web que se desarrolló permitieron tener un diagnóstico confiable y rápido teniendo en cuenta los resultados previstos por las técnicas que se implementaron.

Asimismo, el artículo realizado en la India titulado *PERFORMANCE ANALYSIS OF DATA MINING ALGORITHMS FOR BREAST CANCER CELL DETECTION USING NAÏVEBAYES, LOGISTIC REGRESSION AND DECISION TREE* [13], el cual tiene como objetivos encontrar el subconjunto más pequeño de particularidades que garanticen la correcta y exitosa clasificación del tumor canceroso en mama, y por consiguiente estudiar conforme a diferentes métodos de clasificación como el Naive Bayes, Regresión Logística, Árbol de Decisión presentan un mejor resultado.

Así pues, con el conjunto de datos WBC de UCI se llevó a cabo el procesamiento de las actividades como la limpieza y eliminación de datos perdidos, la reducción de la dimensionalidad de los datos aplicando el coeficiente de correlación parcial (siglas en inglés PCC, Partial Correlation Coefficients), la selección y transformación de las variables y la implementación de métodos de clasificación dividiendo el acumulado de datos en un 70% para el adiestramiento y el 30% para llevar a cabo pruebas.

Por último, aplicando los diferentes métodos de clasificación el autor tuvo como resultado una precisión para el Naive Bayes del 94.40%, Regresión Logística el 97.90% y para el Árbol de Decisión el 96.50%, siendo entonces el método de Regresión Logística el que mejor tuvo resultados y llegando a la conclusión que la aplicación de estas técnicas ayudan a encontrar información valiosa generando conocimiento para el diagnóstico en futuros casos.

También, en el artículo científico titulado *USING MACHINE LEARNING ALGORITHMS FOR BREAST CANCER RISK PREDICTION AND DIAGNOSIS* [14] se estudiaron métodos de clasificación y extracción de datos, utilizados en muchas ocasiones para generar diagnósticos y análisis para la toma de decisiones en el campo de la medicina. Por tanto, el objetivo de esta investigación fue evaluar la correlación de la clasificación de datos teniendo en cuenta la efectividad y eficiencia de cada uno de los algoritmos en

cuanto a exactitud y precisión, admitiendo la investigación la aplicación de algoritmos de aprendizaje automático como: Máquinas Vectoriales, Árbol de decisiones (C4.5), Naive Bayes y sobre el conjunto de datos que lleva por nombre Cáncer de Mama de Wisconsin (original).

Por tal motivo, para llevar a cabo la experimentación y el cumplimiento del objetivo propuesto, la experimentación se realizó con la herramienta WEKA aplicando validación cruzada, para evaluar y determinar la efectividad y eficiencia de los modelos, en la que el clasificador C4.5 tuvo un porcentaje de precisión y una tasa de error de 95.13%; el SVM de 97.13% de precisión y una tasa de error de 45; el NB de 95.99% de precisión y una tasa de error de 35.58 y el K-NN de 95.27% de precisión y una tasa de error de 44.77. Con esto pudieron concluir, que el algoritmo de clasificación que mejor resultados tiene es el SMV siendo el que ofrece el mayor porcentaje de precisión y con la tasa de error más baja.

A nivel Nacional

En primer lugar, el artículo realizado en la ciudad de Bucaramanga titulado *ANÁLISIS COMPARATIVO DE PREDICCIÓN DENTRO DE BASES DE DATOS DE CÁNCER: UNA APLICACIÓN DE APRENDIZAJE AUTOMÁTICO* [15], el cual tuvo como objetivo realizar comparaciones en términos de rendimientos en las predicciones de los algoritmos y técnicas Regresión Logística, K – Vecinos más cercanos, el algoritmo de agrupamiento K-means, Random Forest, Máquinas de Soporte Vectorial, Análisis de Discriminantes Lineales, Naive Bayes y Perceptrón Multicapa sobre una plataforma que soporta 11 cánceres de mama procedentes de la Universidad de Wisconsin con sus respectivos datos.

De allí que, para cumplir con los objetivos del estudio, se procesaron los datos en los entornos de desarrollo Anaconda Spyder y H20. Luego, aplicaron cada uno de los modelos en la cual, las medidas para llevar a cabo la comparación del desempeño de los resultados se describieron con la media aritmética y la desviación estándar. En cuanto al análisis de los modelos, tomaron del conjunto de datos el 80% para el entrenamiento y el 20% de los datos restantes para las validaciones teniendo como resultado; media:

0.906044 – desviación estándar: 0.079336 para la Regresión Logística, media:0.834066 – desviación estándar:0.116214 para el Análisis discriminante lineal, media: 0.782967– desviación estándar: 0.127118 para k – vecinos más cercanos, media:0.841209 – desviación estándar: 0.127810 para Naive Bayes, media: 0.931800 – desviación estándar: 0.066529 para Máquinas de soporte vectorial, media: 0.931800 – desviación estándar: 0.066529, siendo este último el que mejor desempeño muestra.

Por otra parte, en cuanto al análisis de los modelos de aprendizaje no supervisado, implementan teniendo en cuenta las condiciones para los parámetros de cada uno de los modelos un resultado en precisión para el K-means (0.71), Random Forest (0.69) y Perceptrón Multicapa (0.95), siendo este último el que mejor resultado presenta.

Seguidamente, se seleccionó el artículo realizado en la ciudad de Pasto titulado *CARACTERIZACIÓN DE LA SUPERVIVENCIA DE MUJERES CON CÁNCER INVASIVO DE CUELLO UTERINO USANDO MINERÍA DE DATOS* [16]. Para la realización de la investigación se utilizó la base de datos de mujeres diagnosticadas con cáncer cervicouterino invasivo entre 1998 y 2002 del Registro Poblacional de Cáncer del Municipio de Pasto (Colombia) y la metodología CRISP - DM.

Para la realización de esta investigación, se aplicaron tres tareas de minería de datos; clasificación, asociación y agrupamiento. En cuanto a la tarea de clasificación, construyeron un modelo aplicando la técnica del árbol de decisión utilizando el algoritmo J48 que a su vez implementa el algoritmo C4.5, que permitió extraer las características de aquellos pacientes que sobrevivieron al cáncer y los que no sobrevivieron y para la evaluación del modelo dividieron los datos en dos conjuntos, uno de entrenamiento y otro de pruebas con lo que obtuvieron un porcentaje del 93.2% instancias correctamente clasificadas.

Asimismo, desarrollando la tarea de asociación descubrieron patrones de ocurrencia y características comunes entre las mujeres. Por ello, utilizaron el algoritmo a priori evaluando los resultados con un soporte mínimo del 10%, la confianza del 80% (ambas métricas que permiten medir la calidad de los patrones o reglas), un incremento de 0.5 y un número de 25 reglas a generar, teniendo como resultado una confianza del 100%. Por

último, desarrollaron la tarea de agrupación con la finalidad de encontrar grupos similares o clusters utilizando el algoritmo k-means y evaluando los resultados se utilizó un conjunto de datos de entrenamiento. Al implementar k-means los autores tomaron para k tres valores; k=2, k=4 y k=6, siendo para k=2 en el que encontraron los clusters más homogéneos.

Finalmente, se concluyó que existen modelos vinculados a las condiciones socioeconómicas de las mujeres del municipio de Pasto y a las características de las mujeres con cáncer invasivo de cuello uterino como sobrevivientes, si superan los 52 meses posteriores al momento del diagnóstico del cáncer, reiterando que de esta manera los hallazgos encontrados en otras investigaciones.

Asimismo, se seleccionó la investigación realizada en la ciudad de Medellín que lleva por título *DETECCIÓN AUTOMÁTICA DE MICROCALCIFICACIONES EN UNA MAMOGRAFÍA DIGITAL, USANDO TÉCNICAS DE INTELIGENCIA ARTIFICIAL* [17]. En esta investigación, los autores tomaron de la base de datos Nijmegen, un conjunto de imágenes de mamografías compuesto por 40 imágenes de 12 bits por pixel y 100 micrones por pixel, el cual contiene 13 clusters de microcalcificaciones benignas y 27 malignos. Por lo tanto, realizaron procesamientos digitales de imágenes a través del enfoque gaussiano de filtrado, con el cual llevaron a cabo comparaciones en las que lograron realzar el contraste entre las microcalcificaciones, es decir, pequeñas manchas de calcio que se pueden observar a través de una mamografía, y el tejido sano del seno.

Para esto, fue necesario reducir el ruido en las imágenes para luego llevar a cabo una segmentación mediante filtros DoG, con el propósito de recortar las zonas correspondientes a la mamá o al seno para que esta sea usada en procedimientos posteriores. De esta manera, se aplicó el algoritmo K-vecinos más cercanos, para clasificar y así poder establecer el nivel de benignidad o malignidad de las microcalcificaciones, validando los resultados a través de las curvas ROC, teniendo como resultado 17 características que fueron evaluadas mediante la técnica de correlación de datos con la finalidad de identificar en este caso 7 características de microcalcificaciones sospechosas.

Posteriormente, los autores concluyeron que los métodos aplicados para la realización de esta investigación, presentan la gran ventaja de la implementación de algoritmos de segmentación, con las se detectaron microcalcificaciones en las mamografías, lo cual permite minimizar el nivel de complejidad del algoritmo de clasificación implementado.

A nivel regional

Se seleccionó el artículo realizado en la ciudad de Barranquilla y titulado *METHOD BASED ON DATA MINING TECHNIQUES FOR BREAST CANCER RECURRENCE ANALYSIS*, en donde los autores utilizaron el conjunto de datos denominado Breast Cancer Data Set recogido del repositorio de datos UCI, teniendo como propósito de la investigación presentar la comparación algoritmos de clasificación de J48 y random forest, Naive Bayes y Naive Bayes Simple, SMO Poli-Kernel y SMO RBF-Kernel, integrado con el algoritmo de agrupamiento Simple K-Means para la generación de un modelo que permite la clasificación exitosa de pacientes que son o no recurrente luego de haber sido previamente operadas para el tratamiento de dicha enfermedad [18].

Por consiguiente, metodológicamente para el desarrollo de la investigación llevaron a cabo un preprocesamiento de los datos debido a que estos no estaban balanceados con respecto a su variable clase llamada recurrencia el cual contiene los valores no recurrentes el cual representa el 70.27% de los registros y el valor recurrentes representan el 29.7% de los registros. Es por eso, que utilizaron el filtro SMOTE para balancear los datos teniendo como resultado una distribución homogénea teniendo el valor no recurrente el 50.7% de los registros y el valor recurrente el 49.3% y consecutivamente una vez balanceados los datos verificaron si había datos atípicos y datos faltantes.

Posteriormente, llevaron a cabo el proceso de entrenamiento y prueba de los métodos de clasificación en donde compararon los algoritmos DT, Naive Bayes y Máquinas vectoriales a través de las métricas de precisión, evaluación de cobertura, tasa de verdaderos positivos, tasa de falsos positivos, determinando de dicha comparación que el mejor resultado fue proporcionado por los modelos SMO Poly-Kernel + Simple K-Means con un

98.5% de precisión, de recuperación, y de TPRATE y 0,2% de FPRATE. Cabe resaltar que todos estos procesos fueron realizados en la herramienta de minería de datos WEKA.

Finalmente, se concluyó que estas herramientas computacionales basadas en minería de datos son muy significativas e importantes para la localización de la recurrencia del tumor en mama de mujeres que habían sido previamente tratadas.

2.4.2 REFERENCIA TEÓRICA

2.4.2.1 CÁNCER DE MAMA

El cáncer de mama es una enfermedad que consiste en la multiplicación de células cancerígenas que se alojan en los tejidos mamarios que hacen parte de la mama conectados por tubos delgados llamados conductos. Existe una variedad de cáncer de mama, los cuales van a depender de las células mamarias que son vulnerables al padecimiento del cáncer, las mamas están compuestas de tres partes indispensables: a) los lobulillos que son glándulas que producen la leche materna y que en la mayoría de las veces es la parte de la mama donde comienza a aparecer el cáncer, b) los conductos, como su nombre lo indica se encargan de transportar la leche producida por los lobulillos a través de tubos al pezón, es otra parte sensible a producirse el cáncer de mama c) el tejido fibroso y adiposo que forman el tejido conectivo que tiene como función primordial sostener las partes de la mama en su conjunto.

El cáncer de mama se puede esparcir externamente fuera del tejido fibroso y adiposo a través de los vasos sanguíneos que irrigan sangre al interior de la mama, a diferencia de los vasos linfáticos que son los encargados del transporte del líquido llamado linfa. La mama también está provista de racimos llamados ganglios linfáticos como pequeñas estructuras alojadas cerca de las axilas y muy cerca del contorno de la mama que coadyuvan a evitar y/o combatir enfermedades infecciosas. Ahora bien, cuando el cáncer de mama se esparce a otras partes del cuerpo, se denomina metástasis.

Frecuentemente el cáncer de mama que más azota a las mujeres es el carcinoma ductal in situ, este tipo de cáncer se inicia en las células de los conductos, otro tipo de cáncer de

mamá es el carcinoma lobulillar muy común en ambos senos. Por otra parte, se conoce el cáncer inflamatorio, caracterizado por alcanzar a enrojecer e hinchar la mama manteniéndola caliente en todo momento.

El cáncer de mama es el problema de salud más importante a nivel mundial y es el más común entre las mujeres. La prevención a primera vista resulta algo complicado porque se desconoce la causa limitando su detección temprana, por lo que se recomienda que las mujeres con frecuencia debe realizarse examen de mamografía que es considera a nivel del orbe como el mecanismo de descubrimiento o localización oportuna, ya que la ciencia ha demostrado que este tipo de exploración tiene a reducir el índice de mortalidad por esta causa en mujeres confirmando su efectividad en dicho índice entre un 30 y 70% [19], sin embargo tiene ciertas limitaciones para los observadores humanos pero cuenta con limitaciones de observadores humanos (radiólogos) dar resultados precisos y consistentes debido a la cantidad de mamografías que se realizan [20].

2.4.2.1.1 IDENTIFICACIÓN DE TONOS CLARO, OSCURO Y GRIS

Para esta parte del proceso investigativo, todas las imágenes radiológicas se muestran en una placa que es un soporte plástico, todavía se usa en algunos lugares, pero con menos frecuencia y las imágenes se obtienen mediante una combinación de rayos “X” y luz sobre la membrana fotográfica, que a su vez, genera un retrato a la vista que luego se procesa en el cuarto oscuro mediante la concentración de químicos y posteriormente se cuelga “literalmente”, hasta su secado [21]. En este sentido, en caso de una urgencia, la placa se comentará químicamente húmeda; de ahí el significación de lectura húmeda para una interpretación inmediata [22].

Asimismo, esta forma de trabajar se mantuvo durante décadas, pero tenía dos grandes inconvenientes: a) se requería mucho espacio para almacenar cada vez más placas y b) Las placas sólo podían colocarse en un solo lugar, y ese lugar no era necesariamente el requerido para el tratamiento del paciente [19]. Luego surgió la radiografía digital, reemplazando la placa fotosensible siendo esta procesada por un mecanismo de lectura electrónico para capturar imágenes digitales que pueden almacenarse en el disco duro de un ordenador que funciona como servidor [23].

Desde entonces, los estudios radiológicos que se realizaban bajo el mecanismo de lectura electrónica eran almacenados como un historial del paciente en ordenadores donde fue posible archivar y conservar las imágenes. Es por eso que este sistema se denominó Archivo, comunicaciones y archivo de imágenes. Por consiguiente, todo tipo de imágenes se pueden almacenar y recuperar con los sistemas PACS incluidas las radiografías convencionales (CR), las imágenes de tomografía computarizada (TC), las imágenes de ultrasonidos y las resonancias magnéticas (MRI).

2.4.2.1.2 RADIOGRAFÍA CONVENCIONAL (PLACA SIMPLE)

Aquí es importante señalar que las imágenes creadas con radiación ionizante (rayos X), pero sin la adición de medios de contraste bien sea bario (Ba) o iodo (I) como oligoelemento, a menudo se denominan radiografías tradicionales o películas simples [24]. Generar estas imágenes resulta muy económico y se consigue realizar en cualquier espacio utilizando un mecanismo provisto de un dispositivo portátil o de mano.

Sigue siendo la prueba de imagen más utilizada en la actualidad. Demandan el uso de equipos de rayos X, en efecto, es un procedimiento de adquisición de imágenes (placa, marco o placa) y un sistema (equipo químico o de imágenes digitales) digital para procesar la imagen resultante). Las indicaciones radiográficas tradicionales más comunes son una radiografía de tórax, una radiografía de abdomen normal y prácticamente cualquier fotografía preliminar del conjunto óseo humano a fin de detectar la apariencia de una fractura o complicaciones en las articulaciones [19].

Se ha demostrado que las dosis altas de radiación ionizante (mucho más altas que las que se usan en la radioterapia médica) causan mutaciones celulares que pueden causar varios cánceres y deformaciones. Las valoraciones ependimales de índices viles de radiación dependen de la evaluación del riesgo, pero solo se deben realizar las pruebas radiográficas de diagnóstico necesarias y se deben evitar los estudios que utilicen rayos X en situaciones potencialmente teratogénicas como el embarazo. Generalmente se acepta que sí existe [21].

2.4.2.1.3 TOMOGRAFÍA COMPUTARIZADA (TC)

El TC, fue presentado por primera vez en la década de 1970, este producto representa un nuevo avance en el campo de las imágenes médicas. Mediante el uso de un marco o pórtico (gantry) en el que un haz de rayos X giratorio y varios detectores se colocan en disímiles configuraciones (girando continuamente alrededor del paciente), y algoritmos informáticos complejos para procesar datos, es posible formatear una gran cantidad de imágenes bidimensionales en forma de capas y en variados planos.

Una vez adquiridas, las imágenes de TC, se pueden procesar en distintas ventanas para potenciar la visualización de la imagen de varios tipos de procesos patológicos, se trata de una utilidad que mejora todo lo relacionado con las imágenes digitales en general, denominadas post-proceso o posprocesamiento. El post-proceso permite trabajar más profundamente con datos sin procesar o datos brutos (raw data) para visualizar los cambios observados con más detalle. Los pacientes pueden evitar la reexposición porque no es necesario repetir el estudio.

La obtención de las imágenes de TC requiere el uso de un escáner muy costoso, un gran espacio de instalación y una considerable capacidad de procesamiento por computadora compleja [23]. Sin embargo, la tomografía computarizada es una técnica de imagen importante para estudios de imágenes transversales y está disponible en muchos centros médicos, aunque todavía no se ha convertido en un procedimiento verdaderamente manejable. Los escáneres de TC son más costosos y más complejos que los aparatos de rayos X tradicionales y, al igual que estos últimos, utilizan radiación ionizante (rayos X) para crear imágenes.

2.4.2.1.4 ECOGRAFÍA

El ultrasonido usa energía sonora a frecuencias más altas de lo que los humanos pueden escuchar para crear imágenes, en lugar de rayos X como los tradicionales rayos X y tomografías computarizadas. Utiliza un transductor que produce ecografías y registra señales de ultrasonido. El ordenador integrado en el dispositivo procesa la propia señal

según sus particularidades. Las imágenes de ultrasonido se crean digitalmente y se logran almacenar fácilmente en el sistema PACS.

Los dispositivos de ultrasonido son relativamente económicos en comparación con los dispositivos de tomografía computarizada y resonancia magnética. Se pueden encontrar en la mayoría de los centros de salud y se pueden llevar fácilmente en la mano. Este método es especialmente útil para imágenes de mujeres, mujeres embarazadas y niños en edad fértil, ya que la radiación ionizante no se usa en la ecografía..

Este método es particularmente útil para obtener imágenes de mujeres en edad fértil, mujeres embarazadas y niños, ya que la radiación ionizante no se usa en ultrasonido. Asimismo, la ecografía es muy oportuna para visualizar tejidos blandos y distinguir entre estructuras sólidas y quísticas [21].

También se usa comúnmente para biopsias guiadas por imágenes y es un método no invasivo para estudiar el flujo sanguíneo. En general, se considera que la ecografía es un método de obtención de imágenes muy seguro y no tiene efectos negativos importantes cuando se utiliza en el diagnóstico médico [20].

2.4.2.1.5 MAMOGRAFÍA

Es una imagen de la glándula mamaria que se obtiene mediante el mamógrafo por acción de rayos X, para lograr la mayor resolución posible de las estructuras internas que es utilizada por los médicos para visualizar signos iniciales que coadyuven al diagnóstico maligno o benigno de cáncer de mama. En este sentido, afirma el médico alemán Albert Salomón, que las mamografías frecuentes son excelentes ensayos de que dispone el médico para darse cuenta de la prevalencia del cáncer de mama, así lo publicó en su artículo en el año 1913, destacando además, la importancia de las radiografías de las mamas para descubrir la extensión del tumor de los ganglios linfáticos axilares [22].

Las mamografías regulares ayudan a reducir la mortalidad por tumor de los ganglios linfáticos axilares o cáncer de mama en la medida que sea detectado a tiempo. La

frecuencia de prevención y visitas es fundamental para tratar y eliminar la enfermedad. Estos problemas pueden hacerse axiomáticos mucho antes de que se hagan evidentes.

A este respecto, las mamografías es un sistema de presión por placas que aplanan la glándula mamaria a través de una máquina especial de RX, por lo general la mujer no sentirá dolor y en caso de que aparezca va a depender de la tecnología, es decir y el exámen generalmente se realiza en una serie de pasos no superior a cuatro radiografías para captar la mayor parte del tejido y asegurar que las impresiones sean efectivas que permitan garantizar una decisión diagnóstica [25].

El estudio consiste en colocar la mama en la placa mamográfica; acto seguido, se presionará con otra placa para igualar las densidades y hacer más homogéneo el tejido para la radiación del haz de rayos X. Es el tiempo que por demás es muy mínimo donde la mujer siente que pudiera ser la más incómoda de la prueba. Según los especialistas, se debe evitar la realización de este exámen cuando la paciente esté en el período menstrual, ya que los senos son más sensibles [19].

2.4.2.1.6 BIOPSIA POR ASPIRACIÓN CON AGUJA FINA (FNA)

Una biopsia por aspiracion con aguja fina (FNA) permite extraer pequeñas muestras ya sea de liquido o masa en zonas sospechosas con atraves de una jeringa con aguja fina y hueca con la finalidad de examinar posibles celulas cancerosas [26]. Este tipo de procedimientos pueden llevarse a cabo con ayuda de pantallas de ecografias por lo que este tipo de metodos tambien es conocido como biopsia guiada por ecografia [26]. Entre las principales ventajas de este metodo o procedimientos radican en que es facil realizar, no requiere de incisiones o cortes en la piel por lo que no deja cicatriz y es muy rapido de ralizar, sin embargo, las biopsias por aspiracion con aguja fina por si solo no dan un resultado o diagnostico claro por lo que es necesario realizar diferentes tipos de biopsia [26].

2.4.2.1.7 RESONANCIA MAGNÉTICA (RM)

La resonancia magnética (MRI) no permite radiación ionizante es decir RX, por tanto, es un exámen imagenológico que se realiza con la energía potencial que se almacena en los átomos de hidrógeno del cuerpo a través de cuerpos magnéticos y ondas de alta frecuencia para producir y crear niveles de energía locales y específicos de tejido necesarios para que los programas informáticos avanzados cree imágenes en 2D o 3D respectivamente.

Los escáneres de RM no son tan comunes como las tomografías computarizadas; son costosos y deben ubicarse en áreas arquitectónicamente específicas para funcionar correctamente [27]. En general, también conllevan altos costos de mantenimiento. Sin embargo, la RM no utiliza radiación ionizante y, en comparación con la TC, ofrece un contraste mucho mayor entre los diferentes tipos de tejidos blandos [28]. La resonancia magnética se usa ampliamente en neurología y es particularmente adecuada para visualizar tejidos blandos como músculos, tendones y ligamentos..

Coexisten preocupaciones de seguridad acerca de los campos magnéticos considerablemente fuertes de los escáneres de RM, tanto en relación con objetos que pueden estar en el cuerpo, como marcapasos, como en relación con proyectiles ferromagnéticos que se encuentran cerca del escáner. Por ejemplo, las bombas metálicas de oxígeno. También se conocen los efectos secundarios de las ondas de radiofrecuencia generadas por estos escáneres, y siempre deben tenerse en cuenta los posibles efectos secundarios de algunos agentes de contraste utilizados en los estudios de RM [23]. Teniendo esto en cuenta, existen cinco densidades básicas desde la menos hasta la más densa [21], como se ilustra en la tabla 1.

Tabla 1. Las cinco densidades básicas

Densidad	Aspecto
Aire	Absorbe la menor cantidad de radiaciones y percibe con un tono más oscuro que en los RX convencionales.
Grasa	Se observa de color gris, ligeramente más oscuro que los tejidos blandos.
Líquido	La sangre y el músculo poseen igual consistencia en las radiografías convencionales.

Calcio	Se presenta y es absorbido por los huesos que demandan mayor cantidad de radiación (RX).
Metal	Totalmente absorbido por los rayos X y aparece más blanco.

Fuente: *[21], adaptado por los investigadores (2021).*

Uno de los usos principales de la tomografía computarizada TC es la capacidad de expandir la escala de grises, que puede distinguir entre estas cinco o más densidades básicas. El lector debe tener en cuenta que cuanto más denso es el objeto, más rayos X requiere y más claro se aprecia la imagen de rayos X. Cuanto menor es la densidad de los objetos, menos rayos X absorben y más oscuro los colores que aparecen en las radiografías. Desafortunadamente, el término específico que se usa para representar lo que parece blanco y lo que parece negro en una imagen depende del método de obtención de imágenes [21], lo que se explica en la tabla 2.

Tabla 2. Términos utilizados en cada modalidad (Blanco y negro)

Modalidad	Términos utilizados para blanco	
Radiografías convencionales	Opaco; significa que existe un aumento de la densidad.	Translúcido, implica una disminución de la densidad.
TC	Hiperdenso; significa que se está en presencia de una intensidad aumentada.	Hipodenso; implica la presencia de una intensidad disminuida.
RM	Cuando la intensidad se presenta brillante, indica que se está en presencia de una señal aumentada.	Cuando la intensidad se presenta oscura; indica que se está en presencia de una señal disminuida.
Ecografía	Cuando hay un efecto Ecodenso, señala que las ondas ultrasónicas están aumentadas.	Cuando hay un efecto Ecolúcido, señala que las ondas ultrasónicas están disminuidas.
Medicina nuclear	Incremento de la captación del marcador.	Disminución de la captación del marcador.
Estudios con bario (Ba)	Radiopaco.	Radiolúcido.

Fuente: *[21] adaptado por los investigadores (2021)*

2.4.2.2 MINERÍA DE DATOS

Las bases de datos en los últimos tiempos se ha considerado una herramienta fundamental para las organizaciones en general, porque permite generar y almacenar datos de forma organizada y efectiva. La clave está en descubrir patrones y algoritmos y es aquí donde entra en juego la minería de datos que se remonta a la década de los años 60 cuando los estadísticos de la época se ocupaban de términos como pesca de datos (data fishing), minería de datos o arqueología de datos (data mining o data archeology), más tarde en la década de los 80, Rakesh Agrawal, Gio Wiederhold, Robert Blum y Gregory Piatetsky-Shapiro, se dan a la tarea de fortalecer la minería de datos, y sacarle el mayor provecho a las ventajas que aporta al relacionar un conjunto de técnicas y tecnología para la exploración de información de una manera automatizada, las bondades de la Data Mining, fueron efectivas que ya para el año 2002 más de cien empresas a nivel global implementaron este sistema para encontrar soluciones a patrones repetitivos que explican el comportamiento de los datos apoyado en esta tecnología.

En un tiempo más reciente, el desarrollo de aplicaciones que utilizan la minería de datos se ha vuelto más innovador tecnológicamente y producto de ello su consolidación al compatibilizarse con una serie de técnicas y herramientas que contribuyen en la mejora y crecimiento de las organizaciones y en la toma de decisiones. Aunque la Data Mining o Minería de Datos para la organización de datos es relativamente joven, en la actualidad, se utiliza en todos los ámbitos motores de la sociedad. En el sector salud, por ejemplo, se destaca por la efectividad en la detección precoz del cáncer de mama al diagnosticar a partir de la información contenida en la base de datos de imágenes mamográficas [27].

Por otra parte, en este sistema de organización de datos, también se tiene el término Minería de Datos Inteligente (Intelligent Data Mining), que se refiere al estudio y concentración de métodos de aprendizaje automático para detallar específicamente patrones presentes en los datos [27], de esta manera se da origen a una variedad de métodos para el análisis de datos estadísticos [29], en consecuencia del aumento en el volumen de información almacenada en las bases de datos, estos métodos comenzaron a presentar dificultades en cuanto a la eficiencia y escalabilidad y es aquí donde cobra fuerza la minería de datos.

Queda claro entonces, que el análisis de datos tradicionales supone la construcción de variables validadas contra los datos, mientras que en la Data Mining, los patrones son automáticamente extraídos de los datos [27]. Esto significa, que en la actualidad la minería de datos se manipula para resolver diversos problemas como: clasificar correos electrónicos como spam, opiniones indicaciones y/o explicaciones a través de la red amigable Facebook, exploraciones en repositorios Google, entre otros. Se percibe así, como las máquinas aprenden sin una programación previa mediante el uso de redes neuronales como una imitación del cerebro humano [23].

Por otra parte, Data Mining o minería de datos, se apoya en una serie de pasos que abarcan un sistema de descubrimiento integral que potencia el conocimiento [28] los cuales se resumen a continuación:

- Comprensión del dominio de la aplicación, conocimientos relevantes que se utilizaran y objetivos del usuario.
- Selección de un conjunto de datos en el que ejecutar el proceso de nuevos hallazgos.
- Limpieza y preprocesamiento de los datos.
- Diseño de una estrategia adecuada para manejo del ruido, así como también de valores incompletos, fuera de rango e inconsistentes.
- Elección de la tarea, en lo que tiene que ver con categorización y asociación.
- Selección de los algoritmos a utilizar.
- Transformación y reduciendo las dimensiones de los datos. Aquí se hace necesario el uso de formas que requiere el algoritmo hacia la búsqueda de atributos útiles.

Todos los pasos anteriores deben ser revisados cuidadosamente, tomando en cuenta que se asigna un conjunto de datos y con base en ellos, obtener una respuesta oportuna al considerar aspectos de regresión y clasificación de la información; donde sea necesario producir una cantidad infinita de datos, tener en cuenta una salida binaria para la interpretación de datos buenos y malos que permita a su la aceptación o no de los mismos. En consecuencia, si los pasos que potencian el conocimiento no son

supervisados, el aprendizaje estaría en el limbo a tal punto de no saberse por qué característica se agruparán los datos.

En el mismo orden de ideas, hay procesos que estructuran y detallan cada una de las actividades a realizar entre las cuales podemos encontrar las siguientes metodologías de minería de datos:

En primer lugar, ***la extracción de conocimiento en bases de datos (Knowledge Discovery in Databases o KDD)***, a través de un proceso de identificación potencialmente útil y comprensible de patrones ocultos en los datos, se aprovechan procesos inherentes a varias áreas operativas como la estadística, la computación gráfica, y del conocimiento como la inteligencia artificial, usando bases de datos como materia prima [30]. En el mismo orden de ideas, la extracción de conocimientos en bases de datos, también se define como la integración de un conjunto de áreas cuya finalidad es identificar el conocimiento obtenido de bases de datos que brinden una orientación al proceso de toma de decisiones [31]. Esta metodología fue presentada por primera vez en el año 1996 por Fayyad [30] como un procesos iterativo y de intercambio de información entre sus procesos el cual consta de 5 fases como son [32]:

- Selección: es en esta donde se adquiere el conjunto de datos a estudiar y se determinan las variables de estudio.
- Pre procesamiento: en esta fase se realiza el formateo de los datos con la finalidad de obtener la mayor consistencia entre los datos.
- Transformación: es en esta fase donde se transforman los datos a un formato único con la finalidad de seleccionar y aplicar los modelos o métodos a utilizar para la extracción de información.
- Data Mining o minería de datos: dependiendo de los objetivos planteados en la investigación y de los datos obtenidos una vez se han pre procesados y transformados se utilizan una serie de técnicas con el propósito de obtener patrones ocultos entre los datos o en términos generales de descubrir nueva información.
- Interpretación y evaluación: en esta fase final se evalúan los datos para obtener los mejores resultados y por consiguiente evaluar dicha información o patrones

minados. Asimismo se interpretan los resultados y con esto comprender los resultados de la investigación convirtiéndola en conocimiento útil.

Por otro lado, la metodología ***SEMMA*** es el acrónimo en inglés de Muestra, Explorar, Modificar, Modelar y Evaluar hace referencia a los procesos tenidos en cuenta para la mejora de proyectos de minería de datos propuesta por SAS Institute Inc [33]. Esta metodología considera las 5 fases siguientes [32]:

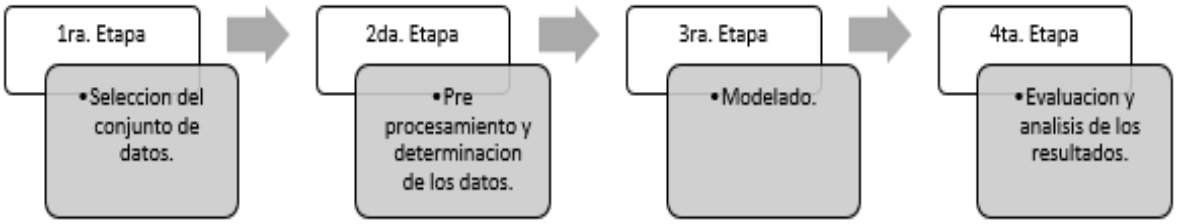
- Muestra: en esta fase se considera como muestra una parte representativa del total de datos, lo suficiente para que se considere que se puedan obtener datos significativos pero así mismo relativamente pequeño para que su manipulación, procesamiento y estudio sean los más fácil y rápidos.
- Explorar: consiste en la exploración de los datos con la finalidad de comprenderlos, obtener patrones o características iniciales como la tendencia o datos atípicos en el conjunto de datos.
- Modificar: en esta fase se modifican los datos, es decir, mediante la selección y creación de de variables estas se transforman pensando en la siguiente fase del modelado de los datos.
- Modelar: esta fase consiste en utilizar software y/o herramientas de minería de datos para explotar los datos y obtener información útil.
- Evaluar: en esta fase se evalúa qué tan confiables y útiles son los resultados obtenidos mediante los resultados del modelado de datos determinando el desempeño de los modelos creados.

Por último, ***el proceso estándar Cross-Industry para minería de datos (CRoss-Industry Standard Process for Data Mining o CRISP-DM)*** es una metodología abierta y flexible específicamente para proyectos de minería de datos [34], es decir, no es necesario una secuencia precisa ya que la gran mayoría de las 6 fases que esta metodología comprende: a) visión del negocio, b) claridad, c) preparación, modelado, evaluación y distribución de los datos se retroalimentan entre sí. Asimismo esta metodología, contempla la herramienta IBM® SPSS® Modeler útil para realizar informes de cada una de las fases y posee rutas de trabajo en donde se resumen detalladamente las tareas que encaminan hacia los resultados de cada una de las fases en particular [34].

Cabe resaltar, que esta es la metodología más utilizada en la ejecución de proyectos de minería de datos [35].

De forma general, y según la literatura revisada, la minería de datos se compone de cuatro etapas como vemos en la siguiente ilustración:

Ilustración 1. Generalidades de la Minería de Datos



***Fuente**: elaboración propia, los investigadores (2021).*

2.4.2.2.1 APLICACIÓN DE LA MINERÍA DE DATOS

Teniendo en cuenta los artículos y otras investigaciones reflejadas en este documento se puede determinar lo siguiente:

Primeramente, la minería de datos como se ha definido anteriormente busca explotar los datos para descubrir información útil, de relevancia y que sea significativa para la toma de decisiones. Ahora bien, para entender cómo se aplica la minería de datos hay que comprender primero los siguientes conceptos: tareas, modelos y técnicas. Dicho esto, se entiende por tarea como aquella actividad a resolver, es decir, el tipo de problema el cual se le requiere buscar una solución a partir de un conjunto de datos, por ejemplo: determinar a qué personas se le aprueba o no un crédito de vivienda = clasificación. En segundo lugar, se entiende por modelos al resultado de la aplicación de un algoritmo a un conjunto de datos de entrenamiento y de prueba, mediante una tarea en específico, por ejemplo: train.kknn(fórmula = salario ~ ., data = creditosVivienda, kmax = 9). Por último, se entiende como técnicas al método utilizado para la construcción del modelo, por ejemplo: kNN – k vecinos más cercano.

Teniendo claro lo anterior, existen dos tipos de tareas y modelos: predictivos y descriptivos.

Predictivos: este tipo de tareas se encargan de predecir un evento futuro de uno o más atributos. En este tipo de tareas predictivas encontramos:

- **Clasificación de datos:** este atributo es utilizado en la metodología Data Mining o CRISP-DM, para clasificar de datos de acuerdo a sus atributos y el nivel del riesgo bajo, medio y alto, lo cual va a depender de la información histórica de los datos. A este respecto, se indica que dentro de los modelos predictivos clasificar los datos es encontrar propiedades identitarias entre un conjunto de objetos. Teniendo esto en mente, construir el modelo predictivo implica utilizar un conjunto de atributos y clasificarlos de acuerdo con la instancia a la cual pertenece.

  Por consiguiente, el propósito de la clasificación a través de la metodología Data Mining o CRISP-DM, se corresponde con el análisis y posterior monitoreo de datos de entrenamiento que faciliten clasificar por sus características los datos disponibles que pudieran ser desconocidas en otras instancias. El método se denomina monitorizado, ya que se conoce la clase de pertenencia del conjunto de entrenamiento y se informa al modelo si la clasificación realizada por él es correcta o no. La construcción del modelo se retroalimenta a partir de esta información proporcionada por el supervisor [36].

  Sobre todo las notaciones principalmente manejadas para las tareas de clasificación tiene una relación directa con los algoritmos de inducción. Actualmente concurren una variedad de enfoques de inducción y algoritmos, el presente estudio destacan aquellos orientados a generar árboles de decisión considerado otro método de enseñanza y aprendizaje . monitoreado que efectivamente crea árboles de decisión a partir de un conjunto de capacitación y entrenamiento.

  Un sistema típico para construir árboles de decisión es ID3, que al utilizar la teoría de la información tiene la particularidad de minimizar el número de pruebas para clasificar un objeto. Este sistema utiliza formas o técnicas heurísticas, así, las decisiones ID3 son consideradas decisiones más o menos simples. En el mismo

sentido, se observa que una extensión de ID3 es C4.5 [37], lo que traduce en extender el dominio de clasificación de atributos categóricos a numéricos. Un paso importante en la construcción del árbol de decisiones es la poda, que elimina las ramas innecesarias, lo que da como resultado una clasificación más rápida y una mejor precisión de la clasificación de datos [38]. En consideración, hay varios algoritmos de clasificación de datos, incluidos métodos estadísticos de regresión lineal; algoritmos de aprendizaje automático, genéticos; lógica difusa y redes neuronales.

- **Árboles de decisión:** Los árboles de decisión son uno de los modelos más fáciles de entender y más utilizados en la categoría de aprendizaje supervisado [29]. Para tomar una decisión, se transita en el árbol de arriba abajo, y cada nodo se prueba repetidamente para cada atributo hasta que se llega a una conclusión que coincide con una de las hojas del árbol. Por lo tanto, el árbol se puede representar con un estilo del tipo “Si Entonces”, que son fáciles de entender e interpretar para el usuario.

  Otra ventaja de usar árboles de decisión es la resistencia a la pérdida de datos sobre ciertos atributos. Además, el algoritmo es robusto incluso en presencia de ruido en los datos [37]. Cada nodo del árbol corresponde a un atributo del repositorio. Se seleccionan de forma recursiva en función de lo que se denomina escala de impurezas de los nodos.

Descriptivos: este tipo de tareas son utilizados para obtener información entre la relación de varias variables.

- **Agrupación de datos:** este modelo de agrupación o agrupamiento tiene sentido en la medida que permite agrupar datos en conjunto similares a los valores de sus atributos. La agrupación permite identificar territorios crasamente poblados, de acuerdo con una medida de distancia establecida [36]. De esta manera, se intenta extender la similitud de las instancias en cada grupo y minimizar la similitud entre los grupos [39]. La técnica de agrupamiento se ha estudiado en los campos de la estadística, las bases de datos espaciales y la minería de datos. Dos de los

algoritmos de agrupación en clúster más utilizados son los mapas autoorganizados y redes de Kohonen.

Se trata de un modelo de red neuronal con la capacidad de crear mapas de características similares a los del cerebro, Los mapas autoorganizados se basan en un aprendizaje competitivo y no controlado, lo que significa que por lo general no se requiere de la intervención humana y se sabe muy poco sobre las propiedades de la información ingresada. A este respecto, SOM proporciona una topología de datos en diferentes dimensiones mediante unidades neuronales para sintetizar la representación [29].

Las neuronas suelen formar un mapa bidimensional de modo que el mapeo transfigura un inconveniente multidimensional en el espacio en un plano. La pertenencia de mantener la topología personifica que el mapeo mantiene distancias recíprocas entre los puntos. Los puntos que están muy juntos en el espacio de entrada original se asignan a las neuronas vecinas en mapas autoorganizados, los cuales resultan ser muy útil como herramienta para analizar clases multidimensionales de datos [27], y también tiene la capacidad de generalizar, lo que significa que la red puede reconocer datos de entrada que nunca antes había encontrado.

K-means es un método iterativo que intenta formar k clústeres, donde k está preestablecido antes de que comience el proceso. K-means comienza dividiendo los datos en k subconjuntos no vacíos, automatiza el centroide de cada partición como el centro del grupo y asigna todos los datos al grupo con el centroide más cercano, seguidamente, reparte iterativamente información hasta no haber cambios de datos en los grupos de una iteración a la otra [38].

- **Reglas de asociación:** una regla de asociación es una declaración sobre la coincidencia de ciertos eventos en una base de datos [40]. En la minería de datos, esta asociación es fundamental en localizar reglas de la forma (A1 y A2 y...y Am) => (B1 y B2 y...y Bn), donde Ai y Bj. son valores de la naturaleza del conjunto de datos. Son varios guarismos que son útiles para la detección coincidencialmente

de reglas de asociación, caracterizando el algoritmo A priori como uno de los más utilizados.

- **Reconocimiento de patrones en la minería de datos:** es la ciencia de encontrar modelos analíticos con el propósito de describir o clasificar datos y mediciones, así como inferir a partir de conjunto de datos, mecanismos de apoyo al procesos de toma de decisión [29], la estadística es vista como un enfoque, ya que facilita el reconocimiento de patrones. Las estadísticas se utilizan para derivar un modelo a partir de datos o una medición. El objetivo es seleccionar un atributo que permita clasificar el patrón en uno o más grupos o clases.

  Acerca de los antes expuestos, hay muchos casos exitosos en los que la minería de datos se ha aplicado a la medicina mediante el reconocimiento de patrones. Otras investigaciones han escaneado imágenes de lesiones cutáneas para ayudar en el diagnóstico de melanoma [41]. Asimismo, se ha analizado la marca en pacientes con pie plano para diagnosticar y monitorear la recuperación del paciente después de la cirugía [42]. Por último, otro estudio sugirió utilizar información de la base de datos de Medline para obtener referencias a genes y proteínas para respaldar los resultados coincidentes [24].

2.4.3 ASPECTO LEGAL Y NORMATIVO

La política en Colombia acerca de la prevención y riesgos del cáncer en general y el cáncer de mama en particular, posee una sólida base internacional siendo pilares fundamentales los instrumentos vinculantes de la Organización Mundial de la Salud (OMS) y la Organización Panamericana de la Salud (OPS). La base normativa nacional reposa en la Constitución Política que consagra en su articulado la seguridad social como un servicio obligatorio para toda la población bajo la potestad del Estado, asimismo, la atención de la salud y el saneamiento ambiental. En este contexto, el estado colombiano se ve obligado a garantizar como lo ordena la carta magna a garantizar a todas las personas del territorio nacional propios y extraños los servicios de protección y recuperación de la salud, conforme a los principios de eficiencia, universalidad y solidaridad y en los términos que establezcan la leyes.

Instrumentos internacionales vinculantes, Colombia ha incorporado los tratados más importantes sobre el cáncer y sus factores de riesgo en su legislación nacional. Estos tratados incluyen:

- El Convenio núm. 139 de 197 sobre el cáncer en el lugar de trabajo obliga a las partes contratantes a determinar periódicamente las sustancias y carcinógenos cuya exposición está prohibida en el lugar de trabajo o que están sujetos a autorización o control.
- Convenio núm. 161 de 1985 sobre los servicios de salud en el trabajo.
- Convenio núm. 162 de 1986 sobre el uso de Asbesto en condiciones de seguridad.
- Convenio sobre productos químicos núm. 170 de 1990.
- Convenio de Estocolmo sobre contaminantes orgánicos persistentes.
- Convenio Marco de la OMS para el Control del Tabaco.

Normatividad Nacional

- Ley 09 de 1979. Con la que se dictan medidas sanitarias.
- Ley 30 de 1998, con la cual se dictan normas ambientales prohibitivas, referidas a residuos peligrosos y se dictan otras disposiciones.
- La Ley 715 de 2001, que dicta normas orgánicas sobre recursos y competencias de conformidad con los artículos 151, 288, 356 y 357 de la constitución política y se dictan otras disposiciones para organizar la prestación de los servicios educativos y de salud.
- Ley 1122 de 2007, Reforma del sistema general de Seguridad Social en Salud.
- Ley 1335 de 2009. Disposiciones mediante las cuales se previene el daño a la salud de los menores y la población no fumadora y se estipulan políticas públicas para la prevención del consumo de tabaco y el abandono de la adicción al tabaco del Tabaquismo y sus derivados en Colombia.
- Ley 1355 de 2009, por la que se define la obesidad y las enfermedades crónicas no transmisibles asociadas a ella como prioridades de salud pública y se dictan medidas para su control, tratamiento y prevención.
- Ley 138 de 2010, que establece acciones para la atención global del cáncer en Colombia.

- Ley 1388 de 2010, para el derecho a la vida de los niños con cáncer en Colombia.
- Ley 1 38 de 2011, con la que se reforma el sistema de Seguridad Social en Salud y se dictan otras disposiciones.

2.4.4 ASPECTO ÉTICO

La necesidad de establecer nuevas formas de gestionar la educación acordes a las necesidades de la sociedad actual, combinada con condiciones específicas, ha obligado a los profesionales de la práctica educativa a combinar valores éticos y sociales para generar conocimiento útil y persistente sin que se evidencie un vacío relevante en la consideración ética. Por tanto, hacer ingeniería no es lo mismo que ser Ingeniero, no consiste solamente en hacer un buen diseño, una buena estructura, una buena programación informática y un buen rendimiento de la máquina y el sistema, más bien, se trata de comprender el papel de la ingeniería en la sociedad, en la mejora de las condiciones de vida y la calidad de vida de las personas, así como en el desarrollo sostenible, y ejercer la profesión en este contexto.

Más allá de la ingeniería, ser ingeniero tiene que ver con actuar de acuerdo con los valores superiores contenidos en la Declaración de Principios Éticos del Ingeniero: Veracidad, Integridad, Responsabilidad y Precisión, según la Asociación Colombiana de Facultades de Ingeniería (ACOFI). Estos son los valores básicos de la práctica de la ingeniería que deben incorporarse al trabajo diario de ingeniería. En este sentido, los ingenieros de cada una de las disciplinas tienen entonces el deber primordial de contribuir al análisis, la reflexión y el diálogo desde todos los ángulos con el fin de crear un clima ético, una ciudadanía que podría verse como el primer mandamiento, la acción técnica sin restricciones debe ser. tomar el control.

De ello se desprende que “saber qué”, además del “querer ético”, que se le exige al ingeniero para garantizar la moralidad de la acción, se encuentra dentro de los límites del “querer hacer” como virtud que descuida el dónde, cuándo quién y cómo hacer algo, que va más allá de una educación cívica ocasional. En el día a día del investigador se presentan la mayoría de los dilemas éticos a resolver y en este momento crucial el futuro

profesional del ingeniero debe contar con más información y más apoyo de la institución a la que pertenece o egresa.

2.4.5 LA ÉTICA EN LA INVESTIGACIÓN

La ética juega un papel esencial, donde alguno con intereses especiales abandona la ética en una investigación y corrompe la ciencia y sus productos [43], desde esta perspectiva, en el proceso educativo, las personas se construyen a sí mismas y colectivamente. Esto se debe a que allí se establecen los significados de la coexistencia, la obligación, el compromiso y el descubrir, en esa articulación de la vida social con el entorno. Están inmersos entonces los criterios, modelos, principios que formaron al hombre digno con ética y moral para defenderse en la vida siendo proactivo, si la escuela como formadora de valores no brinda estos nutrientes, la demanda del entorno terminará por quitar sus deseos y aspiraciones.

Hoy es un aspecto neurálgico que la ética de primera y segunda generación, que todavía representa un significativo papel, se anulen los espacios de reflexión acerca de la conducta ética y moral de la persona. El ideal de lo práctico cumpliría su función, pero para las generaciones de relevo sería ineludible pensar en una correspondencia dinámica con la realidad del entorno. Aquí es donde puede entrar la bioética, ya que se requería el desempeño holístico de un ingeniero.

Existen códigos deontológicos universales que en la actualidad se ocupan de lo ético y lo moral en cualquier profesión que regulan la independencia, la obligación de decir la verdad, el deber de mantener el secreto profesional. Sobre todo, un profesional llamado a servir al prójimo debe ser una persona humanizada, sensible, honesta y ante todo cumplidora de sus deberes y respetuosa de sus derechos.

También debe ser particularmente ecuánime en todas las declaraciones, especialmente aquellas que son públicas y se relacionan con aspectos técnicos de su profesión. Por tanto, el ingeniero informático debe adherirse a la normativa ética en relación con su profesión. Asimismo, debe ser consciente en los procedimientos administrativos, técnicos,

operativos en el uso de la información que trata y se confía, esto es confidencialidad, que es un derecho y una obligación profesional cardinal..

2.5 MARCO METODOLOGICO

2.5.1 DISEÑO METODOLÓGICO

La investigación científica se entiende como un proceso de ida y vuelta, es decir, el camino para indagar la realidad social apoyado en fases previamente interconectadas que atienden a una secuencia determinada de pasos según el tipo de estudio que se desea desarrollar. Al inicio del proceso investigativo es primordial definir los motivos que originaron la intención investigativa, luego describir la problemática con sentido lógico que despierta el interés del investigador [44]; En este estudio se consideró de gran interés contribuir a la seguridad de la salud humana; dado que el “cáncer de mama” es una de las principales causa de muerte en mujeres en Colombia, por tanto, coadyuvar con un diagnóstico oportuno, reduce de forma significativa el número de casos que conducen a la muerte de las pacientes. De esta manera se estaría dando forma al propósito de la presente investigación.

De acuerdo a lo anterior, el paradigma donde se inserta esta investigación es el positivista, aplicando un modelo de positivismo, que considera el descubrimiento de hechos de fenómenos sociales sin priorizar el estado subjetivo del individuo. En este sentido, el enfoque que tendrá la investigación es cuantitativo por ser un proceso sistemático y ordenado que permite proyectar el estudio de acuerdo a una estructura lógica de decisiones que orienta el cómo obtener respuestas adecuadas al problema de indagación propuesto.

Teniendo en mente lo anterior y para efectos de esta investigación, es posible identificar una serie de elementos comunes, lógicamente estructurados ya que se trabajará con datos o grupos de datos que requieren de una mirada crítica, sobre aspectos relacionados con la exactitud, precisión y representatividad. Esto es, la crítica de datos estadísticos para luego, expresar numéricamente el resultado de la medición de la variable en estudio, lo cual es posible a través de programas computarizados, se utilizará entonces el método

de minería de datos (Data Mining), el método heurístico CRISP-DM, lo que se persigue por una parte, evaluar la magnitud y la confiabilidad del fenómeno estudiado con mayo facilidad, asimismo, identificar patrones de interés del estudio que permitirán obtener información significativa para detectar y/o diagnosticar el “cáncer de mama”

2.5.2 CLASIFICACIÓN DE LA INVESTIGACIÓN

Esta clasificación está basada según los fundamentos metodológicos que se presentan a continuación:

2.5.2.1 SEGÚN EL PROPÓSITO O RAZÓN

Se desarrollará una investigación aplicada destinada a la búsqueda de conocimiento y por tanto, se pretende resolver un problema práctico para su aplicación inmediata.

2.5.2.2 SEGÚN EL NIVEL DE CONOCIMIENTOS

En la fase la investigación será descriptiva en la que se buscara una explicación al cómo de los hechos y explicativa en la que se busca una explicación al porqué de los hechos, se describen los hechos como son observados, las características fundamentales de un conjunto homogéneo de fenómenos empleando criterios sistemáticos pretendiendo explicar cómo ocurre y se presenta el fenómeno y en qué condiciones, lo fundamental es presentar una interpretación correcta de esa realidad.

2.5.2.3 SEGÚN LA ESTRATEGIA EMPLEADA

En la primera fase será documental, se apoya en la recopilación de antecedentes o estudios previos acerca de la misma temática a través de documentos gráficos formales e informales, los materiales de consulta suelen ser las fuentes bibliográficas, iconográficas, fonográficas y algunos medios magnéticos.

Posteriormente, en una segunda fase se hará una investigación de campo, referida a la fuente que origina la información o donde se recogerán los datos, es decir en el lugar

área, espacio, ambiente, institución, comunidad, entre otras, para obtener información con el objeto de comprender la hipótesis o descubrir relaciones desconocidas entre los hechos examinados, recogiendo los datos en forma directa de la realidad apoyado en la metodología CRISP-DM.

2.5.4 TÉCNICAS DE INVESTIGACIÓN

Para nutrir la investigación se utilizará:

- ***La investigación documental***, se realiza a través de la consulta de documentos: libros, revistas, periódicos, memorias, anuarios, registros, códigos, constituciones, entre otros. Partiendo de esta investigación se hace necesario ahondar en los enlaces de la web y repositorios digitales relacionados al tema.

- ***Investigación de campo,*** para obtener datos reales en el momento en que ocurren los investigadores de este estudio deben dirigirse al lugar donde ocurre el fenómeno o donde está ubicado en hecho y objeto para luego procesarlos. Se trata entonces de indagar el fenómeno social con el objeto de observar y comprender mejor la hipótesis de estudio. En suma, manejo y selección de patrones inherentes a la minería de datos.

2.5.3 POBLACIÓN Y MUESTRA

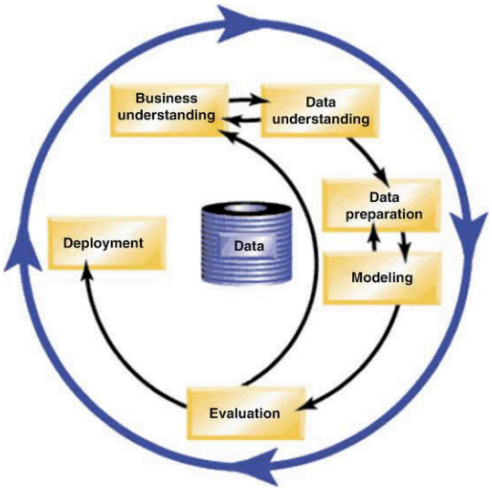
Población: la población la cual será objeto de estudio en esta investigación bien puede considerarse infinita, debido a que no es posible determinar un listado de registros en específicos debido al tipo de variables que conforman el conjunto de datos con el que se trabajara. Por lo tanto, quedan excluidos otros tipos de cáncer u otras enfermedades, sin embargo, estas pudieran ser consideradas para trabajos futuros.

Muestra: la muestra será tomada con la técnica de muestreo probabilístico ya que corresponde a un muestreo de tipo aleatorio simple, es decir, al azar los cuales permitirán extraer características e información para el cumplimiento de los objetivos propuestos.

2.5.5 UNA VISIÓN GENERAL DEL MÉTODO PROPUESTO CRISP-DM

El método o metodología ***CRISP-DM***, tiene tiene como característica fundamental la arquitectura tecnológica del método como se puede apreciar en la ilustración 2, expuesta continuación:

Ilustración 2. Modelo CRISP-DM



Fuente: *Manual CRISP-DM de IBM SPSS Modeler (2012).*

Como se puede observar en la ilustración anterior, el modelo CRISP-DM en su arquitectura tecnológica resalta seis fases. Los detalles del método son los siguientes [36]:

- ***La primera fase,*** denominada ***comprensión del negocio,*** el cual comprende las siguientes actividades:
    - Determinar metas y objetivos del negocio.
        - Objetivos del negocio.
        - Criterios de éxito del negocio.
    - Valoración de la situación.
        - Inventario de recursos.
        - Requisitos, supuestos y restricciones.
        - Gestión del riego.
        - Beneficios.
    - Determinar objetivos y metas de la minería de datos.
        - Objetivos de la minería de datos.

        - Criterios de éxito.
    - Realizar plan del proyecto.
        - Plan del proyecto.

Cabe resaltar que esta fase puede considerarse como la más importante ya que es en esta en el que se alinean los propósitos de la investigación, es decir, se determinan el porqué hacer el proyecto desde la perspectiva empresarial para luego determinar el que hacer de la investigación conforme a los propósitos de la minería de datos con respecto a la variable de estudio determinando así los criterios de la investigación, la definición de la problemática y el diseño del plan de trabajo.

- ***La segunda fase,*** corresponde a la ***comprensión de los datos***, el cual comprende las siguientes actividades:
    - Recolectar los datos iniciales.
        - Reporte de recolección de datos.
    - Descripción de los datos.
        - Reporte de la descripción de los datos.
    - Exploración de los datos.
        - Reporte de la exploración de los datos.
    - Verificación de la calidad de los datos.
        - Reporte de la calidad de los datos.

Es en esta fase en donde se determinan las primeras hipótesis de la investigación y se obtiene el primer vistazo de los datos y su estructura.

- ***La tercera fase,*** lleva por nombre ***preparación de los datos***, el cual se describen las características de las variables y los datos como también sus estructuras. Esta fase comprende con las siguientes actividades:
    - Relacionar los datos.
        - Inclusión/exclusión de datos.
    - Estructurar los datos.
    - Integrar los datos.
        - Reporte de la exploración de los datos.
    - Formateo de los datos.

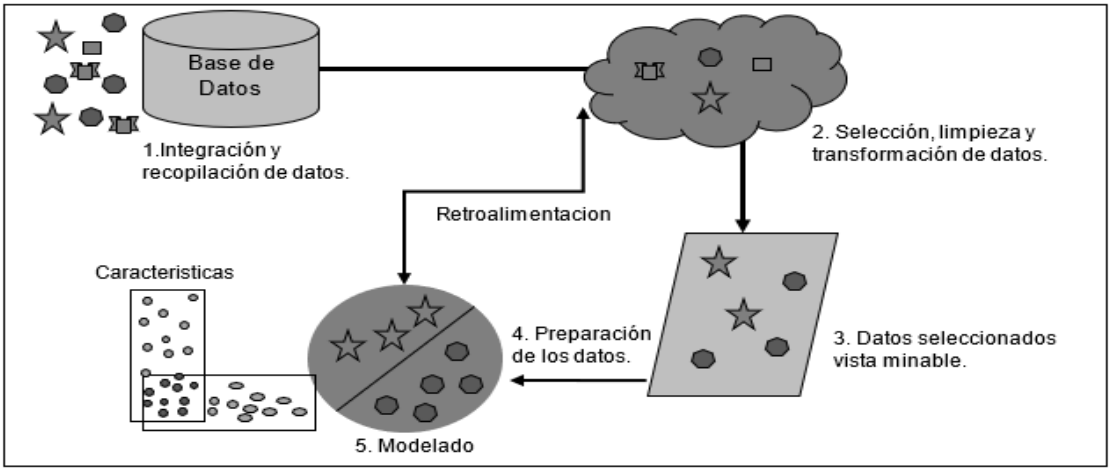
- Reporte de la calidad de los datos.

Cabe mencionar, que es en esta fase en la que se establece el conjunto de datos final para su posterior procesamiento.

- ***La cuarta fase,*** lleva por nombre ***modelado***, Esta fase comprende con las siguientes actividades:
    - Seleccionar técnica de modelado.
        - Técnicas seleccionadas.
        - Supuestos.
    - Establecer plan de pruebas.
        - Plan de pruebas.
    - Construir modelos.
        - Parámetros.
        - Construcción del modelo.
        - Descripción del modelo.
    - Evaluar el modelo.
        - Reporte de evaluación del modelo.
        - Revisión de parámetros.

Esta fase se retroalimenta con la fase 3 preparación de los datos ya que del tratamiento de los datos, es decir, desde la limpieza y transformación de los datos determinara la construcción de los modelos y la calidad de sus resultados, este proceso se puede ver ejemplificado en la siguiente ilustración:

Ilustración 3. Retroalimentación entre fases



Fuente: *elaboración propia, los investigadores (2021).*

- ***La quinta fase,*** la fase de la ***evaluación,*** contempla las siguientes actividades:
  - Evaluar los resultados.
    - Valoración de los resultados.
    - Modelos aprobados.
  - Revisión de los procesos.
  - Determinar los próximos pasos.
    - Listado de las acciones a tomar.

En esta etapa, no solo evaluamos los modelos construidos sino también, todos los procesos hasta este punto para asegurar el logro de los objetivos propuestos.

- ***La sexta fase,*** llamada ***implementación***, comprende las siguientes actividades:
  - Plan de implementación.
  - Plan de monitoreo.
  - Informe final
    - Resultados finales.
    - Modelos aprobados.
  - Revisión del proyecto.
    - Documentación de experiencias.

Por último, esta fase también es conocida como despliegue y es esta en la que toda la información descubierta y el pasos pasos de las actividades se organiza, se presenta y se discuente con el cliente final para posteriormente implementarlo o utilizarlo.

2.6 RESULTADOS ESPERADOS

Uno de los campos del saber y del conocimiento que se beneficia de la ciencia, es la medicina por su constante interacción entre la tecnología y las matemáticas, ya que de esta forma se pueden potenciar las dificultades y complicaciones más complejas con resultados y argumentos profundos y duraderos a favor de la humanidad; el aprendizaje automático y la minería de datos en el aquí y en el ahora representan para las áreas del conocimiento una realidad direccionada a la solución de problemas de salud, al brindar métodos y procedimientos innovadores que permiten un diagnóstico preciso de las enfermedades y/o padecimientos del ser humano.
Teniendo en cuenta esta premisa, se espera con esta investigación que los resultados estén direccionados a:

- Diagnosticar de manera oportuna, mediante la construcción y aplicación de modelos a través del uso de métodos como la minería de datos, el tipo de cáncer de mama (Benigno o Maligno).

- Prevenir el diagnóstico erróneo por factores humanos, ya que la detección de los carcinomas visibles mediante análisis retrospectivos en pacientes a menudo es complicada; las lesiones varían desde cambios en los tejidos blandos, hasta calcificaciones de diferentes formas y morfología, significando una evidente presencia de malignidad; por lo tanto, se requieren variadas lecturas del mismo examen.

- Mediante el uso de la técnica de la minería de datos obtener resultados confiables y fiables para así mejorar los tiempos de respuestas en los diagnósticos de los pacientes de La Liga Contra el Cáncer – Seccional Cesar.

- Implementar el aplicativo web con todos los módulos necesarios para que este pueda ser de apoyo en el diagnóstico oportuno y de manera sistemática en la Liga contra el cáncer – seccional cesar.

- Incentivar a la comunidad académica a realizar investigaciones relacionadas o similares a esta temática con el fin de que se puedan aplicar las nuevas tecnologías emergentes a problemas comunes en cuanto a la salud y a la sociedad en general.

2.7 CRONOGRAMA DEL PROYECTO

Ilustración 4. Cronograma

ACTIVIDADES	SEMANAS																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1.. Entendimiento del negocio																		
1.1. Evaluar la situación																		
1.2. Determinar los objetivos del negocio																		
1.3. Determinar los objetivos de la mineria de datos																		
1.4. Reunión con el Ingeniero de sistemas																		
2. Compresion de los datos																		
2.1. Describir los datos iniciales																		
2.2. Explorar los datos																		
2.3. Verificar datos																		
2.4. Reunión con el Ingeniero de sistemas																		
3. Preparacion de los datos																		
3.1. Seleccionar los datos																		
3.2. Liempieza de los datos																		
3.3. Construcción de los datos																		
3.4. Integrar los datos																		
3.5. Formatear los datos																		
4. Modelado																		
4.1. Seleccionar las ténicas de modelado																		
4.2. Diseñar pruebas de los modelos																		
4.3. Construir los modelos (descriptivo y predictivo)																		
4.4. Evaluar los modelos																		
4.5. Reunión con el Ingeniero de sistemas																		
5. Evaluación																		
5.1. Evaluar los resultados																		
5.2. Revisión del proceso																		

Fuente: *elaboración propia, los investigadores (2021).*

SECCIÓN III: DESARROLLO CIENTÍFICO TECNOLÓGICO

3.1. DESARROLLO DE LAS FASES DE LA METODOLOGÍA PROPUESTA

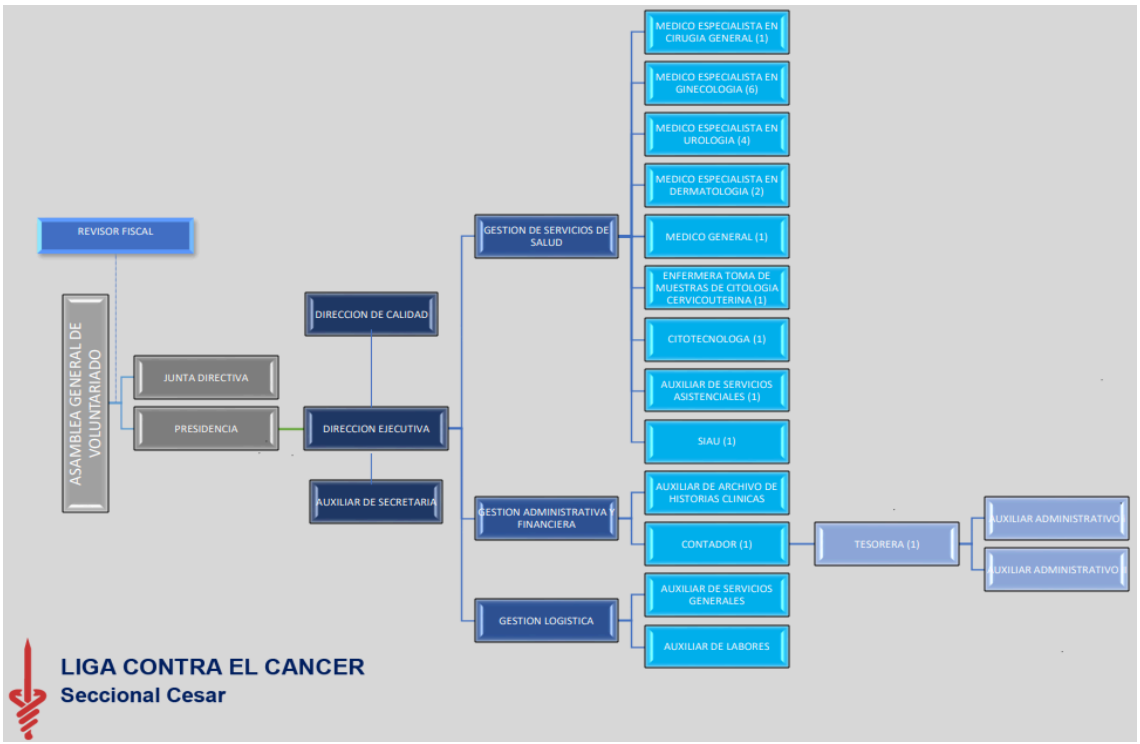
3.1.1. COMPRENSIÓN DEL NEGOCIO

3.1.1.1. Compilación de la información de la empresa

La atención integral a una persona que padezca de cáncer busca garantizar a la población afectada un diagnóstico oportuno que ayude a prevenir y detectar esta enfermedad en su etapa más temprana y asimismo puedan tener un tratamiento adecuado. Es por eso, que la Liga Contra el Cáncer – Seccional Cesar fundada el 3 de diciembre de 1979 y afiliada a la Liga Colombiana Contra el Cáncer, como IPS sin ánimo de lucro, tiene como principal propósito educar y concientizar a las personas sobre los distintos tipos de cáncer, sus afectaciones y el poder prevenirlos [45]. De igual manera, esta IPS tiene influencia en todo el departamento del Cesar y sus regiones vecinas.

Por consiguiente, la organización de esta IPS está estructurada de la siguiente manera:

Ilustración 5. Organigrama de la Liga Contra el Cancer - Seccional Cesar



Fuente: *aportada por el Ingeniero de Sistemas de la Liga Contra el Cáncer – Seccional Cesar.*

Área problemática: Ahora bien, en el departamento del cesar el cáncer está entre las enfermedades con mayor índice de mortalidad de la región, precisamente, es la tercera enfermedad con mayor porcentaje de muertes [46]. De igual importancia, durante lo corrido del año 2020, 96 personas padecieron de cáncer de mama, los cuales la mayor incidencia de casos fue en la ciudad de Valledupar [46], como podemos ver en la siguiente tabla:

Tabla 3. Casos de Cáncer de mama en el Cesar

No. diagnósticos	Ciudad
70	Valledupar
5	Agustín Codazzi
4	Aguachica
2	Chiriguaná
2	Gamarra
2	Rio de Oro
2	La Paz
2	San Alberto
1	Bosconia
1	Curumaní
1	La Gloria

1	La Jagüa de Ibirico
1	Manaure
1	Pailitas
1	San Diego

Fuente: *N. Baute Barrios, El Pilón (2021).*

Por otra parte, la actual pandemia por COVID -19 ha impactado de manera negativa a las diversas entidades y sus programas de detección temprana del cáncer así como la continuidad en los tratamientos especializados [47]. Por lo tanto, en la Liga Contra el Cáncer – Seccional Cesar la situación no es diferente y de hecho, está IPS ha recurrido a cerrar sus puertas de manera temporal con respecto a la atención de sus usuarios desde finales del mes de marzo del 2020, actuando con respecto al Decreto 253 del 23 de marzo del 2020 de la alcaldía de Valledupar, hasta principios del mes de junio del mismo año, pero sin embargo, debido al temor generalizado por los contagios entre la población y el cumplimientos de diversas medidas y protocolos de bioseguridad los tiempos de consultas para cada persona fueron condicionados en un periodo de tiempo de entre 30 a 20 minutos por seguridad tanto de los profesionales de la salud como de los usuarios en dichas consultas citológicas [48].

Solución actual: En este orden de ideas, para afrontar esta problemática se han implementado mecanismos como lo son las tele consultas y visitas de atención domiciliarias y la Liga Contra Colombiana contra el Cáncer en coordinación con su afiliada en el departamento del cesar, han creado estrategias como la creación de eventos educativos a través de medios virtuales, capacitación a sus profesionales de la salud, elaboración de campañas publicitarias como la que lleva por consigna “Ve a tiempo, que el COVID no te detenga”, se crearon 12 contenidos que tuvieron la finalidad de incentivar la detección temprana del cáncer de mama, videos con testimonios reales de pacientes, concursos de conocimientos y el evento virtual “Gran Encuentro por la Vida” dirigido a todo público en general [49].

En cuanto al apoyo psicosocial, con el objetivo de mejorar la calidad de vida y fortalecer el bienestar psicológico de los pacientes afectados por esta enfermedad, se llevaron a cabo en el año 2020 donaciones de prótesis mamarias y brasieres en todas las seccionales o

sedes de la Liga Colombiana Contra el Cáncer [49], como podemos ver en la siguiente tabla:

Tabla 4. Brasieres y Prótesis donados por la Liga

TOTAL PRÓTESIS Y BRASIERES DONADOS EN EL 2020		
Liga	**Prótesis**	**Brasier**
Liga Colombiana Contra el Cáncer	26	26
Liga Contra el Cáncer Seccional Cúcuta	18	18
Liga Contra el Cáncer Seccional Riohacha	6	6
Liga Contra el Cáncer Seccional Sucre - Sincelejo	3	3
Liga Contra el Cáncer Seccional Cesar - Valledupar	16	16
Liga Contra el Cáncer Seccional Valle Unicancer - Cali	12	12
Capítulo de Maicao	2	2
Liga Contra el Cáncer Seccional Nariño	3	3
Liga Contra el Cáncer Seccional Huila	3	3
TOTAL	89	89

Fuente: *Liga Colombiana contra el Cáncer, Informe de Gestión General (2020).*

En síntesis, son varios los mecanismos que está IPS con su entidad madre han implementado para la mejora continua de la población afectada por el cáncer de mama y otros tipos de cáncer, primando siempre la salud de sus usuarios a través de normas de bioseguridad generando un impacto positivo en lo psicológico, emocional y físico de los pacientes.

3.1.1.2. Definición de los objetivos comerciales

Esta investigación se realiza para promover el cuidado en la salud y detectar de manera temprana el cáncer de mama a través de la aplicación de técnicas de minería de datos. Por esto, se requiere cumplir con los siguientes objetivos:

- Diagnosticar los tipos de cáncer malignos y benignos.
- Desarrollar e implementar un aplicativo web que integre los resultados de la investigación.

Criterios de éxito: *se considera que estos objetivos han tenido éxito desde una perspectiva de negocio si se logra integrar los modelos a desarrollar en un aplicativo web que permita a los profesionales de la salud en la Liga Contra el Cáncer Seccional Cesar diagnosticar el cáncer de mama.*

3.1.1.3. Valoración de la situación

Para el cumplimiento de estos objetivos es primordial entender lo mejor posible las problemáticas antes planteadas y valorar los recursos que permitirán solventar dicha situación. Por otro lado, a pesar de todo el esfuerzo de los profesionales de la salud con los que la IPS cuenta, no está de más aportar de conocimiento y tecnología. Por lo tanto, se tiene en cuenta los siguientes factores:

Datos: los datos con los que se construirán los modelos que se aplicaran, serán obtenidos del repositorio de datos libre UCI y a su vez los datos que se utilizarán para probar los modelos una vez se hayan creados serán proporcionados por la Liga Contra el Cáncer – Seccional Cesar.

Riesgo**:** a pesar de ser una investigación sin ánimos de lucro y con la finalidad de apoyar en la detección temprana del cáncer de mama, está no está exenta de riesgos, los cuales pueden ser:

- Alta complejidad en el desarrollo del proyecto.
- Mala planificación.
- Expectativas poco realistas.
- Mal manejo del tiempo.
- Datos insuficientes o pocos relevantes.
- Falta de recursos necesarios.

Todo lo antes mencionado puede hacer peligrar el desarrollo de la investigación y con ello la no consecución de los objetivos propuestos.

3.1.1.4. Inventario de recursos

Para llevar a cabo cada una de las fases, actividades o tareas de la metodología se disponen de las siguientes herramientas:

Hardware: para llevar a cabo el almacenamiento y procesamiento de los datos, se dispone del equipo que contiene las siguientes características:

- Procesador: Ryzen 5 2500U
- Memoria Ram: 4GB DDR4
- Disco duro: 1000 GB Sata III de 5400 RPM
- Tarjeta gráfica: AMD Radeon 535 with 2GB Dedicated VRAM

Software: se utilizarán los lenguajes de programación R y Python, el entorno de desarrollo RStudio y Jupyter, la herramienta Power BI y por último la hoja de cálculo Excel en Windows 10. Por otro lado, todos estos softwares que se utilizaran para la aplicación de las técnicas de minería de datos son gratuitos (a excepción de la licencia de Excel y Windows 10) y que no necesitan de requisitos de hardware tan altos, por lo que con la descripción del hardware mencionado en el ítem anterior se considera que es suficiente para el desarrollo de la investigación.

Datos: En cuanto al conjunto de datos el entrenamiento de los modelos será utilizado el conjunto de datos denominado “Breast Cancer Wisconsin (Diagnostic) Data Set” obtenido del repositorio de datos libres UCI.

3.1.1.5. Requisitos, supuestos, límites y exclusiones

Requisitos: para la Liga Contra el Cáncer – Seccional Cesar quien tiene la necesidad de realizar diagnósticos de manera temprana del cáncer de mama, el producto de esta investigación será la creación de modelos predictivos y un aplicativo que web que los integre, para así apoyar la gestión de los profesionales de la salud en la IPS y que es un producto útil y totalmente escalable de tal manera que se pueden analizar otros tipos de cáncer u otras enfermedades. Entonces, por todo esto se requiere:

- Aplicar correctamente la metodología de minería de datos CRISD – DM para la extracción de información que permita el cumplimiento de los objetivos propuestos.
- Crear material informativo para promover el autocuidado.
- El aplicativo web debe ser amigable y escalable.
- El sistema debe ser multiplataforma.
- Documentar apropiadamente cada una de las tareas que se aplicaran a lo largo de la realización de la investigación..

Supuestos:

- Disponibilidad inmediata del personal que conforma el equipo de trabajo.
- Los entregables del proyecto deben ser aprobados por nuestro asesor y profesor Álvaro Oñate Bowen.
- Se asume que se van a analizar datos sobre el cáncer de mama.
- Se desarrollarán e implementarán modelos que permitan diagnosticar el tipo de cáncer de mama maligno/benigno.
- Se creará un aplicativo web que contenga los resultados de la investigación.
- Se contará con un documento final que refleja los resultados de la investigación.

Límites y exclusiones: para el desarrollo de la investigación se tendrá en cuenta lo siguiente:

- No se analizarán datos que no hagan parte de los conjuntos de datos seleccionados para el entrenamiento y la prueba de los modelos.
- Los datos con los que se cuentan fueron obtenidos mediante la técnica denominada biopsia por aspiración con aguja fina, por lo tanto, los datos que serán utilizados para realizar nuevos diagnósticos deberán ser extraídos con esta misma técnica.
- Solo se analizarán datos concernientes únicamente al cáncer de mama.

3.1.1.6. Riesgos y contingencia

Es importante tener en cuenta que puede afectar a la investigación para así poder proveer una situación que no convenga y que pueda afectar al desarrollo de esta, por lo que es importante definirlas antes de dar inicio. Para esto, se requiere responder las siguientes preguntas:

- **¿Qué sucede si el proyecto dura más de lo programado?**

Si la investigación dura más de lo programado se perderá interés por parte de la Liga Contra el Cáncer – Seccional Cesar en cuanto al desarrollo de la investigación generando así misma desconfianza en la capacidad de los responsables.

- **¿Qué sucede si en la realización del proyecto se detectan problemas presupuestarios?**

Si bien esta investigación no tiene un enfoque desde una visión de negocio con fines económicos y que gran parte del hardware ya se posee y en cuanto los softwares de desarrollo que se utilizarán son completamente gratuitos, es importante contar con un presupuesto moderado y, sí es el caso, un presupuesto para poder transportarse hacia la Liga Contra el Cáncer – Seccional Cesar con el propósito de realizar visitas ya sea para enseñar avances o recopilar datos. Por consiguiente, si no se tiene en cuenta lo anterior puede dificultar la comunicación entre los responsables del proyecto y sus destinatarios.

- **¿Qué sucede si los datos son de escasa calidad o cobertura?**

Para este caso, si los datos no son suficientes simplemente no será posible el logro de los objetivos propuestos.

- **¿Qué sucede si los resultados son menos dramáticos que los esperados?**

Si los resultados no son los esperados, la investigación no tendrá el impacto que se desea tener y por lo tanto no tendrá ninguna relevancia en cuanto a la gestión llevada a cabo por los profesionales de la salud en la Liga Contra el Cáncer – Seccional Cesar.

3.1.1.7. Análisis de costes/beneficios

Seran beneficiarios directos del proyecto, todas y cada una de las mujeres y hombres indistintamente de su nivel socioeconómico, creencias, culturas y costumbres, ya que este tipo de proyecto ayuda detectar, educar, concientizar y tratar el cáncer de mama, para poder así mejorar la calidad de vida de la población afectada y asimismo reducir la mortalidad por cáncer. Igualmente, este proyecto es de gran ayuda para aquellas personas profesionales de la salud, entidades u organizaciones, encargadas de establecer políticas de salud pública en cuanto a la prevención y el diagnóstico oportuno, como también es de interés para aquellas instituciones académicas en las que se puede incentivar el desarrollo de proyectos que traten esta misma problemática o similares, para el beneficio de la población en general.

Dicho esto, para el desarrollo de la investigación, se requieren herramientas informáticas, materiales y otros servicios. Por lo tanto, a continuación se expondrá un análisis de costos detallados de la siguiente manera:

Tabla 5. Análisis de Costos

Descripción	Unidad Cantidad	Costo	
		Unitario	Total
Equipos			
- Computador portátil	2	$1.699.999	$3.399.998
- Impresora Epson Multifunción	1	$550.000	$550.000
Total, equipos			**$3.949.998**
Materiales			
- Caja de resma carta x10 U	1	$149.900	$149.900
- Caja de lápiz x12 U	1	$6.600	$219.999
- Caja de bolígrafos x12 U	1	$9.100	$9.100
- Carpeta archivadora	1	$26.000	$26.000
Total, materiales			**$404.999**
Servicios públicos			
- Internet Claro 75 Megas x1año	1	$77.900	$77.900
- Luz	12 meses	$90.000	$1.080.000
- Agua	12 meses	$2.000	$24.000
Total, servicios públicos			**$1.181.900**
Software			
- Microsoft 365 x1año	1	$219.000	$219.000
- Licencia de Windows 10 x1año	1	$98.000	$98.000
- Vps bluehosting	12 meses	$28.000	$336.000
- Herramientas de procesamiento de datos (Rstudio, Phyton, otros)	1	$0	$0
- Herramientas de desarrollo (php, angular, otros)	1	$0	$0
Total, Software			**$677.000**
Recursos humanos			
- Autores	2	$0	$0
- Director del proyecto	1	$0	$0
Total, recursos humanos			$0
Otros			
- Transporte/pasajes			$200.000
Total, Otros			$200.000
COSTO TOTAL DEL PROYECTO			**$6.413.897**

Fuente: *elaboración propia, los investigadores (2021).*

3.1.1.8. Determinación de los objetivos de la minería de datos

Esta investigación contempla los siguientes objetivos desde la perspectiva de la minera de datos:

- Analizar los datos mediante el análisis exploratorio para comprender los datos.

- Construir modelos que permitan caracterizar y diagnosticar el tipo de cáncer de mama maligno/benigno.
- Evaluar los modelos creados para obtener los mejores resultados.

Criterios de éxito: *se considera que estos objetivos han tenido éxito desde una perspectiva de la minería de datos al crear modelos que superen un 90% de porcentaje de exactitud y fiabilidad que permitan diagnosticar el tipo de cáncer de mama.*

3.1.1.9. Plan de proyecto

El plan de la investigación con respecto a las fases de la metodología CRISP-DM es el siguiente:

Tabla 6. Plan de proyecto de minería de datos

Fase	Tiempo	Responsables	Riesgos
Compresión del negocio	1 semana	Andres Gonzalez Juan Almenares	Mala definición de la situación
Comprensión de los datos	1 semana	Andres Gonzalez Juan Almenares	Poco entendimiento de los datos
Preparación de los datos	2 semanas	Andres Gonzalez Juan Almenares	Eliminación de datos relevantes
Modelado	4 semanas	Andres Gonzalez Juan Almenares	Poco conocimiento y mala aplicación de los parámetros de los modelos.
Evaluación	2 semanas	Andres Gonzalez Juan Almenares	Mala interpretación de los resultados.
Implementación y puesta en marcha	4 semanas	Andres Gonzalez Juan Almenares	Resultados poco relevantes.

Fuente: *elaboración propia, los investigadores (2021).*

3.1.2. COMPRENSIÓN DE LOS DATOS

3.1.2.2. Recopilación de los datos iniciales

El conjunto de datos denominado *Breast Cancer Wisconsin (Diagnostic) Data Set* que se utilizara para la realización del proyecto se almacena en el tipo de archivo *.csv* tiene un peso total de 128 KB (131.072 bytes), contiene 32 atributos y 569 instancias o registros. Este conjunto de datos fue obtenido del repositorio libre llamado Centro de Aprendizaje Automático y Sistemas Inteligentes (conocido por sus siglas en inglés como UCI - Center for Machine Learning and Intelligent Systems) y fue creado en 1995 por el *Dr. William H. Wolberg del Departamento de Cirugía General de la Universidad de Wisconsin - Centro de Ciencias Clínicas, W. Nick Street del Departamento de Ciencias de la Computación de la Universidad de Wisconsin y Olvi L. Mangasarian del Departamento de Ciencias de la Computación de la Universidad de Wisconsin* en base a características calculadas de imágenes digitalizadas, el cual, dichas características representan las propiedades de células en masas mamarias en imágenes como las siguientes [50]:

Ilustración 6. Células Malignas

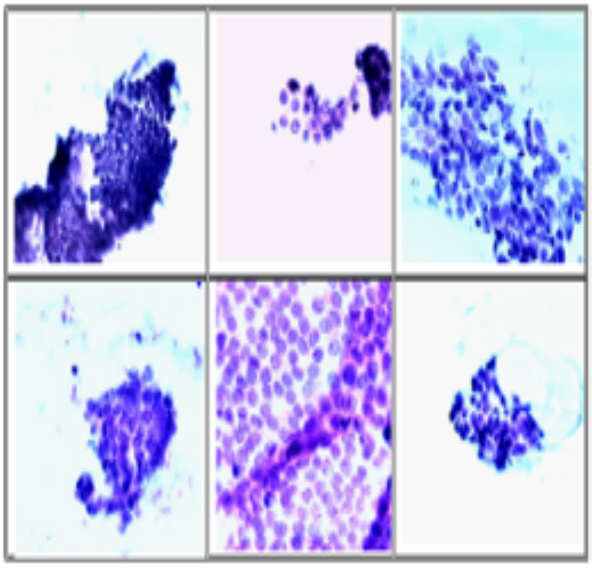
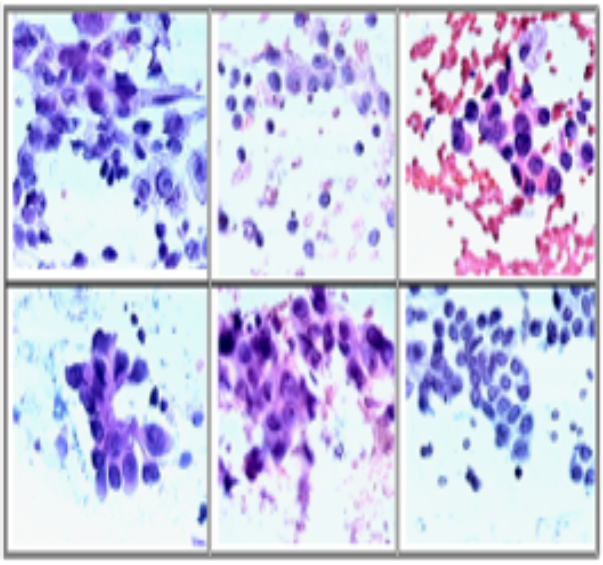


Ilustración 7. Células Benignas



Fuente: *O. L. Mangasarian y W. H. Wolberg (1990).*

Por consiguiente, estos cálculos fueron obtenidos a través del método del árbol multisuperficie (conocido por sus siglas en inglés, MSM-T - Multi Surface Method-Tree), método que mediante programación lineal crea árboles de decisión. Por otro lado, este conjunto de datos contiene los siguientes atributos [50]:

Tabla 7. Descripción de Atributos

No	Atributos	Descripción
1.	Id	Identificador del registro
2.	Diagnosis	Diagnóstico del cáncer (M=maligno / B=benigno)
3.	Radius_mean	Media de la distancia de las celulas desde el centro hasta el borde.
4.	Texture_mean	Media de la intensidad de las escalas grises de la celulas, es decir, textura de las celulas..
5.	Parimeter__mean	Media del perímetro de las celulas, es decir, longitud del contorno de las celulas.
6.	Area_mean	Media del área de las celulas.
7.	Smoothness_mean	Media de la distancia de las celulas desde el centro hasta cada punto del borde, es decir, la suavidad de las celulas.
8.	Compactness_mean	Media de la compacidad o densidad de las celulas (se calcula: perímetro ^2 / área - 1.0)
9.	Concavity_mean	Media de la concavidad de las celulas, es decir, las endiduras o curvas de las celulas.
10.	Concave.points_mean	Media de concavidad de las porciones del contorno de las celulas, es decir, curvaturas hacia dentro de las celulas.
11.	Symmetry_mean	Media de la forma simétrica de las celulas, es decir, proporciones ordenadas de las celulas.
12.	Fractal_dimension_mean	Media fractal del perímetro de las celulas (se calcula: "aproximación de la línea de costa o contorno" – 1).
13.	Radius_se	Error estándar de la distancia de las celulas desde el centro hasta el borde.
14.	Texture_se	Error estándar de la intensidad de las escalas grises de las celulas.
15.	Perimeter_se	Error estándar del perímetro de las celulas.
16.	Area_se	Error estándar del área de las celulas.
17.	Smoothness_se	Error estándar de la distancia de las celulas desde el centro hasta cada punto del borde.
18.	Compactnees_se	Error estándar de la compacidad o densidad de la celulas.
19.	Concavity_se	Error estándar de la concavidad de la celulas.
20.	Concave.points_Se	Error estándar de concavidad de las porciones del contorno de las celulas.
21.	Symmetry_se	Error estandar de la forma simétrica de las celulas.
22.	Fractal_dimension_se	Error estandar fractal del perímetro de las celulas.
23.	Radius_worst	Peor o el mayor valor medio para la media de radio de las celulas.
24.	Texture_worst	Peor o el mayor valor medio de la textura de las celulas.
25.	Perimeter_worst	Peor o el mayor tamaño medio del perímetro de las celulas.

26.	Area_worst	Peor o la mayor área de las celulas.
27.	Smoothness_worst	Peor o la mayor variación media de la suavidad de las celulas.
28.	Compactness_worst	Peor o mayor valor medio de compacidad de las celulas.
29.	Concavity_worst	Peor o mayor valor medio de concavidad de las celulas.
30.	Concave.point_worst	Peor o mayor valor medio de las porciones cóncavas del contorno de las celulas.
31.	Symmetry.worst	Peor o mayor valor medio de la simetría de las celulas.
32.	Fractal_dimenssion_worst	Peor o mayor valor medio de la dimensión fractal de las celulas.

Fuente: *UCI Machine Learning Repository.*

Ahora bien, para la optimización de los resultados es necesario seleccionar aquellos atributos del conjunto de datos que se consideren significativos para la extracción de información mediante la implementación de modelos en fases posteriores y aquellos que atributos que no se consideren relevantes para la investigación y se puedan excluir. Por lo tanto, teniendo en cuenta la revisión bibliográfica de artículos científicos sólo se excluirá el atributo Id o el identificador de cada registro, teniendo entonces un total de 31 variables con las que se trabajarán.

3.1.2.3. Descripción de los datos

Una vez hemos determinado los atributos relevantes para la investigación es necesario describir el tipo de atributo. Entonces, los tipos de datos que tenemos en nuestro conjunto de datos son los siguientes:

Tabla 8. Tipos de Atributos

Atributos	Tipos de atributos
Diagnosis	Cualitativo, Nominal
Radius_mean	Cuantitativo, Continuo
Texture_mean	Cuantitativo, Continuo
Parimeter__mean	Cuantitativo, Continuo
Area_mean	Cuantitativo, Continuo
Smoothness_mean	Cuantitativo, Continuo
Compactness_mean	Cuantitativo, Continuo
Concavity_mean	Cuantitativo, Continuo

Concave.points_mean	Cuantitativo, Continuo
Symmetry_mean	Cuantitativo, Continuo
Fractal_dimension_mean	Cuantitativo, Continuo
Radius_se	Cuantitativo, Continuo
Texture_se	Cuantitativo, Continuo
Perimeter_se	Cuantitativo, Continuo
Area_se	Cuantitativo, Continuo
Smoothness_se	Cuantitativo, Continuo
Compactnees_se	Cuantitativo, Continuo
Concavity_se	Cuantitativo, Continuo
Concave.points_Se	Cuantitativo, Continuo
Symmetry_se	Cuantitativo, Continuo
Fractal_dimension_se	Cuantitativo, Continuo
Radius_worst	Cuantitativo, Continuo
Texture_worst	Cuantitativo, Continuo
Perimeter_worst	Cuantitativo, Continuo
Area_worst	Cuantitativo, Continuo
Smoothness_worst	Cuantitativo, Continuo
Compactness_worst	Cuantitativo, Continuo
Concavity_worst	Cuantitativo, Continuo
Concave.point_worst	Cuantitativo, Continuo
Symmetry.worst	Cuantitativo, Continuo
Fractal_dimenssion_worst	Cuantitativo, Continuo

Fuente: *elaboración propia, los investigadores (2021).*

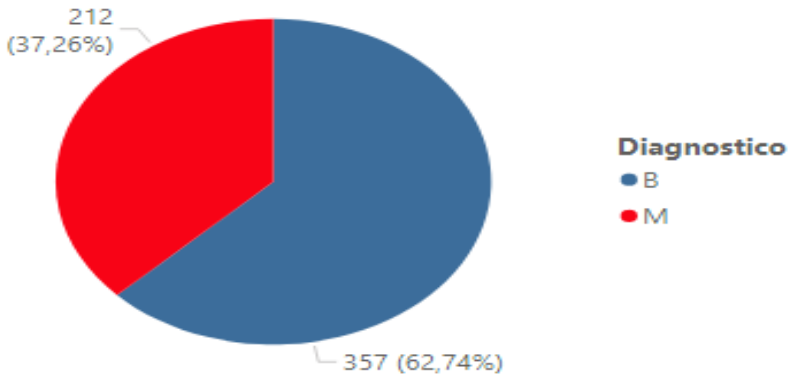
3.1.2.4. Exploración de los datos

Esta fase es necesaria e importante, porque es en esta que obtendremos de manera gráfica el primer acercamiento a la información o patrones ocultos, la organización y estructura general de los datos mediante tablas de frecuencias, diagramas de tortas o diagramas de distribución. Para ello, en primer lugar, se deben generar hipótesis o

preguntas que nos permitirán determinar los supuestos iniciales de la investigación, como vemos a continuación:

- **¿Cuántos pacientes fueron diagnosticados con cáncer benigno y maligno?**

Gráfico 1. Cáncer Benigno vs Maligno



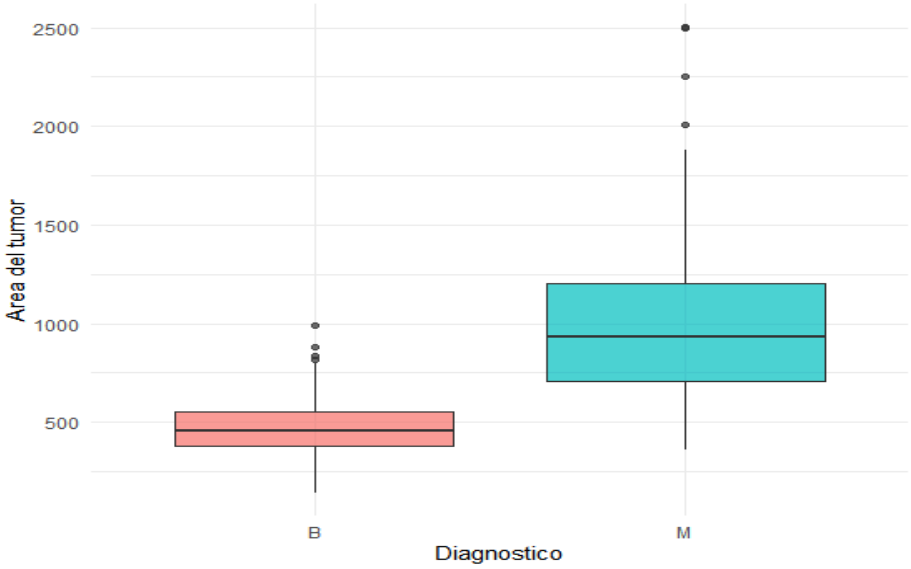
Fuente: elaboración propia, los investigadores (2021).

En el gráfico 1, se observa el porcentaje de los diagnósticos que tenemos en el conjunto de datos teniendo:

- Para células benignas: se muestran en el gráfico de color azul, se tienen 357 registros representando el 62.74% de los casos.
- Para células malignas: se muestran en el gráfico de color rojo, se tienen 212 registros representando el 37.26% de los casos.

- **¿Cómo es la distribución del área de las células mamarias con respecto a los diagnósticos (B=benigno / M=maligno)?**

Gráfico 2. Área de las células y Diagnóstico



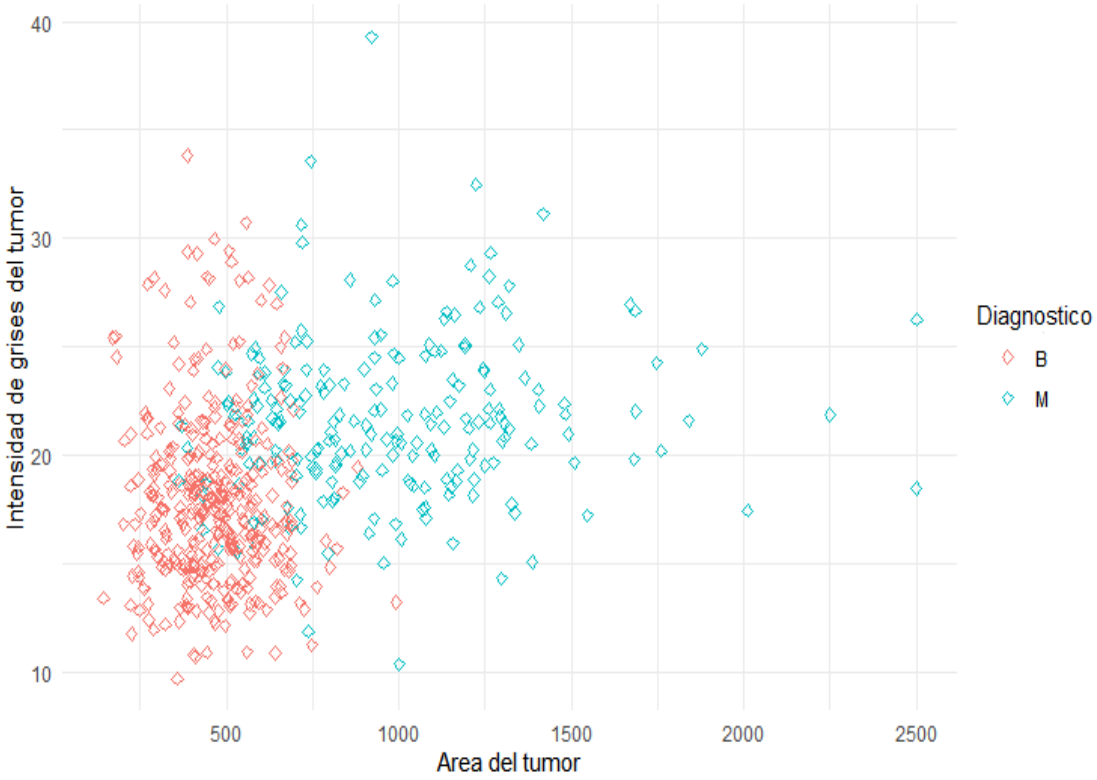
Fuente: *elaboración propia, los investigadores (2021).*

En el gráfico 2, se observa que para los casos de células malignas (M) la distribución de los datos indica que estas tienen áreas mayores a las células benignas (B) teniendo como valores atípicos, es decir, registros que difieren a los datos de las demás observaciones como un grupo los siguientes:

- Benignos: 4 valores atípicos.
- Malignos: 3 valores atípicos.

- **¿Existe alguna relación entre el área y la intensidad en la escala de grises entre las células?**

Gráfico 3. Área de las células vs Escala de grises

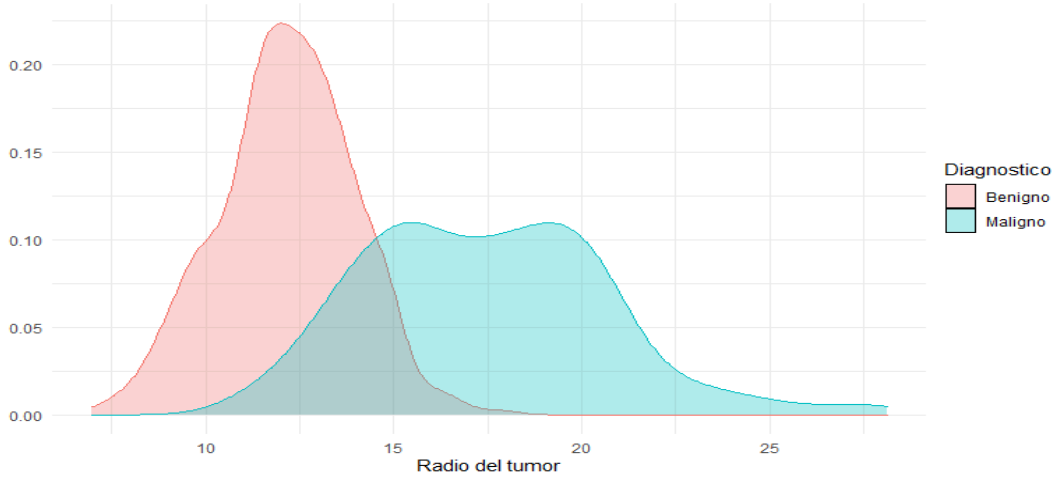


Fuente: *elaboración propia, los investigadores (2021).*

En la gráfica 3, se observa un diagrama de dispersión en la que los datos del área de las células con respecto a la intensidad de los tonos grises del tumor se encuentran demasiados dispersos, con lo que se puede afirmar que no existe ningún tipo de correlación o relación entre estas variables.

- **¿El radio de las células está relacionado con el tipo de diagnóstico?**

Gráfico 4. Radio de las células y Diagnóstico

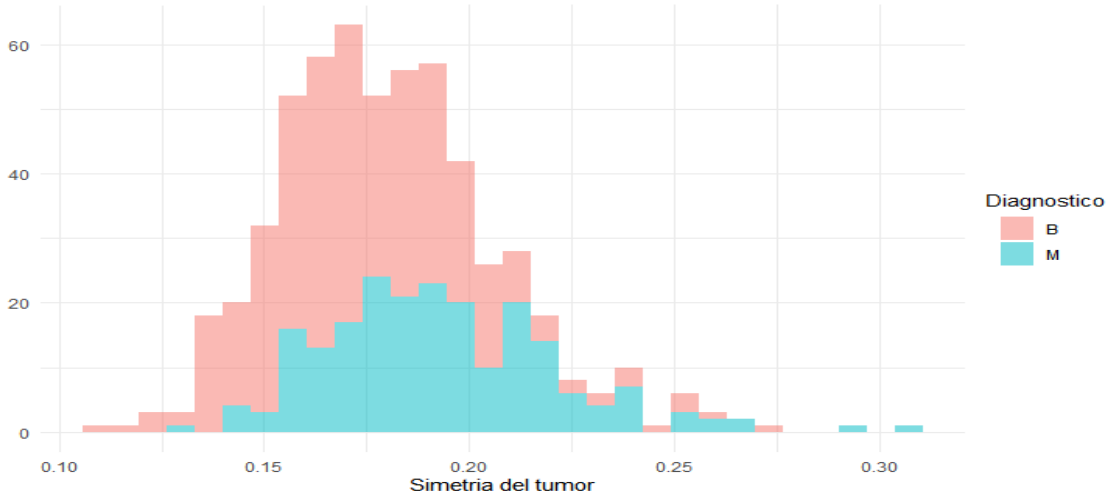


Fuente: *elaboración propia, los investigadores (2021).*

En la gráfica 4, se observa que, para el tipo de diagnóstico benigno el radio de las células no son mayores al radio de las células malignas, por lo que se puede concluir que desde el centro hasta el borde de estas células, es decir la extensión circular, son mayores en células malignas.

- **Es probable que la distorsión simétrica de las células esté relacionada con el diagnóstico.**

Gráfico 5. Distorsión simétrica vs Diagnóstico



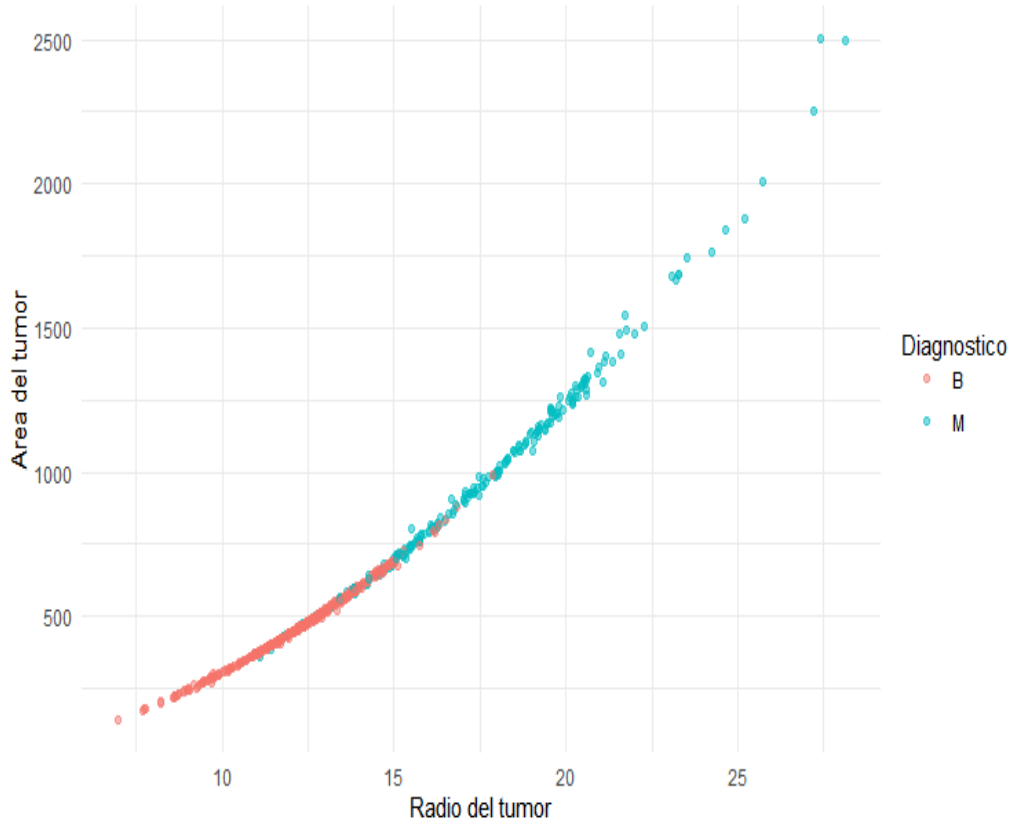
Fuente: *elaboración propia, los investigadores (2021).*

En el gráfico 5, se observa la frecuencia de los diagnóstico benignos (M) y malignos (M) con respecto a las simetrías de las células en la que se puede concluir que las células

malignas tienes una distorsión simétrica muy irregular (forma de las células) con respecto a las células benignas, propias de lo que se puede considerar su naturaleza como células cancerígenas.

- **La relación entre el área y el radio de las células determina el diagnóstico del cáncer.**

Gráfico 6. Área vs el Radio de las células

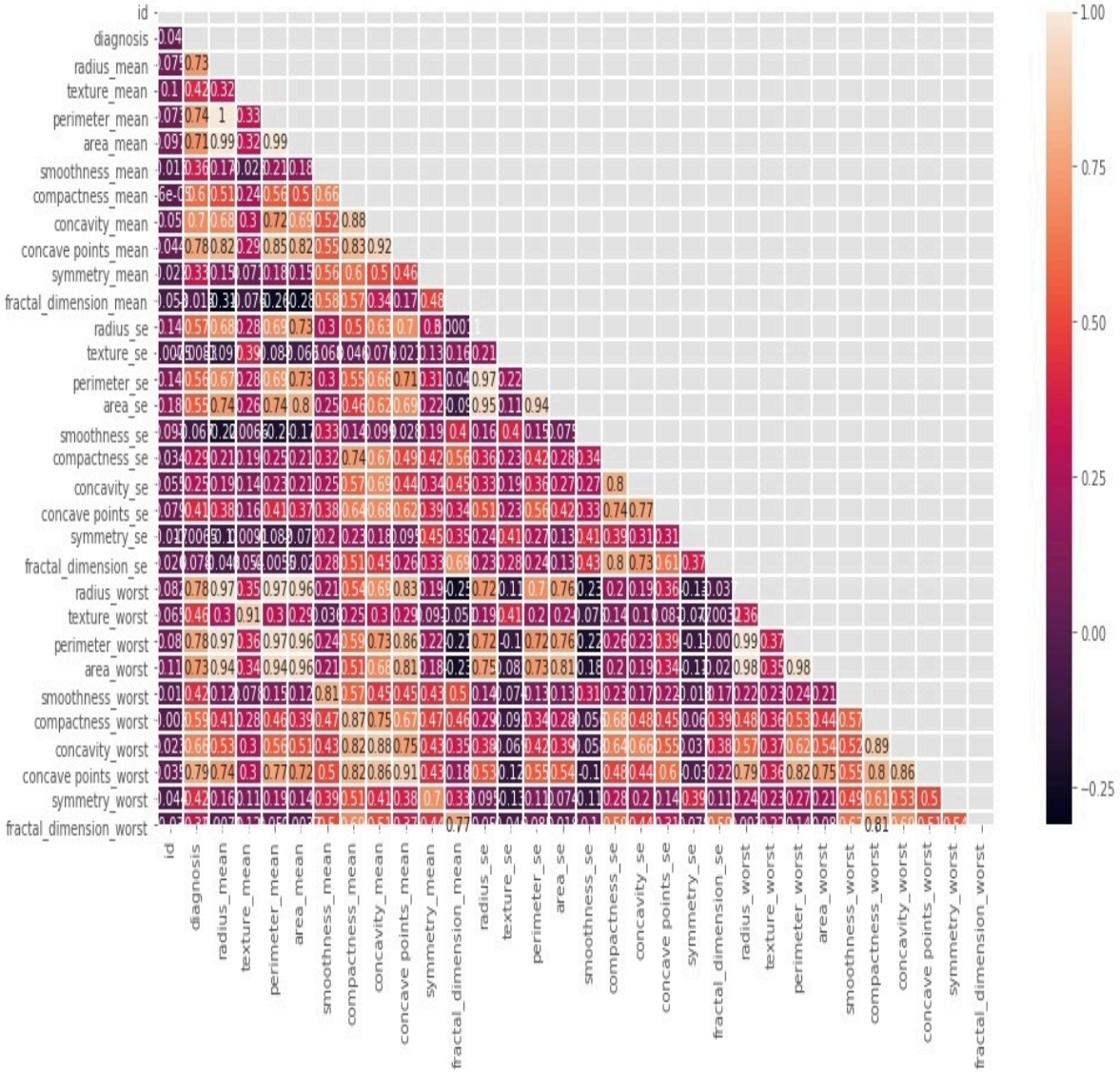


Fuente: *elaboración propia, los investigadores (2021).*

En el gráfico 6, se observa una relación directa entre el área del tumor con respecto a su radio, es decir, que mientras más grande sea el área de las células mayor será su circunferencias. Asimismo, la gráfica indica cómo al tener un área y radio menor estas células son mayormente benignas, mientras que al tener un área y radio mayor estas células son malignas y mucho más grandes.

Del gráfico anterior, surge la idea de representar en una matriz de correlación todas las variables de nuestro conjunto de datos cómo vemos a continuación:

Gráfico 7. Matriz de correlación de variables

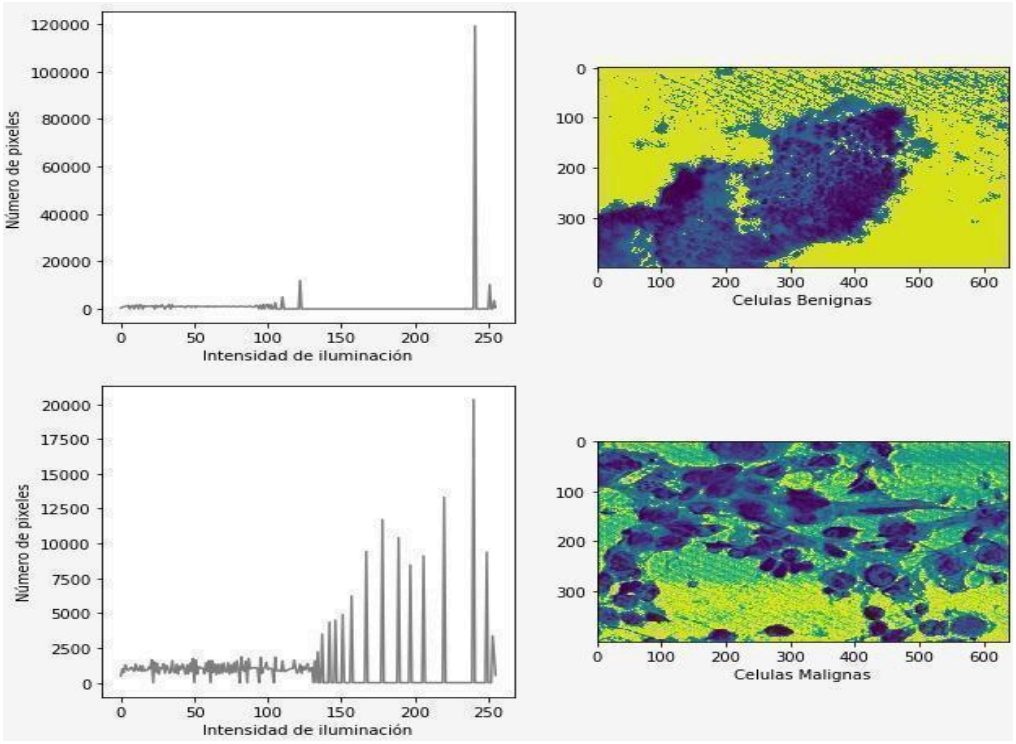


Fuente: *elaboración propia, los investigadores (2021).*

En la matriz del gráfico 7, podemos evidenciar la correlación entre las variables donde aquellos puntajes más cercanos a 1.00 indican que existe una correlación directa (representado por el color más claro), mientras que los puntajes más cercanos a -1.00 indican que existe una correlación inversa (representado por el color más oscuro) y por último, los puntajes comprendidos entre 0.2 a -0.2 indican una correlación nula, es decir, no existe relación entre las variables.

De igual manera y aprovechando las imágenes que tomó el Dr. Wolberg, el Prof. Mangasarian y Nick Street como fuente para la creación del conjunto de datos se obtiene el siguiente análisis:

Gráfico 8. No. de pixeles vs Intensidad de luz



Fuente: *elaboración propia, los investigadores (2021).*

En la gráfica 8, se observa una comparativa entre células benignas y malignas aplicando la técnica de ecualización del histograma, en la cual, podemos obtener la distribución uniforme de cada píxel con respecto a la intensidad de luz en la imagen digitalizada, evidenciando para este que caso una mayor sobreexposición anómala o dispersa de luz en la gráfica e imagen inferior que corresponde a células malignas.

3.1.2.5. Verificación de la calidad de los datos

En este mismo orden de ideas, a pesar que de este conjunto de datos cuenta con un número de instancias o registros relativamente pequeño cabe resaltar que estos datos fueron obtenidos de fuentes reales para el diagnóstico del cáncer de mama a partir de imágenes digitalizadas por el Dr. Wolberg y sus colegas el Prof. Mangasarian y Nick Street consolidando un sistema para el diagnóstico y pronóstico citológicos remotos del

cáncer de mama también conocido por sus siglas en inglés como Xcyt (System for Remote Cytological Diagnosis and Prognosis of Breast Cancer) y utilizado por muchos años por Wolberg en su carrera clínica [50].

Por ende, este sistema consistía en agregar tejidos de masas mamarias mediante la técnica conocida como biopsia por aspiración con aguja fina o FNA (consiste en extraer tejido de la región que se quieren examinar o estudiar), digitalizar las muestras en archivos de imágenes, mientras que el usuario a través del software procede a dibujar el contorno de las células para calcular las características como el radio, el área, la simetría etc., de cada una de las muestras obteniendo así un total de 569 casos con 30 características que conforman el conjunto de datos Breast Cancer Wisconsin (Diagnostic) Data Set [50]. Con respecto al pronóstico del cáncer de mama en benigno o maligno, a través del método del árbol multisuperficie antes mencionado, se calculó la probabilidad del tipo de cáncer de mama en los pacientes resultando así el atributo del diagnóstico en el conjunto de datos [50].

3.1.3. PREPARACIÓN DE LOS DATOS

En este punto, después de la recolección, descripción y exploración de los datos se procede a lo que se conoce como preprocesamiento de los datos que consiste en la selección, limpieza y formateo de los datos para optimizar la construcción de los modelos.

3.1.3.1. Seleccionar datos

Teniendo en cuenta que todos los atributos son prometedores para la investigación a excepción del atributo Id, estas variables están relacionados con los objetivos planteados y de igual manera se tomarán todos los registros del conjunto de datos el cual quedará conformado con 31 variables y 569 registros. A continuación, se muestran las variables con las que se trabajaran y un resumen estadístico de los datos que las conforman:

Atributos cualitativos, nominales:

Tabla 9. Atributos cualitativos

Atributos	Etiquetas	Descripción	Moda
Diagnosis	B / M	Benigno / Maligno	B

Fuente: *elaboración propia, los investigadores (2021).*

Atributos cuantitativos, continuos:

Tabla 10. Atributos cuantitativos

Atributos	Valor mínimo	Primer cuartil	Mediana	Media	Tercer cuartil	Valor máximo
Radius_mean	6.981	11.700	13.370	14.127	15.780	28.110
Texture_mean	9.71	16.17	18.84	19.29	21.80	39.28
Parimeter__mean	43.79	75.17	86.24	91.97	104.10	188.50
Area_mean	143.5	420.3	551.1	654.9	782.7	2501.0
Smoothness_mean	0.05263	0.08637	0.09587	0.09636	0.10530	0.16340
Compactness_mean	0.01938	0.06492	0.09263	0.10434	0.13040	0.34540
Concavity_mean	0	0.02956	0.06154	0.08880	0.13070	0.42680
Concave.points_mean	0	0.02031	0.03350	0.04892	0.07400	0.20120
Symmetry_mean	0.1060	0.1619	0.1792	0.1812	0.1957	0.3040
Fractal_dimension_mean	0.04996	0.05770	0.06154	0.06280	0.06612	0.09744
Radius_se	0.1115	0.2324	0.3242	0.4052	0.4789	2.8730
Texture_se	0.3602	0.8339	1.1080	1.2169	1.4740	4.8850
Perimeter_se	0.757	1.606	2.287	2.866	3.357	21.980
Area_se	6.802	17.850	24.530	40.337	45.190	542.200
Smoothness_se	0.001713	0.005169	0.006380	0.007041	0.008146	0.031130
Compactnees_se	0.002252	0.013080	0.020450	0.025478	0.032450	0.135400
Concavity_se	0	0.01509	0.02589	0.03189	0.04205	0.39600
Concave.points_Se	0	0.007638	0.010930	0.011796	0.014710	0.052790
Symmetry_se	0.007882	0.015160	0.018730	0.020542	0.023480	0.078950
Fractal_dimension_se	0.0008948	0.0022480	0.0031870	0.0037949	0.0045580	0.0298400
Radius_worst	7.93	13.01	14.97	16.27	18.79	36.04
Texture_worst	12.02	21.08	25.41	25.68	29.72	49.54
Perimeter_worst	50.41	84.11	97.66	107.26	125.40	251.20
Area_worst	185.2	515.3	686.5	880.6	1084.0	4254.0
Smoothness_worst	0.07117	0.11660	0.13130	0.13237	0.14600	0.22260
Compactness_worst	0.02729	0.14720	0.21190	0.25427	0.33910	1.05800
Concavity_worst	0	0.1145	0.2267	0.2722	0.3829	1.2520

Concave.point_worst	0	0.06493	0.09993	0.11461	0.16140	0.29100
Symmetry.worst	0.1565	0.2504	0.2822	0.2901	0.3179	0.6638
Fractal_dimenssion_worst	0.05504	0.07146	0.08004	0.08395	0.09208	0.20750

Fuente: *elaboración propia, los investigadores (2021).*

3.1.3.2. Limpiar los datos

El conjunto de datos con el que se cuenta no tiene valores erróneos o caracteres especiales que puedan afectar a la construcción de los modelos. Asimismo, este conjunto de datos no tiene valores faltantes como vemos en la siguiente tabla:

Tabla 11. Estructura y datos faltantes

Estructura de los datos		Datos faltantes	
diagnosis	object	diagnosis	0
radius_mean	float64	radius_mean	0
texture_mean	float64	texture_mean	0
perimeter_mean	float64	perimeter_mean	0
area_mean	float64	area_mean	0
smoothness_mean	float64	smoothness_mean	0
compactness_mean	float64	compactness_mean	0
concavity_mean	float64	concavity_mean	0
concave points_mean	float64	concave points_mean	0
symmetry_mean	float64	symmetry_mean	0
fractal_dimension_mean	float64	fractal_dimension_mean	0
radius_se	float64	radius_se	0
texture_se	float64	texture_se	0
perimeter_se	float64	perimeter_se	0
area_se	float64	area_se	0
smoothness_se	float64	smoothness_se	0
compactness_se	float64	compactness_se	0
concavity_se	float64	concavity_se	0
concave points_se	float64	concave points_se	0
symmetry_se	float64	symmetry_se	0
fractal_dimension_se	float64	fractal_dimension_se	0
radius_worst	float64	radius_worst	0
texture_worst	float64	texture_worst	0
perimeter_worst	float64	perimeter_worst	0
area_worst	float64	area_worst	0
smoothness_worst	float64	smoothness_worst	0
compactness_worst	float64	compactness_worst	0
concavity_worst	float64	concavity_worst	0
concave points_worst	float64	concave points_worst	0
symmetry_worst	float64	symmetry_worst	0
fractal_dimension_worst	float64	fractal_dimension_worst	0

Fuente: *elaboración propia, los investigadores (2021).*

Cabe mencionar que en caso de haber encontrado valores faltantes, estos serían reemplazados por la moda para el atributo cualitativo (es decir por el valor que más se

repite), mientras que para los atributos cuantitativos por la media de valores respectivamente.

3.1.3.3. Construcción de nuevos datos

En este apartado sólo se considera transformar el tipo de archivo del conjunto de datos, de *.csv* a *.xlsx* para facilitar la manipulación de los mismos y este no requiere la construcción de atributos derivados y de nuevo registros.

3.1.3.4. Integración de datos

Como se menciona en el apartado anterior no se consideró necesario crear una nueva estructura, reorganizar, fusionar y adicionar nuevos datos ya que con los cuentan están organizados y óptimos para su procesamiento.

3.1.3.5. Formateo de datos

Por último, en este apartado debemos preguntarnos, ¿Qué modelos se pretenden implementar?, ¿requieren estos modelos la transformación de una o más variables?. Respondiendo antes estos interrogantes, aplicaremos 7 modelos los cuales son Random Forest, KNN, SVM, Naive Bayes, Regression Logistic, Decision Tree y Gradient Boosting Tree. Para ello, se considera cambiar los valores cualitativos a valores cuantitativos nuestra variable predictora diagnóstica debido a que modelos como el de Regression Logistic requieren este tipo de variables para su implementación.

3.1.4. MODELADO

3.1.4.1. Escoger las técnicas de modelado

En primer lugar, antes de la construccion e implementacion de los modelos es importante definir las técnicas con las que se trabajarán, las cuales son: Random forest, k-nearest neighbors, Support Vector Machines, Decision trees, Gradient Boosting Trees, LogisticRegression y Naive Bayes.

Random Forest: también conocido como bosques aleatorios, es un algoritmo de tipo predictivo que aplica la técnica de bagging, la cual consiste en combinar diferentes árboles que toman observaciones y variables aleatorias, seleccionando registros al azar aplicando un muestreo con reemplazo para crear distintos conjuntos de datos, luego crea

un árbol de decisión para cada conjunto de datos, teniendo como resultado diferentes árboles dejándolos crecer en profundidad, es decir sin podar, con diferentes registros y variables [12]. Por último, se predicen nuevos datos usando lo que se conoce como voto mayoritario donde se clasifica como positivo si la mayoría de los árboles predice observaciones como positivo o se clasifica como negativo si la mayoría de los árboles predicen observaciones como negativo [12].

k-Nearest Neighbors, KNN: también conocido como K vecinos más cercanos, es un algoritmo de clasificación supervisado que consiste en clasificar las nuevas instancias o registros en la clase más frecuente según el número k vecinos más próximos [12]. Cabe resaltar que este algoritmo pierde precisión si se utilizan datos irrelevantes. De igual manera, es importante señalar que si se utilizan datos categóricos este algoritmo devolverá la categoría a la que pertenece el registro nuevo o desconocido, si se utilizan datos continuos, este devolverá la media de los valores de los vecinos más cercanos y por último, si este algoritmo de clasificación es binaria, la clasificación de los nuevos registros se fundamenta en el voto (mayoría) para decir el correctamente el resultado por lo que es recomendable elegir k o el número de vecinos más cercanos impar para evitar empates [12].

Support Vector Machines, SVM: también como Máquina de Soporte Vectorial y al igual que Knn es un algoritmo de clasificación supervisado[12]. Este algoritmo trabaja de la siguiente manera: a través de un conjunto de datos en el que cada instancia pertenece a una de los 2 posibles categorías o clases, este algoritmo construye un modelo que predice a qué categoría pertenece una instancia nueva o desconocida[12]. De igual manera, este algoritmo al igual que otros algoritmos supervisados toman los datos de entrada como una lista ordenada de p números denominado vector p-dimensional. Por último, este algoritmo a través de un plano separa los datos clasificados en una clase u otra [12].

Decision Trees: también conocido como árboles de decisión, es un modelo de tipo predictivo que gráficamente representan condiciones y acciones de manera secuencial categorizando estas condiciones con el fin de resolver el problema [12]. Estos árboles de decisión son utilizados ampliamente en análisis crediticios y diagnósticos médicos [12].

Naive Bayes, NB: también es conocido como bayesiano ingenuo ya que este toma por hecho que cada variable de entrada es independiente [15]. Este es un algoritmo

supervisado basado en el teorema de bayes muy útil para llevar a cabo predicciones calculando la posibilidad de cada clase y la probabilidad de clases teniendo en cuenta cada valor a partir de los datos de entrenamiento [15].

Gradient Boosting Tree: también conocido por como Árboles de aumento de gradiente, está conformado por un conjunto de árboles de decisión llamados ensemble que son entrenados de forma secuencial generando para cada árbol en particular observaciones distribuidas en nodos generando la estructura de un árbol hasta llegar a un nodo final [50]. Este método trabaja de manera iterativa donde cada árbol aprende de los errores del árbol anterior de manera que cada predicción a raíz de una nueva observación es el resultado de la predicción de cada árbol que conforma el conjunto [51].

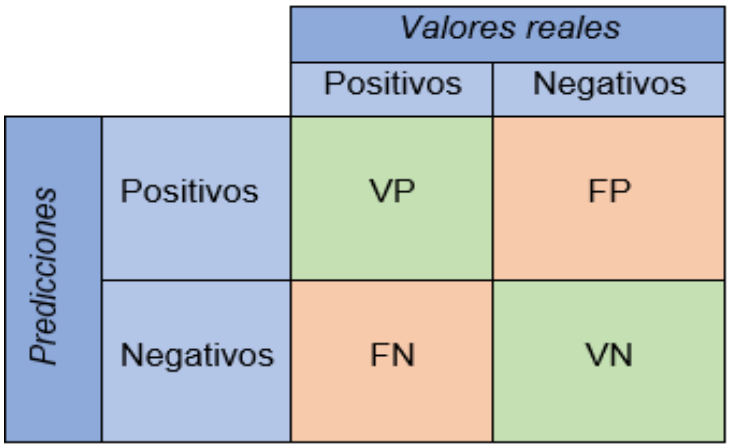
Logistic Regression, RL: también conocido como Regresión Logística, este algoritmo basado en la regresión lineal permite a través de variables cuantitativas estimar la probabilidad de una variable cualitativa binaria muy útil y frecuentemente utilizado en la medicina [52]. Por ejemplo, clasificar a un individuo sufre o no de alguna enfermedad cardiovascular en función a su peso corporal [52].

3.1.4.2. Generar Plan de prueba

Para garantizar la calidad en los resultados de nuestros modelos aplicaremos las siguientes pruebas:

Matriz de confusión: ésta nos permite medir el desempeño de los modelos a implementar. En base a las investigaciones consultadas, esta matriz es representada de la siguiente manera:

Tabla 12. Matriz de confusión



Fuente: *Consulta bibliográfica.*

Donde:

- VP o verdaderos positivos, nos muestra la cantidad de casos positivos que fueron predichos y clasificados correctamente como positivos en nuestro modelo, por ejemplo, el paciente tiene cáncer y el modelo demuestra que si tiene cáncer.
- FN o falsos negativos, nos muestra la cantidad de casos positivos que fueron predichos y clasificados incorrectamente como negativos en nuestro modelo, por ejemplo, el paciente tiene cáncer y el modelo demuestra incorrectamente que no tiene cáncer.
- FP o falsos positivos, nos muestra la cantidad de casos negativos que fueron predichos y clasificados incorrectamente como positivos en nuestro modelo, por ejemplo, el paciente no tiene cáncer y el modelo demuestra incorrectamente que si tiene cáncer.
- VN o verdaderos negativos, nos muestra la cantidad de casos negativos que fueron clasificados correctamente como negativos en nuestro modelo, por ejemplo, el paciente no tiene cáncer y el modelo demuestra correctamente que no tiene cáncer.

Por consiguiente, de lo anterior se pueden calcular las siguientes métricas:

- *Exactitud:* es el porcentaje de casos que fueron predichos y clasificados correctamente. Se calcula de la siguiente manera:

$$Exactitud = \frac{VP+VN}{TOTAL}$$

- *Precisión:* es el porcentaje de casos positivos que fueron predichos y clasificados correctamente. Se calcula de la siguiente manera:

$$Precisión = \frac{VP}{VP+FP}$$

- *Sensibilidad:* conocido también como exhaustividad o recall, es el porcentaje de casos positivos con respecto al total de casos positivos predichos y clasificados correctamente. Se calcula de la siguiente manera:

$$Recall = \frac{VP}{VP+FN}$$

- Puntuación F1: es una puntuación que representa la media aritmética entre la precisión y la exhaustividad. Se calcula de la siguiente manera:

$$F1 = 2 * \frac{precisión*recall}{precision+recall}$$

- Especificidad: es el porcentaje de casos negativos predichos y clasificados correctamente. Se calcula de la siguiente manera:

$$Especificidad = \frac{VN}{FP+VN}$$

- Tasa de falsos negativos: es el porcentaje de casos positivos predichos y clasificados incorrectamente como negativos. Se calcula de la siguiente manera:

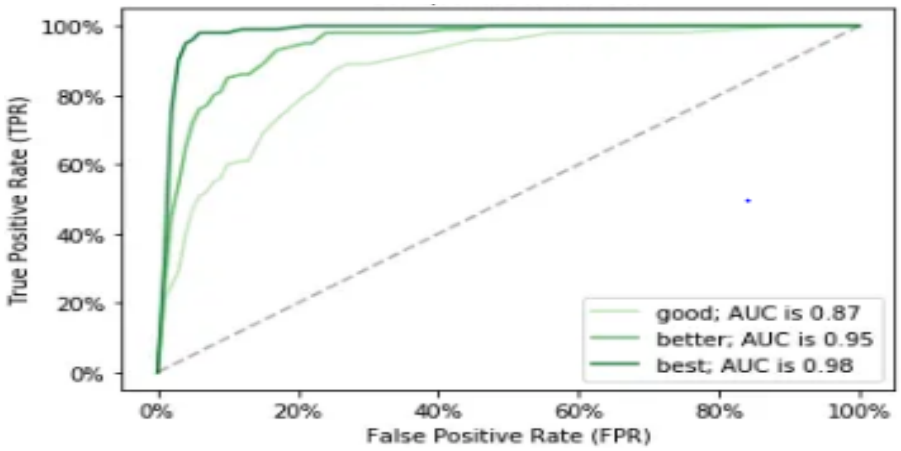
$$FN = \frac{FN}{VP+FN}$$

- Tasa de falsos positivos: es el porcentaje de casos negativos predichos y clasificados incorrectamente como positivos. Se calcula de la siguiente manera:

$$FP = \frac{FP}{VN+FP}$$

- Área bajo la curva o AUC: es la relación entre el recall y la tasa de falsos positivos, en la que se obtiene lo que se conoce como la curva de característica operativa del receptor o ROC, la cual, nos permite ver en a través de un gráfico similar a la ilustración 8, dicha relación, en la que la línea diagonal representa casos ideales o de suposición de los casos predichos en la que cualquier valor que esté por debajo de esta es considerado erróneo mientras que los valores que se encuentran por encima de esta son considerados correctos.

Ilustración 8. Curva ROC



Fuente: *Consulta bibliográfica.*

3.1.4.3. Construir el modelo

Una vez definidas las técnicas a implementar procederemos a construir nuestros modelos con el lenguaje de programación Python usando la librería Scikit-learn a través del entorno de trabajo Jupyter. Esta librería de código abierto contiene una serie de algoritmos, funciones y métodos que nos permiten el procesamiento de nuestro conjunto de datos, el entrenamiento, las pruebas y la optimización de nuestros modelos.

Por lo tanto, en primer lugar, a nuestra variable diagnosis que contiene B (benigno) y M (maligno) lo convertimos a una variable numérica ya que esta manera facilita el tratamiento de nuestros modelos, donde B será igual a 0 y M será igual a 1. Luego, separamos nuestro conjunto de datos con nuestra clase, es decir, la clase diagnosis en el conjunto de datos *X* y el resto de las variables en el conjunto de datos *Y*. Una vez tengamos separado nuestra variable predictora diagnosis de las demás variables, se procede a establecer los conjunto de datos de entrenamiento y prueba con la función

train_test_split asigando el 80% de los datos para el entrenar los modelos y el 20% restante para realizar las pruebas. Estas variables llevan por nombre *X_train* y *X_test*. Asimismo, para nuestra clase diagnosis la almacenaremos en las variables *y_train* y *y_test* para predecir los diagnósticos.

Seguidamente, se crea la función *Classifiers()* en donde agregaremos un total de 7 modelos y evaluaremos sus resultados para obtener de manera rápida y eficaz los modelos más óptimos a implementar. Estos modelos son: *LogisticRegression(),SVM(), KNeighborsClassifier(), RandomForestClassifier(), DecisionTreeClassifier(), GradientBoostingClassifier(), GaussianNB()* y con la funcion *croos_val_score()* evaluamos la exactitud de nuestro entrenamiento, teniendo los siguientes resultados:

Ilustración 9. Modelos

```
MODELO:  LogisticRegression TIENE UN PUNTAJE DE ENTRENAMIENTO DE  55.00000000000001 % DE EXACTITUD
MODELO:  KNeighborsClassifier TIENE UN PUNTAJE DE ENTRENAMIENTO DE  75.0 % DE EXACTITUD
MODELO:  SVC TIENE UN PUNTAJE DE ENTRENAMIENTO DE  66.0 % DE EXACTITUD
MODELO:  DecisionTreeClassifier TIENE UN PUNTAJE DE ENTRENAMIENTO DE  100.0 % DE EXACTITUD
MODELO:  GaussianNB TIENE UN PUNTAJE DE ENTRENAMIENTO DE  65.0 % DE EXACTITUD
MODELO:  RandomForestClassifier TIENE UN PUNTAJE DE ENTRENAMIENTO DE  99.0 % DE EXACTITUD
MODELO:  GradientBoostingClassifier TIENE UN PUNTAJE DE ENTRENAMIENTO DE  100.0 % DE EXACTITUD
```

Fuente: *elaboración propia, los investigadores (2021).*

Como vemos, en la imagen anterior los modelos que mejor resultados obtuvieron fueron GradientBoostingClassifier(), DecisionTreeClassifier() con un porcentaje del 100% de exactitud para ambos casos y RandomForestClassifier() con porcentaje 99% de exactitud.

Es importante mencionar que se aplicó de manera conjunta los modelos para descartar aquellos que no son utilies para el complimiento de los objetivos debido al bajo porcentaje obtenido.

Por lo tanto, conociendo que modelos presentaron mejor desempeño se procede a construirlos de manera individual con los mismos parámetros antes mencionados aplicando la función *gnb.predict()* en nuestro conjunto de pruebas *X_test* y con nuestra clase predictora contenida en las varibales *y_train* y *y_test*. El rendimiento de los modelos se calculó con la función *gnb.score()* y se obtuvieron los siguientes resultados:

Random Forest: para este modelo se obtuvo, una precisión para el set de entrenamiento de 1.0, para el set de prueba de 0.96, con un error absoluto medio de 0.043, error cuadrático medio de 0.043, raíz del error cuadrático 0.20.

Matriz de confusión:

Tabla 13. Matriz confusión Random Forest

	Positivos	**Negativos**
Positivos	58	0
Negativos	5	51

Fuente: *elaboración propia, los investigadores (2021).*

Esta matriz nos indica que:

- Verdaderos Positivos (VP): nos muestra la cantidad de valores positivos que fueron clasificados correctamente como positivos en nuestro modelo, siendo un total de 58 valores predichos correctamente.
- Falsos Negativos (FN): nos muestra la cantidad de positivos que fueron clasificados incorrectamente como negativos 8 en nuestro modelo, siendo un total de 5 valores predichos correctamente.
- Falsos Positivos (FP): no muestra la cantidad de negativos que fueron clasificados incorrectamente como positivos en nuestro modelo, siendo 0 los predichos incorrectamente.
- Verdaderos Negativos (VN): nos muestra la cantidad de negativos que fueron clasificados correctamente como negativos por el modelo, siendo 51 los predichos correctamente.

Por lo tanto, los resultados obtenidos de este modelo fueron:

Tabla 14. Métricas Random Forest

Precisión	**1.0**
Exactitud	**0.95**
Sensibilidad	**0.91**
Puntuación F1	**0.95**

Especificidad	**1.0**
Tasa de falsos negativos	**1.0**
Tasa de falsos positivos	**0**
AUC	**0.99**

Fuente: *elaboración propia, los investigadores (2021).*

Decision Trees: para este modelo se obtuvo una precisión para el set de entrenamiento de 1.0, para el set de prueba de 0.88, con un error absoluto medio de 0.12, error cuadrático medio de 0.12, raíz del error cuadrático 0.35.

Matriz de confusión:

Tabla 15. Matriz de confusión Decision Trees

	Positivos	**Negativos**
Positivos	**49**	**9**
Negativos	**5**	**51**

Fuente: *elaboración propia, los investigadores (2021).*

Esta matriz nos indica que:

- Verdaderos Positivos (VP): nos muestra la cantidad de valores positivos que fueron predichos y clasificados correctamente como positivos en nuestro modelo, siendo un total de 49 casos predichos correctamente.
- Falsos Negativos (FN): nos muestra la cantidad de positivos que fueron predichos y clasificados incorrectamente como negativos 5 en nuestro modelo, siendo un total de 8 casos predichos incorrectamente.
- Falsos Positivos (FP): no muestra la cantidad de negativos que fueron predichos y clasificados incorrectamente como positivos en nuestro modelo, siendo 9 los casos predichos incorrectamente.
- Verdaderos Negativos (VN): nos muestra la cantidad de negativos que fueron predichos y clasificados correctamente como negativos por el modelo, siendo 51 los casos predichos correctamente.

Por lo tanto, los resultados obtenidos de este modelo fueron:

Tabla 16. Métricas Decision Trees

Precisión	**0.85**
Exactitud	**0.87**
Sensibilidad	**0.91**
Puntuación F1	**0.88**
Especificidad	**0.85**
Tasa de falsos negativos	**0.092**
Tasa de falsos positivos	**0.15**
AUC	**1.0**

Fuente: *elaboración propia, los investigadores (2021).*

Gradient Boosting: para este modelo se obtuvo, Una precisión para el set de entrenamiento de 1.0, para el set de prueba de 0.97, con un error absoluto medio de 0.026, error cuadrático medio de 0.026, raíz del error cuadrático 0.16.

Matriz de confusión:

Tabla 17. Matriz de confusión Gradient Boosting

	Positivos	**Negativos**
Positivos	**58**	**0**
Negativos	**3**	**53**

Fuente: *elaboración propia, los investigadores (2021).*

Esta matriz nos indica que:

- Verdaderos Positivos (VP): nos muestra la cantidad de valores positivos que fueron predichos y clasificados correctamente como positivos en nuestro modelo, siendo un total de 58 casos predichos correctamente.
- Falsos Negativos (FN): nos muestra la cantidad de positivos que fueron predichos y clasificados incorrectamente como negativos en nuestro modelo, siendo un total de 3 casos predichos incorrectamente.

- Falsos Positivos (FP): no muestra la cantidad de negativos que fueron predichos y clasificados incorrectamente como positivos en nuestro modelo, siendo 0 los predichos incorrectamente.
- Verdaderos Negativos (VN): nos muestra la cantidad de negativos que fueron predichos y clasificados correctamente como negativos por el modelo, siendo 53 los predichos correctamente.

Por lo tanto, los resultados obtenidos de este modelo fueron:

Tabla 18. Métricas Gradient Boosting

Precisión	**1.0**
Exactitud	**0.97**
Sensibilidad	**0.94**
Puntuación F1	**0.96**
Especificidad	**1**
Tasa de falsos negativos	**0.049**
Tasa de falsos positivos	**0**
AUC	**1.0**

Fuente: *elaboración propia, los investigadores (2021).*

3.1.5. FASE DE EVALUACIÓN

En este apartado compararemos los modelos construidos para determinar si cumplen con los objetivos expuestos dentro de la metodología y poder llevar a cabo la fase final de implementación. Siendo el caso, estos objetivos no son cumplidos es pertinente llevar a cabo una revisión detallada y minuciosa de todos los procesos realizados ya que no se estaría cumpliendo el propósito y la razón de ser de la investigación.

3.1.5.1. Evaluar los resultados

Durante la construcción de los modelos se implementó una función que permite agrupar y ejecutar los 7 modelos con los que se tenían previsto trabajar teniendo como resultado en su gran mayoría altos porcentajes de exactitud. Este proceso previo ayudó a reconocer aquellos modelos que mejores resultados presentaron y de esta manera se agilizan las tareas llevadas a cabo en la fase del modelado. Estos modelos fueron Random Forest,

Decision Trees y Gradient Boosting. La comparación entre las métricas obtenidas se puede ver reflejado en la siguiente tabla:

Tabla 19. Comparación de las métricas de los modelos

Métricas	Random Forest	Decision Trees	Gradient Boosting
Precisión	1.0	0.85	1.0
Exactitud	0.95	0.87	0.97
Sensibilidad	0.91	0.91	0.94
Puntuación F1	0.95	0.88	0.96
Especificidad	1.0	0.85	1
Tasa de falsos negativos	1.0	0.092	0.049
Tasa de falsos positivos	0	0.15	0
AUC	0.99	1.0	1.0

Fuente: *elaboración propia, los investigadores (2021).*

Como podemos observar, estos modelos tienen un alto porcentaje en cuanto a la exactitud y precisión de los modelos. Ahora bien, ¿Cuál de los tres modelos es el más óptimo?. Teniendo en cuenta la métrica puntuación F1 podemos decir que el modelo más óptimo es Gradient boosting.

Con esto, los modelos aprobados y que logran cumplir con los objetivos propuestos son los mencionados RandomForest, Decision trees y Gradient boosting, mientras que regresion logistica, knn, máquinas de soporte vectorial y naive bayes son los modelos que serán descartados por su bajo desempeño evidenciado en la ilustración 9.

3.1.5.2. Revisar el proceso

Hasta este apartado de la investigación se han ejecutado gran parte de las tareas propuestas por la metodología CRISP-DM sin problema alguno y obteniendo resultados favorables para el diagnóstico del cáncer de mama.

3.1.5.3. Determinar los próximos pasos

Una vez revisados los procesos llevados a cabo en el desarrollo de la investigación se procederá a la fase final de la metodología propuesta, es decir, la fase de la implementación, en la cual, se detallarán los planes contemplados para la aplicación de

los modelos en la liga contra el cancer – seccional cesar, un informe final en donde se expondrán los resultados obtenidos y por último una revisión final de la investigación.

3.1.6. FASE DE IMPLEMENTACIÓN

Esta última fase tiene como propósito, organizar y exponer al cliente final los resultados obtenidos presentando los planes de implementación, monitoreo y un informe del producto final.

3.1.6.1. Plan de implementación

Para la aplicación de esta investigación en un entorno de negocio real como lo es el diagnóstico del cáncer de mama en la liga contra el cancer – seccional cesar se desarrollara un aplicativo web que integre los modelos aprobados en fases posterior, en donde un profesional de la salud podrá llevar a cabo la gestión del diagnóstico de los pacientes y de igual manera, donde cada paciente podrá respectivamente consultar dicho diagnóstico y obtener información sobre el cáncer de mama. Para esto, se capacitará al personal encargado de la entidad en cuanto a la usabilidad del software.

Por otra parte, dicho aplicativo web estará alojado en un servidor privado virtual que facilitará su uso y el proceso de socialización del proyecto.

3.1.6.2. Plan de monitoreo

Se tiene previsto almacenar los datos de los diagnósticos para establecer en un conjunto de datos que permita obtener información estadística, ya sea, a través de gráficos o visualizaciones que ayuden a interpretar los datos históricos que serán extraídos y almacenados en un formato de hoja de cálculo como excel, con la finalidad de que estos datos puedan formar parte de una investigación futura. Este proceso de extracción de datos se tiene planeado llevar a cabo en de manera semestral.

En cuanto a las actividades de socialización del aplicativo web a nuevo personal encargado o tareas de mantenimiento al software, serán llevadas a cabo por solicitud de la liga contra el cáncer teniendo en cuenta la disponibilidad de los investigadores responsables del proyecto.

3.1.6.3. Informe final

Hoy en día, es claro que los medios tecnológicos son fundamentales para llevar a cabo todo tipo de procesos en cualquier tipo escenarios y por ello esta investigación se considera muy importante porque permitirá a la liga contra el cáncer – seccional cesar diagnosticar el cáncer de mama de manera eficaz y oportuna.

De manera resumida, para la realización de la investigación se recopiló información de la Liga Contra el Cáncer – Seccional Cesar, información como su estructura organizacional, su situación actual y gestiones durante la pandemia por COVID – 19, que permitieron establecer los objetivos comerciales o de negocio y por consiguiente los objetivos de la minería de datos.

Seguidamente, se valoró la situación en cuanto a los datos con los que se pretendían trabajar, los beneficios, supuestos, límites, exclusiones, los riesgos y sus contingencias de cara al desarrollo de la investigación.

Por otro lado, se obtuvo el conjunto de datos del repositorio de datos libre UCI denominado “Breast Cancer Wisconsin (Diagnostic) Data Set” y una vez obtenido el conjunto de datos se establecieron hipótesis que permitieron llevar a cabo análisis exploratorios mediante gráficos y visualizaciones de los mismos, con la finalidad de elegir los modelos y técnicas apropiadas la para construcción de los mismos.

Dicho esto, la minería de datos a través de la metodología CRISP-DM nos permitió llevar a cabo tareas y actividades que permitieron desarrollar e implementar modelos con un alto porcentaje de efectividad, precisión y exactitud en cuanto al diagnóstico del cáncer de mama benigno y maligno, logrando cumplir con los criterios de éxito de nuestros objetivos planteados desde la perspectiva de la minería de datos y de negocio obteniendo más de un 90% fiabilidad en los modelos seleccionados y aprobados los cuales fueron: Árboles de Decisión (en inglés, Decision Trees), Bosques Aleatorios (en inglés, Random Forest) y Potenciación del Gradiente (en inglés, Gradient Boosting).

Cabe resaltar que se tuvieron en cuenta un total de 7 modelos, pero solo fueron seleccionados los nombrados anteriormente debido a su alto porcentaje, ya que estos modelos presentaron hasta un 100% de exactitud. Así mismo, para el entrenamiento y

prueba de cada uno de estos modelos no fue posible adquirir datos de pacientes de la Liga Contra el Cáncer - Seccional Cesar para probar los modelos construidos debido a la confiabilidad de los mismos y de igual manera los datos para generar nuevos diagnósticos deberán ser extraídos con la misma técnica con la que fue construido el conjunto de datos con el que se trabajó, es decir, características de las células o tejido mamario a través de biopsias por aspiración con aguja fina.

3.1.6.4. Revisión del proyecto

Es satisfactorio el logro de los resultados obtenidos durante el desarrollo de la investigación, no se presentaron mayores inconvenientes en el desarrollo de cada una de las tareas, actividades y procesos de la metodología propuesta consiguiendo de esta manera llevar a cabo todas estas.

Solo por mencionar una dificultad considerada externa al proyecto o investigación fue la situación de la actual pandemia por COVID – 19 ya que esta era un inconveniente para llevar a cabo reuniones entre nosotros los investigadores y a su vez con nuestro director de proyecto el profesor Alvaro Oñate Bowen y el personal de la Liga Contra el Cancer – Seccional Cesar.

3.2. APLICATIVO WEB: DIAGNOSIS BREAST CANCER

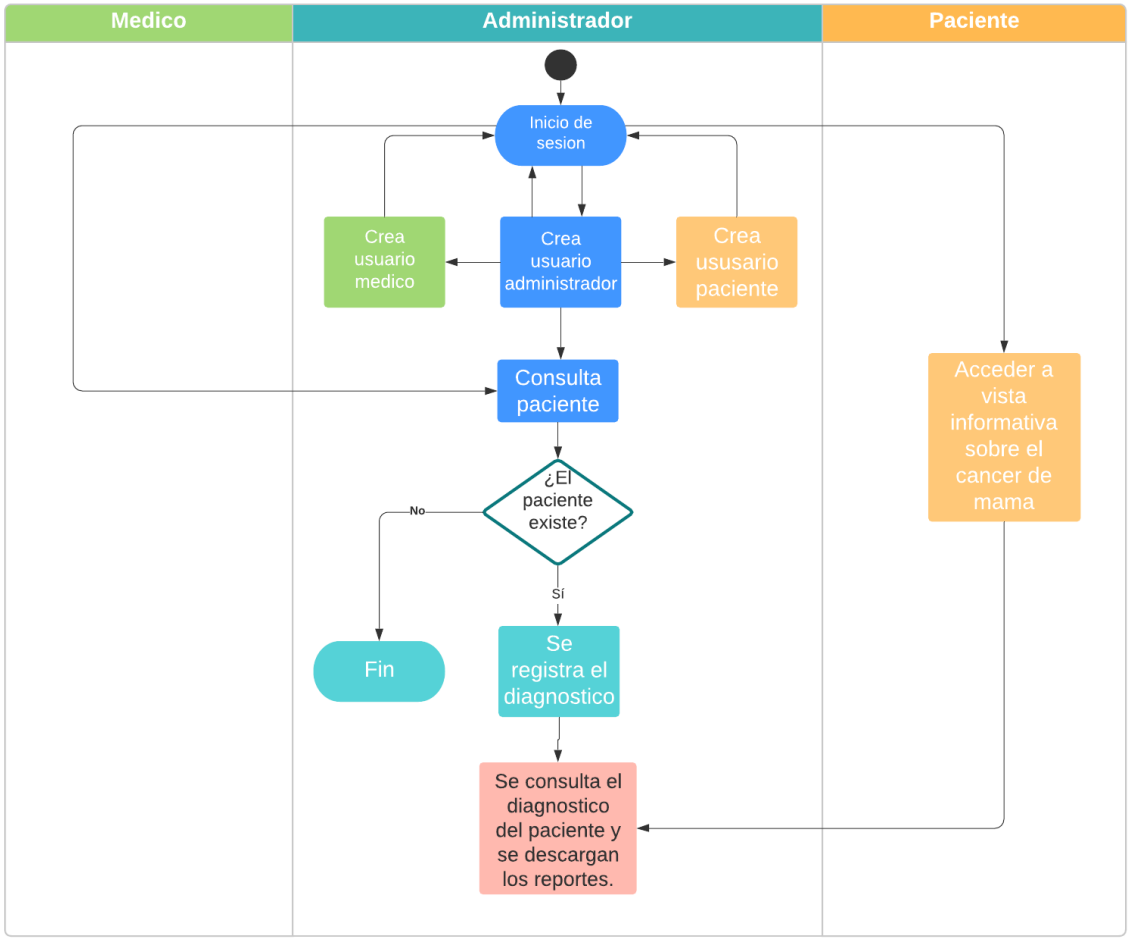
Luego de finalizar cada una de las fases y actividades de la metodología CRISP -DM para el cumplimiento del objetivo al que hace referencia el desarrollo e implementación de un aplicativo web que integre los modelos construidos este lleva por nombre “Diagnosis Breast Cancer”.

Por ello, de manera general el aplicativo web cumple con un proceso de negocio detallado de la siguiente manera:

Primeramente, el sistema contará con 3 roles, administrador, médicos y paciente de tal manera que la primera interacción con el aplicativo web comienza con el administrador. Este deberá iniciar sesión en el sistema y es el único encargado de crear a nuevos usuarios médicos y pacientes como también a nuevos administradores en caso de ser requeridos. Todo usuario nuevo deberá iniciar sesión para poder acceder al sistema, asimismo, tanto los administradores y médicos son los encargados de realizar un

diagnóstico en donde deberán en primer lugar consultar la existencia del paciente en el sistema. Una vez llevado a cabo el registro del diagnóstico será posible llevar a cabo una consulta en donde los 3 roles de usuarios podrán acceder al diagnóstico y descargar el respectivo diagnóstico. Todo esto proceso de negocio lo podemos ver expresado en el siguiente gráfico:

Gráfico 9. Diagrama del proceso de negocio



Fuente: *elaboración propia, los investigadores (2022).*

Por ende, entendiendo el proceso de negocio se llevaron a cabo las siguientes tareas y actividades:

3.2.1. ANÁLISIS DE REQUERIMIENTOS

A continuación se presentan los roles, el Product Backlog o Pila del Producto en el que se describen los requerimientos funcionales del sistema y las historias de usuarios.

- **Roles:**

Tabla 20. Roles del equipo de trabajo

Roles	Asignación
Desarrollador	Andres Gonzalez, Juan Almenares
Cliente	Liga Contra el Cancer – Seccional Cesar
Tester	Andres Gonzalez, Juan Almenares
Tracker	Alvaro Oñate Bowen

Fuente: *elaboración propia, los investigadores (2022).*

- **Product Backlog:**

Tabla 21. Product Backlog

ID	HISTORIA DE USUARIO	PRIORIDAD
1	Como administrador quiero iniciar sesión para acceder al sistema.	alta
2	Como administrador quiero registrar médicos al sistema para que estos puedan llevar a cabo gestiones de diagnósticos y consultas de pacientes.	alta
3	Como administrador quiero registrar pacientes para realizar los diagnósticos.	alta
4	Como administrador quiero registrar a otros administradores para que puedan acceder al sistema.	alta
5	Como administrador quiero realizar diagnósticos para determinar el tipo de cáncer.	alta
6	Como administrador quiero consultar a los usuarios en el sistema para realizar reportes.	alta
7	Como administrador quiero modificar los datos de los pacientes y médicos para mantenerlos actualizados.	alta
8	Como médico necesito iniciar sesión para acceder al sistema.	alta

9	Como médico quiero realizar diagnósticos para establecer el tipo de cáncer de mama.	alta
10	Como médico quiero realizar consultas para generar reportes.	alta
11	Como paciente necesita iniciar sesión para acceder al sistema.	alta
12	Como paciente quiero ver información acerca del cáncer de mama para mantenerme informado.	media
13	Como paciente necesito consultar mis diagnósticos para comprender mi situación.	alta

Fuente: *elaboración propia, los investigadores (2022).*

- **Historias de usuario:**

Tabla 22. Historia de usuario: Login administrador

Identificador: HU001	**Nombre:** Login administrador
Usuario: Administrador	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo:** 70
Descripción: Como administrador quiero iniciar sesión para acceder al sistema.	
Criterios de aceptación: ▪ Nombre de la cuenta (cadena de caracteres, obligatorio y único, de 5 a 15 caracteres). ▪ Contraseña (cadena de caracteres, obligatorio, de 5 a 15 caracteres). ▪ El aplicativo web deberá iniciar sesión. ▪ Una vez iniciada la sesión el administrador podrá cambiar su contraseña. ▪ El sistema deberá notificar a través de un Pop-up el cambio de contraseña.	

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 23. Historia de usuario: Registro de médicos

Identificador: HU002	**Nombre:** Registro de médicos
Usuario: Administrador	
Módulo: Aplicativo Web	

Prioridad: alta	**Esfuerzo:** 80
Descripción: Como administrador quiero registrar médicos al sistema para que estos puedan llevar a cabo gestiones de diagnósticos y consultas de pacientes.	
Criterios de aceptación: ▪ Número de documento (obligatorio, único). ▪ Nombres y apellidos (caracteres, de 3 a 20). ▪ Teléfono (7 a 10 dígitos). ▪ Celular (10 dígitos) ▪ Dirección (cadena de caracteres, 5 a 20 caracteres) ▪ Nombre de la cuenta (cadena de caracteres, obligatorio y único, de 5 a 15 caracteres) y contraseña determinada por defecto. ▪ El sistema deberá notificar a través de un Pop-up el registro exitoso.	

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 24. Historia de usuario: Registro de pacientes

Identificador: HU003	**Nombre**: Registro de pacientes
Usuario: Administrador	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo:** 80
Descripción: Como administrador quiero registrar pacientes para realizar los diagnósticos.	
Criterios de aceptación: ▪ Tipo de documento ▪ Número de documento (obligatorio, único). Sera usuario y contraseña. ▪ Nombres y apellidos (caracteres, de 3 a 20). ▪ Dirección (cadena de caracteres, 5 a 20 caracteres). ▪ Fecha (formato de fecha). ▪ Teléfono (7 a 10 dígitos). ▪ Celular (10 dígitos). ▪ Tipo de sangre. ▪ Departamento.	

▪ Municipio. ▪ Sexo. ▪ Familiares (opcional): datos de familiares de pacientes que tienen o alguna vez han tenido cáncer. ▪ El sistema deberá notificar a través de un Pop-up el registro exitoso.

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 25. Historia de usuario: Registro de administradores

Identificador: HU004	**Nombre**: Registro de administradores
Usuario: Administrador	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo**: 80
Descripción: Como administrador quiero registrar a otros administradores para realizar las gestiones de usuarios, reportes y diagnósticos.	
Criterios de aceptación: ▪ Nombres y apellidos (caracteres, de 3 a 20). ▪ Correo (obligatorio y único, tipo email de 11 a 30 caracteres). ▪ Nombre de la cuenta (cadena de caracteres, obligatorio y único, de 5 a 15 caracteres) y contraseña determinada por defecto.. ▪ El sistema deberá notificar a través de un Pop-up el registro exitoso.	

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 26. Historia de usuario: Diagnosticos del administrador

Identificador: HU005	**Nombre**: Diagnosticos del administrador
Usuario: Administrador	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo**: 100
Descripción: Como administrador quiero realizar diagnósticos para determinar el tipo de cáncer.	

Criterios de aceptación: ▪ El administrador podrá realizar diagnósticos buscando al paciente por su número de identificación. ▪ Los campos correspondientes a los parámetros diagnósticos deben ser numéricos. ▪ El administrador deberá ver los resultados del diagnóstico. ▪ El administrador podrá realizar observaciones y recomendaciones después de realizar el diagnóstico.

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 27. Historia de usuario: Administrador consulta usuarios

Identificador: HU006	**Nombre**: Administrador consulta usuarios
Usuario: Administrador	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo**: 90
Descripción: Como administrador quiero consultar a los usuarios en el sistema para realizar reportes.	
Criterios de aceptación: ▪ El administrador podrá realizar consultas de los pacientes a través del número de identificación y descargar sus respectivos diagnósticos. ▪ El administrador podrá generar una consulta de todos los pacientes con diagnósticos en un determinado rango de tiempo. ▪ El administrador podrá consultar a los demás administradores del sistema.	

Fuente: *elaboración propia, los investigadores (2022)*

.

Tabla 28. Historia de usuario: Actualización de datos

Identificador: HU007	**Nombre**: Actualización de datos.
Usuario: Administrador	
Módulo: Aplicativo Web	

Prioridad: alta	**Esfuerzo**: 80
Descripción: Como administrador quiero modificar los datos de los pacientes y médicos para mantenerlos actualizados.	
Criterios de aceptación: ▪ El administrador podrá actualizar los datos personales de los pacientes y médicos. ▪ El sistema deberá notificar a través de un Pop-up el proceso exitoso.	

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 29. Historia de usuario: Login médico

Identificador: HU008	**Nombre**: Login médico
Usuario: Médico	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo**: 70
Descripción: Como médico necesito iniciar sesión para acceder al sistema.	
Criterios de aceptación: ▪ Nombre de la cuenta (cadena de caracteres, obligatorio y único, de 5 a 15 caracteres). ▪ Contraseña (cadena de caracteres, obligatorio y único, de 5 a 15 caracteres). ▪ El aplicativo web deberá iniciar sesión. ▪ Una ves iniciado sesión el medico podra cambiar su contraseña. ▪ El sistema deberá notificar a través de un Pop-up el cambio de contraseña.	

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 30. Historia de usuario: Diagnósticos de médicos

Identificador: HU009	**Nombre**: Diagnósticos de médico
Usuario: Médico	

Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo**: 100
Descripción: Como administrador quiero realizar diagnósticos para determinar el tipo de cáncer.	
Criterios de aceptación: ▪ El médico podrá realizar diagnósticos buscando al paciente por su número de identificación. ▪ Los campos correspondientes a los parámetros del diagnósticos deben ser numéricos. ▪ El médico deberá ver los resultados del diagnóstico. ▪ El médico podrá realizar observaciones y recomendaciones después de realizar el diagnóstico.	

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 31. Historia de usuario: Medico consulta

Identificador: HU010	**Nombre**: Medico consulta
Usuario: Administrador	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo**: 80
Descripción: Como médico quiero realizar consultas para generar reportes.	
Criterios de aceptación: ▪ El médico podrá realizar consultas de los pacientes a través del número de identificación y descargar sus respectivos diagnósticos. ▪ El médico podrá ver los correos de los administradores con el fin de comunicarse.	

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 32. Historia de usuario: Login paciente

Identificador: HU011	**Nombre**: Login paciente
Usuario: Paciente	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo**: 70

Descripción: Como médico necesito iniciar sesión para acceder al sistema.
Criterios de aceptación: ▪ El nombre de la cuenta de los pacientes y su contraseña será su mismo número de identificación.

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 33. Historia de usuario: Portal informativo

Identificador: HU012	**Nombre**: Portal informático
Usuario: Paciente	
Módulo: Aplicativo Web	
Prioridad: media	**Esfuerzo**: 40
Descripción: Como paciente quiero ver información acerca del cáncer de mama para mantenerme informado.	
Criterios de aceptación: ▪ El paciente podrá ver información de relevancia y actual acerca del cáncer de mama.	

Fuente: *elaboración propia, los investigadores (2022).*

Tabla 34. Historia de usuario: Paciente consulta diagnóstico

Identificador: HU013	**Nombre**: Paciente consulta diagnóstico
Usuario: Paciente	
Módulo: Aplicativo Web	
Prioridad: alta	**Esfuerzo**: 80
Descripción: Como paciente necesito consultar mis diagnósticos para comprender mi situación.	
Criterios de aceptación: ▪ El paciente podrá ver su diagnóstico y descargar su respectivo reporte.	

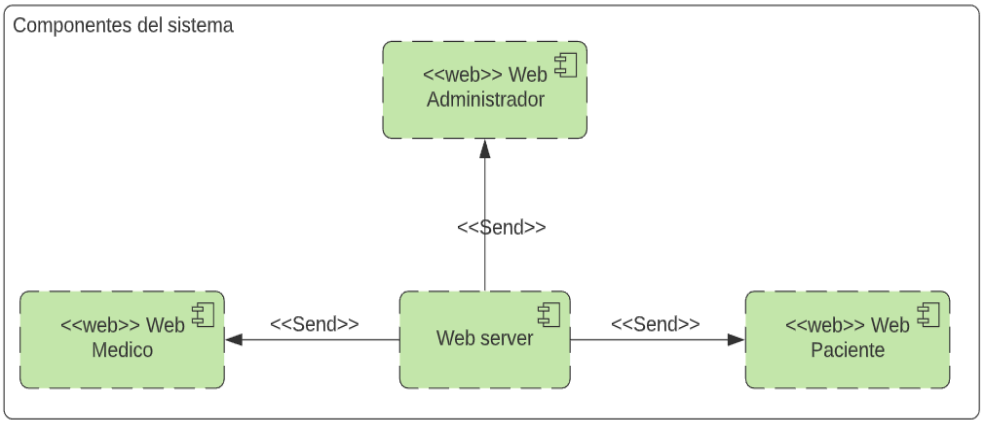
Fuente: *elaboración propia, los investigadores (2022).*

3.2.1. DISEÑO DEL SISTEMA

En este apartado se presentan diagramas de componentes, diagramas de casos de usos, diagramas de clases y el diagrama de la base de datos, los cuales permitirán entender los tipos de usuarios y sus gestiones (usuarios, diagnósticos y los reportes de diagnósticos) donde se evidencian las actividades o procesos llevados a cabo en el aplicativo web. Dichos diagramas fueron realizados mediante el uso de la herramienta web que lleva por nombre Lucidchart en su versión gratuita y que nos permite diseñar y compartir un gran número de diagramas.

Por consiguiente, en primer lugar, tenemos el diagrama de componentes con el cual podemos visualizar una estructura general y las dependencias del aplicativo web.

Gráfico 10. Diagrama de componentes



Fuente: *elaboración propia, los investigadores (2022).*

De tal manera entonces, en la siguiente tabla se muestra de manera detallada las funcionalidades de cada componente.

Tabla 35. Funcionalidades de los componentes

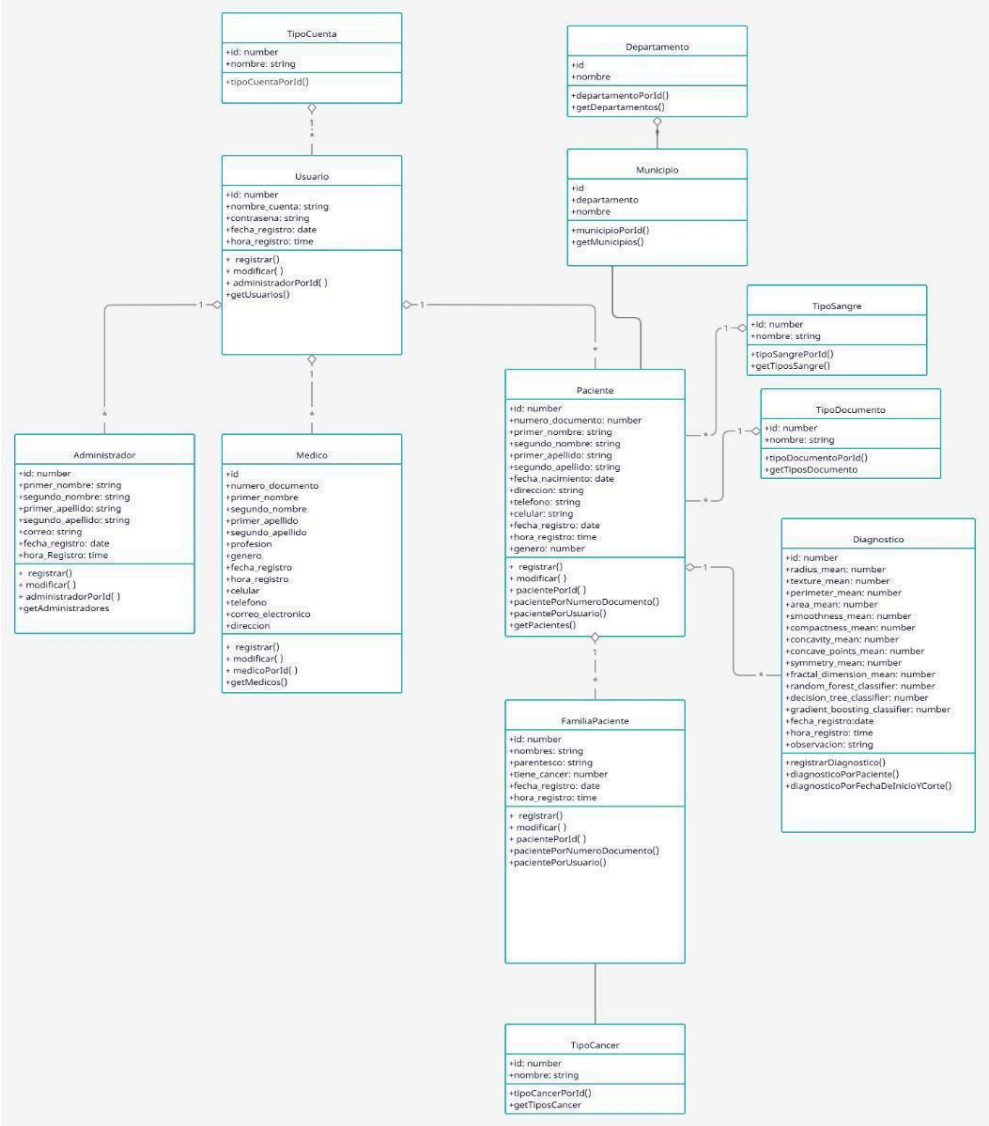
Web Server	Contiene todas las funcionalidades del sistema y proporciona el intercambio de los datos.

Web Administrador	Contiene las funcionalidades referentes a la gestión de usuarios (registros de médicos, pacientes y administradores), diagnósticos (creación de nuevos diagnósticos) y consultas o reportes.
Web Médico	En este componente se registran nuevos diagnósticos y se generan reportes por pacientes o reportes generales.
Web Paciente	En este componente se realiza una consulta sobre el diagnóstico por paciente respectivamente y una vista informativa acerca del cáncer de mama.

Fuente: *elaboración propia, los investigadores (2022).*

Seguidamente, tememos el diagrama de clases donde podemos observar el comportamiento de cada uno de los objetos del aplicativo web, sus atributos y las relaciones entre los objetos.

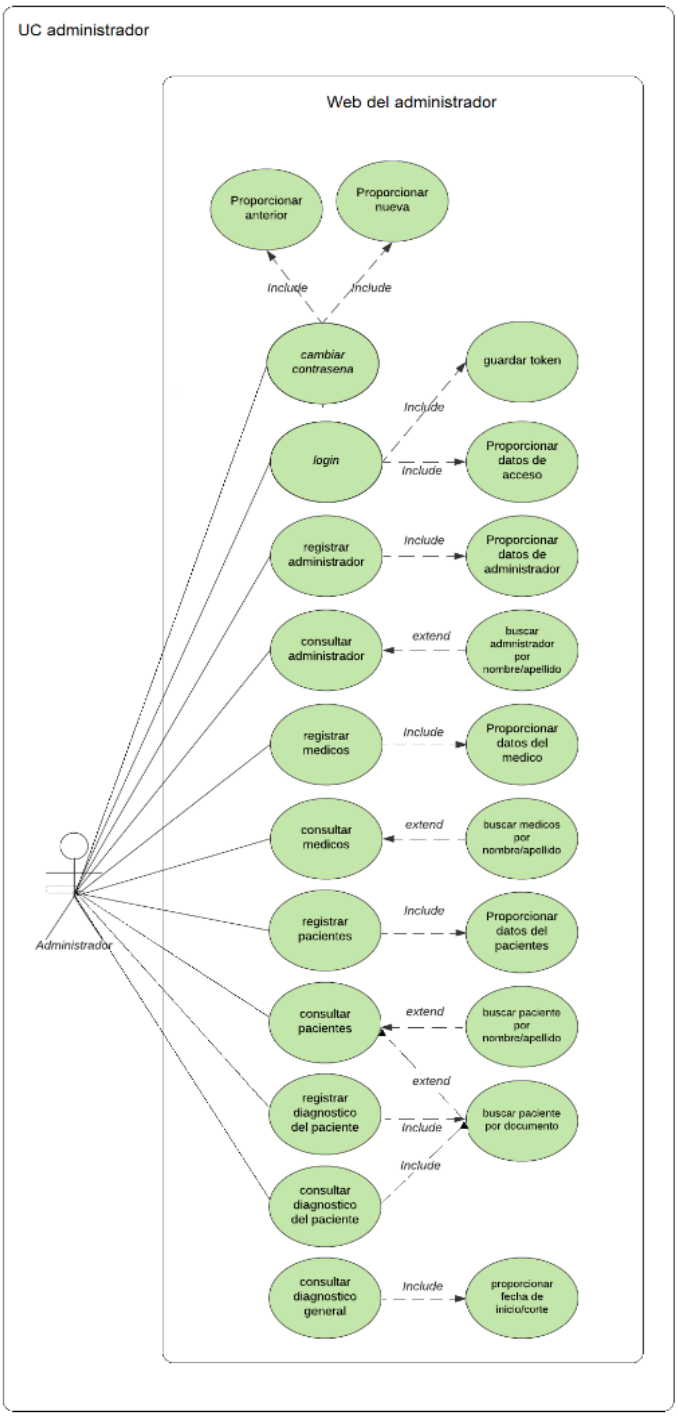
Gráfico 11. Funcionalidades de los componentes



Fuente: *elaboración propia, los investigadores (2022).*

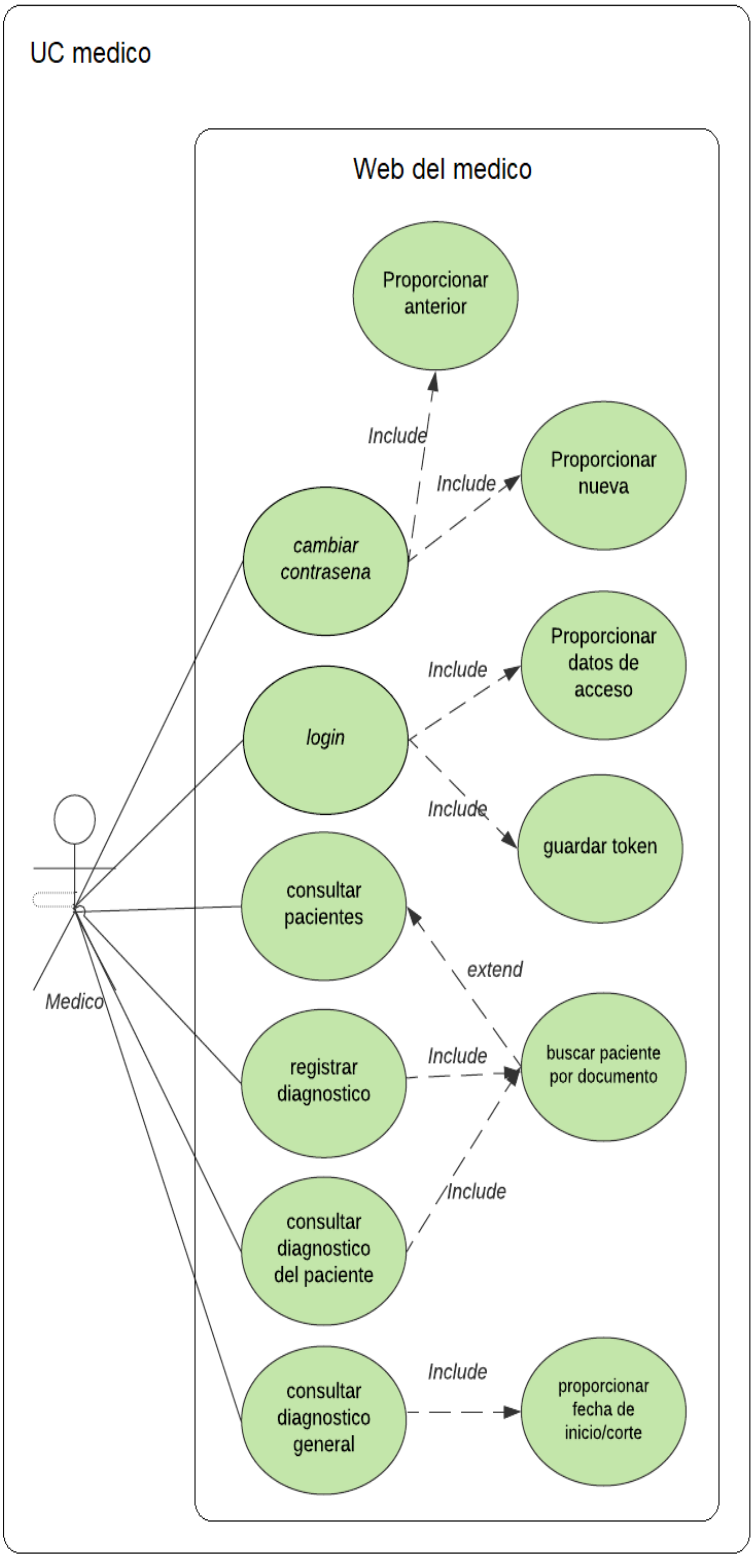
A continuación, tenemos los diagramas de casos de usos. En estos diagrama observamos cada uno de los procesos o actividades llevadas a cabo por los usuarios en sus respectivos roles dentro del aplicativo web.

Gráfico 12. Caso de uso del administrador



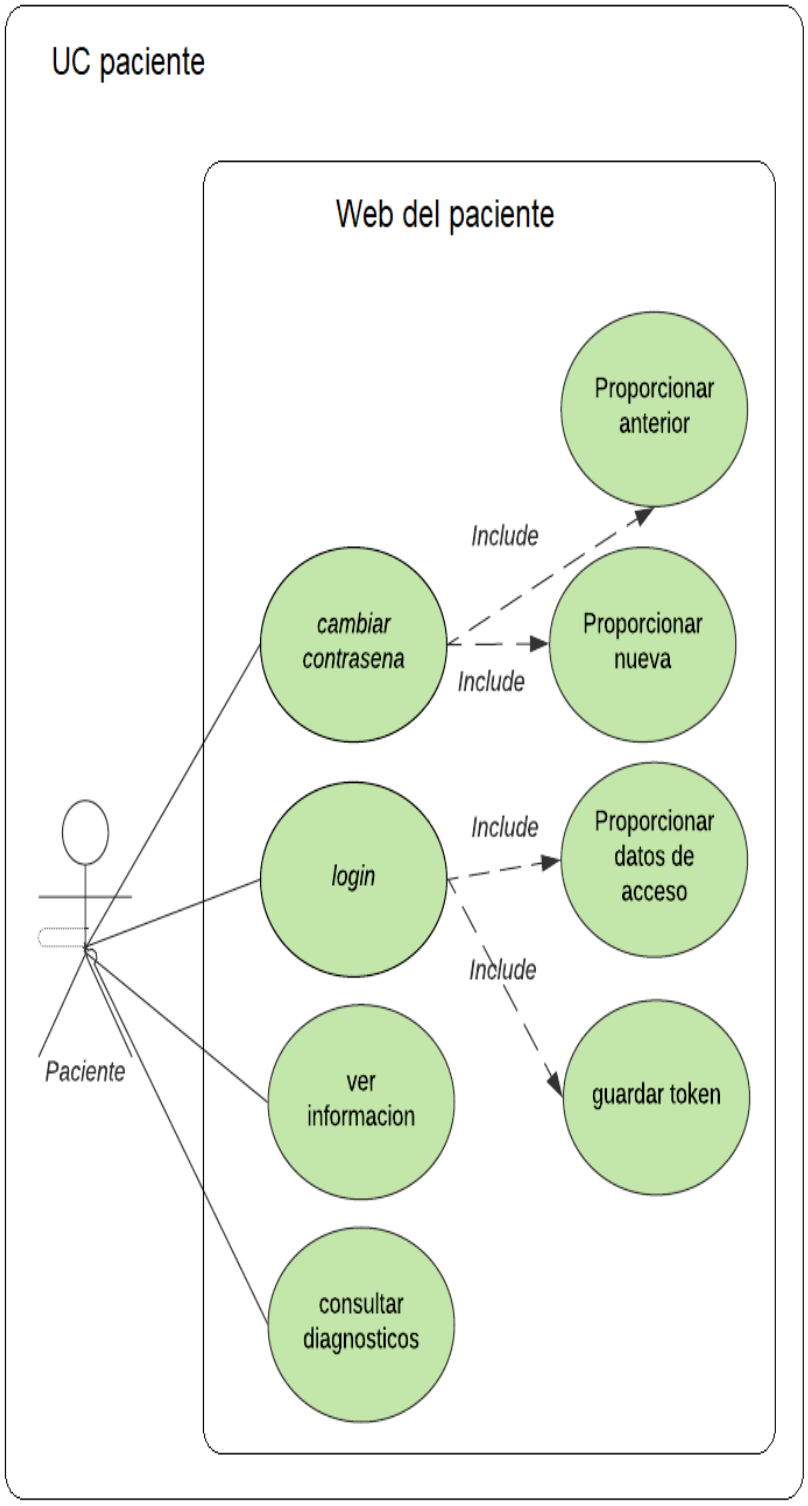
Fuente: *elaboración propia, los investigadores (2022).*

Gráfico 13. Caso de uso del medico



Fuente: *elaboración propia, los investigadores (2022)*

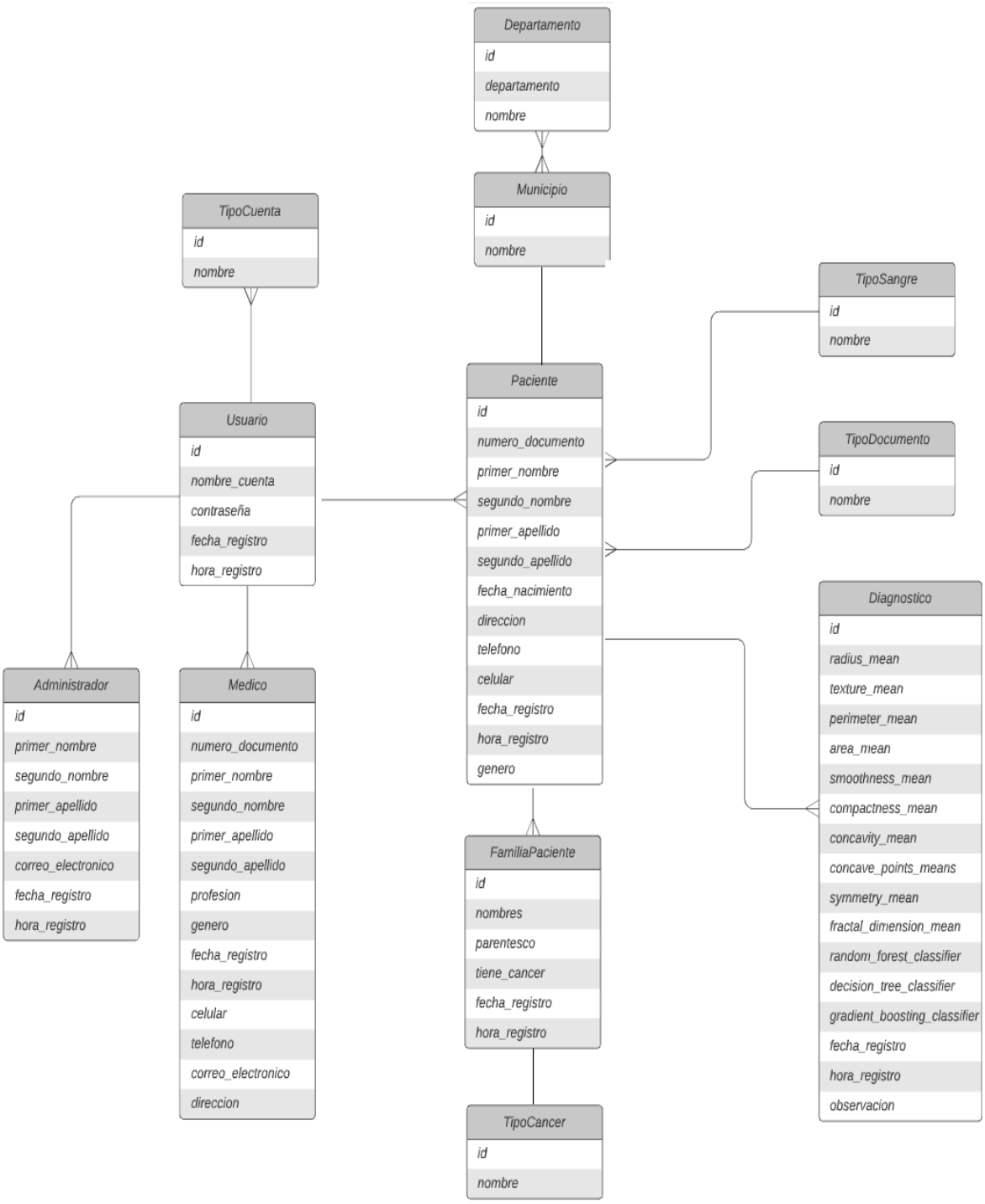
Gráfico 14. Caso de uso del paciente.



Fuente: *elaboración propia, los investigadores (2022).*

Por último tenemos, el diseño de la base de datos estructurada en donde se almacenará de manera organizada los datos que se le ingresen al aplicativo web.

Gráfico 15. Diagrama de la base de datos



Fuente: *elaboración propia, los investigadores (2022).*

3.2.2.1. Front-end

Para el desarrollo del front-end, es decir, el diseño y estructura del aplicativo web se utilizó el framework angular versión 12 con Typescript.

Ilustración 10. Vista al codigo Front-end

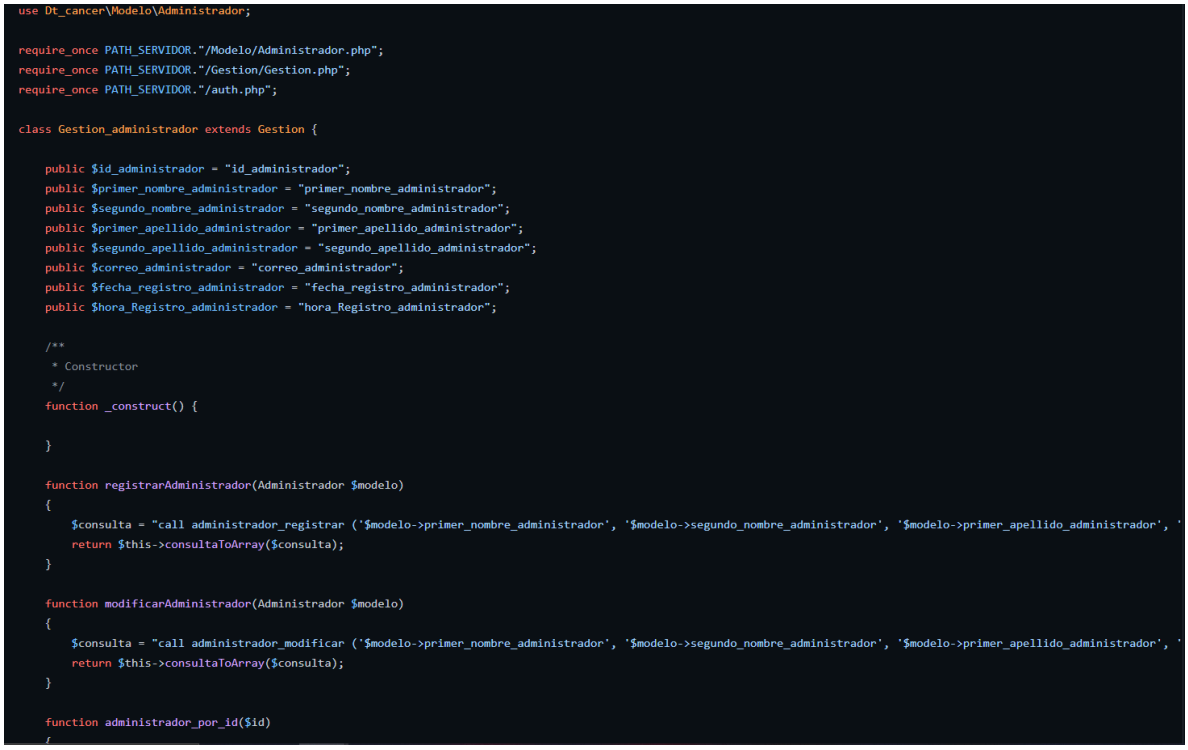
```
<!doctype html>
<html lang="en">
    <head>
        <title>Breast Cancer Diagnosis</title>
        <meta charset="utf-8">
        <meta
            name="description"
            content="Sistema de cobro">
        <meta
            name="keywords"
            content="Fuse,HTML,CSS,Angular,Angular 2,Angular 10,Angular 11,Angular 12,Material,Material 2,Angular Components,Tailwind,Tailwind CSS,TailwindCSS,Admin Template,A
        <meta
            name="viewport"
            content="width=device-width, height=device-height, initial-scale=1.0, minimum-scale=1.0">

        <base href="/">

        <!-- Favicon -->
        <link
                rel="icon"
                type="image/png"
                href="favicon-16x16.png">
        <link
                rel="icon"
                type="image/png"
                href="favicon-32x32.png">

        <!-- Fonts -->
        <link
            href="assets/fonts/inter/inter.css"
            rel="stylesheet">

        <link
            href="https://fonts.gstatic.com"
            rel="preconnect">
        <link
            href="https://fonts.googleapis.com/css2?family=IBM+Plex+Mono:ital,wght@0,400;0,500;0,600;1,400&display=swap"
```

Fuente: *elaboración propia, los investigadores (2022).*

3.2.2.2 Back-end

En cuanto al desarrollo del back-end, es decir, la parte lógica del aplicativo web se utilizó el lenguaje de programación php 7 y python para integrar los módulos de minería de datos al sistema.

Ilustración 11. Vista al codigo Back-end



```php
use Dt_cancer\Modelo\Administrador;

require_once PATH_SERVIDOR."/Modelo/Administrador.php";
require_once PATH_SERVIDOR."/Gestion/Gestion.php";
require_once PATH_SERVIDOR."/auth.php";

class Gestion_administrador extends Gestion {

    public $id_administrador = "id_administrador";
    public $primer_nombre_administrador = "primer_nombre_administrador";
    public $segundo_nombre_administrador = "segundo_nombre_administrador";
    public $primer_apellido_administrador = "primer_apellido_administrador";
    public $segundo_apellido_administrador = "segundo_apellido_administrador";
    public $correo_administrador = "correo_administrador";
    public $fecha_registro_administrador = "fecha_registro_administrador";
    public $hora_Registro_administrador = "hora_Registro_administrador";

    /**
     * Constructor
     */
    function _construct() {

    }

    function registrarAdministrador(Administrador $modelo)
    {
        $consulta = "call administrador_registrar ('$modelo->primer_nombre_administrador', '$modelo->segundo_nombre_administrador', '$modelo->primer_apellido_administrador', '
        return $this->consultaToArray($consulta);
    }

    function modificarAdministrador(Administrador $modelo)
    {
        $consulta = "call administrador_modificar ('$modelo->primer_nombre_administrador', '$modelo->segundo_nombre_administrador', '$modelo->primer_apellido_administrador', '
        return $this->consultaToArray($consulta);
    }

    function administrador_por_id($id)
```

Fuente: *elaboración propia, los investigadores (2022).*

3.2.3. SOCIALIZACIÓN

Ilustración 12. Socialización del proyecto

Fuente: *elaboración propia, los investigadores (2022).*

La socialización del proyecto como se ve en la fotografía anterior, se lleva a cabo entre los responsables del proyecto Andres Gonzalez y Juan Almenares con la directora ejecutiva y de calidad Johana Villero de la Liga Contra el Cáncer – Seccional Cesar.

CONCLUSIONES Y RECOMENDACIONES

Con la realización del proyecto de grado y mediante las bibliografías consultadas y referenciadas en este documento, se determinó que la metodología CRISP - DM resulta muy efectiva para que de manera organizada y estructurada se siguieran una serie de actividades específicas que permitieron planificar, gestionar y ejecutar el proyecto conforme a los objetivos propuestos. Por lo tanto se logró:

En primer lugar, se analizaron las problemáticas y situación actual de la Liga Contra el Cancer - Seccional Cesar y se obtuvo un panorama general de las necesidades de los pacientes y los profesionales de la salud en la actual pandemia, generando de esta manera una mayor comprensión en las actividades llevadas a cabo por esta entidad. Por otro lado, a partir de la exploración del conjunto de datos se obtuvieron las primeras hipótesis con la finalidad de obtener una visión generalizada del conjunto de datos.

En segundo lugar, se lograron crear modelos que permitieron diagnosticar el tipo de cáncer de mama caracterizándose o clasificándolos en benignos y malignos, con un alto grado de exactitud y precisión. De manera general, se construyeron y se evaluaron 7 modelos (Random Forest, KNN, SVM, Naive Bayes, Regression Logistic, Decision Tree y Gradient Boosting), de los cuales, los modelos Decision Tree y Gradient Boosting se obtuvieron un 100% de exactitud y el modelo Random Forest un 99%.

En tercer lugar, se desarrolló un aplicativo web que permite llevar a cabo una gestión de diagnósticos del cáncer de mama integrando los 3 modelos antes mencionados y seleccionados por sus altos porcentajes. Este aplicativo web denominado “Diagnosis Breast Cancer” fue montado en un servidor virtual privado o VPS llamado Blue Hosting y se puede acceder por medio de la dirección http://168.232.165.30/.

En cuanto a los datos utilizados para la construcción de los modelos es importante recalcar que estos datos fueron el resultado de biopsias por aspiración de aguja fina, método que a través de la extracción y su posterior digitalización en imágenes de tejidos mamarios se obtuvieron las variables que conforman el conjunto de datos aplicado, pero como se demuestra en este proyecto y en las bibliografías consultadas, estos modelos pueden clasificar distintos tipos de cáncer en base a otros conjuntos de datos que sean procesados por técnicas diferentes.

Dicho lo anterior, como trabajos futuros este proyecto puede ser complementado en asociación con un centro oncológico, desarrollando un sistema similar por el dr. Wolberg creador del conjunto de datos aplicado en el desarrollo del proyecto a través de la técnica de biopsia por aspiración con aguja fina, para la construcción de un conjunto de datos con pacientes propios de la región; de la misma manera, al aplicativo web es posible integrar otros módulos que permitan diagnosticar diversos tipos de cáncer, por lo que este proyecto puede considerarse como el punto de partida para la creación de un sistema más robusto para el beneficio de la población que permita apoyar y mejorar los procesos de diagnósticos y detección temprana del cáncer u otras enfermedades.

Por otro lado, con la realización del proyecto, se sugieren las siguientes recomendaciones:

Teniendo en cuenta el desarrollo de las fases de la metodología CRISP - DM:

- Analizar previamente un conjunto de datos que permita extraer información relevante antes de desarrollar la investigación para no perder tiempo y recursos, ya que esta metodología implícitamente asume que los datos disponibles son de óptima calidad.
- Tener siempre en cuenta a la bibliografía y antecedentes de la investigación para escoger los modelos apropiados.

Teniendo en cuenta el desarrollo del software:

- Delegar responsabilidades para el desarrollo del frontend y el backend entre los miembros del equipo de trabajo.

Teniendo en cuenta el uso del software:

- Capacitar al personal que utilizará el aplicativo web.
- Los administradores del aplicativo web deben ser personal con conocimientos en el análisis de datos e ingenieros de sistemas, esto para dar soportes al sistema en caso de que sean requeridos.

Por último, para concluir, mediante el uso de las tecnologías de la información se puede mejorar y reducir los tiempos del diagnóstico del cáncer de mama y este tipo de proyecto añade gran valor a los centros oncológicos, sobre todo a la Liga contra el cancer -

Seccional Cesar, brindando de herramientas que permitan tomar decisiones de manera temprana y llevando a la mejora continua de la organización en su misión ante la población y en su que hacer ser.

REFERENCIAS

[1] G. Yu, «Aumentan casos de cáncer a nivel mundial (pero en algunos países el incremento es mayor), según la OMS,» CNN, 17 febrero 2020.

[2] E. R. Rojas, «Mujer: Ataque el cáncer de mama a tiempo ", La Patria, 2 noviembre 2020.

[3] L. T. Borja, «Cáncer: viviendo la batalla rosa,» Las 2 Orillas, 06 noviembre 2020.

[4] «Ministerio de Salud y Protección Social,» 19 octubre 2020. [En línea]. Available: https://www.minsalud.gov.co/Paginas/Detecte-el-cancer-de-mama-a-tiempo.aspx.

[5] S. Mitra y T. Acharya., Data mining: multimedia, soft computing and bioinformatics, Wiley - Interscience, 2003.

[6] j. Hernández Orallo, M. Ramírez Quintana y C. and Ferri Ramírez, Introducción a la Minería de Datos, Madrid: Pearson Education S.A., 2004.

[7] B. A. Akinnuwesi, B. O.Macaulay y B. S. Aribisala, «Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques,» ELSEVIER, vol. 21, 2020.

[8] Syed Jamal Safdar Gardezi, Ahmed Elazab, Baiying Lei, «Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review,» JMIR Publications, vol. 21, nº 7, 23 Abril 2019.

[9] Laboratorio-SI2M, «Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis» ELSEVIER, vol. 127, pp. 293-299, 2018.

[10] D. Devakumari y V. Punithavathi, «Study of Breast Cancer Detection Methods using Image Processing with Data Mining Techniques,» International Journal of Pure and Applied Mathematics, vol. 118, nº 18, pp. 2867-2873, 2018.

[11] E. H. Martínez y R. L. Sanjurjo, «Minera de datos aplicada a la detección de Cáncer de mama,» Universidad Carlos III de Madrid , Madrid.

[12] F. Vázquez y H. M. C. &. P.E., «ESTUDIO DE HERRAMIENTAS DE MINERÍA DE DATOS PARA LA TAREA DE CLASIFICACIÓN,» Tecnológico Nacional de México-Instituto Tecnológico de Cd. Victoria - División de Estudios de Posgrado e Investigación, vol. 14, 2017.

[13] S. K. Mandal, «Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using NaïveBayes, Logistic Regression and Decision Tree,» International Journal Of Engineering And Computer Science, vol. 6, pp. 20388-20391, 2 Febrero 2017.

[14] H. Asria, H. Mousannif, H. A. Moatassime y T. Noel, «Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis ", Procedia Computer Science, vol. 83, pp. 1064 -1069, 2016.

[15] M. Mauricio, B. Dewar Rico y E. Romero Riaño, Análisis comparativo de predicción dentro de bases de datos de cáncer: una aplicación de aprendizaje, Risti, nº E17, pp. 113 - 122, 2019.

[16] R. Timarán Pereira y M. C. Yépez Chamorro, Caracterización de la supervivencia de mujeres con cáncer invasivo de cuello uterino usando minería de datos, REVISTA DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN, RIDI, vol. 7, nº 1, pp. 127-139, Diciembre 2016.

[17] C. A. Madrigal González, R. Prada Vásquez y D. S. Fernández McCann, «Detección Automática de Microcalcificaciones en una Mamografía Digital, Usando Técnicas De Inteligencia Artificial,» TecnoLógicas, pp. 743 - 756, Octubre 2013.

[18] R. C. Morales Ortega, G. Lozan Bernal, P. P. Ariza Colpas, E. Arrieta Rodriguez, E. C. Ospino Mendoza, J. Caicedo Ortiz, M. A. Piñeres Melo, F. E. Mendoza Palechor y M. Roca Vides, «Method Based on Data Mining Techniques for Breast Cancer Recurrence Analysis ", Springer, vol. 12145, 13 Julio 2020.

[19] Rangayyan, R. M., Ayres, F. J., y Desautels, J. L. "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," Journal of the Franklin Institute, vol. 344, no. 3, pp. 312-348. 2017.

[20] Cheng, H.-D., Cai, X., Chen, X., Hu, L. y Lou, X. "Computer-aided detection and classification of microcalcifications in mammograms: a survey," Pattern recognition, vol. 36, no. 12, pp. 2967-2991. 2013.

[21] García, S. M. Radiología básica - Aspectos fundamentales (Segunda Edición). En W. Herring, Learning radiology: recognizing the basics (págs. 1-7). Philadelphia, Pennsylvania: ELSEVIER. 2012.

[22] Amparo Vilarrasa Andrés. Sistema inteligente para la detección y diagnóstico de patología mamaria. En Amparo Vilarrasa Andrés, sistema inteligente para la detección y diagnóstico de patología mamaria. Madrid, pp. 844 -852, 2006.

[23] Andrew, N. Machine Learning, in Machine Learning, C. Stanford University, Ed., ed: Coursera. 2016.

[24] Leonard, J., Colombe, J., & Levy, J. Finding relevant references to genes and proteins in Medline using a Bayesian approach. Bioinformatics, 18(11), 1515–1522, 2002.

[25] Omar Danilo Castrillón, W. A. Diseño de una hiper heurística para la programación de la producción en ambientes JOB SHOP. Ingeniare. Revista chilena de ingeniería, 203-214. 2010.

[26] Equipo-de-redactores-y-equipo-de-editores-médicos-de-la-Sociedad-Americana-Contra-El-Cáncer, «American Cancer Society,» 3 Octubre 2019. [En línea]. Available: https://www.cancer.org/es/cancer/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/biopsia-del-seno/biopsia-del-seno-por-aspiracion-con-aguja-fina.html

[27] Vallejo Delgado N, R. J. Aplicación de técnicas de minería de datos para el diagnóstico prematuro de cáncer. 2012.

[28] Morales, E. Descubrimiento de Conocimiento en Bases de Datos. 2013.

[29] Orallo, J., Ramírez, M., & Ferri, C. Introducción a la minería de datos (p. 680). Pearson Educación. Madrid, pp. 680, 2004

[30] Fayyad, R., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. Advances in knowledge and data mining. Massachusetts: AAAI/MIT Press. 1996

[31] Molina, L. Data mining no processo de extração de conhecimento de bases de Tesis de máster. São Carlos (Brasil): Instituto de Ciências Matemáticas e Computação. Universidade de São Paulo. 1998.

[32] A. Azevedo y M. F. Santos, «KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW,» IDAIS, 2008.

[33] Sas. Institute. Inc., «SAS® Enterprise Miner™,» [En línea]. Available: https://www.sas.com/en_us/software/enterprise-miner.html. [Último acceso: 9 10 2021].

[34] IBM, *Manual CRISP-DM de IBM SPSS,* © Copyright IBM Corporation 1994, 2012..

[35] J. A. Gallardo Arancibia, «Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM».

[36] Chen. C. K., Analysis of Branch Prediction Via Data Compression (págs. 128-137). Cambridge, Massachusetts: ASPLOS VII. 1996.

[37] Quinlan, J. R. Induction of decision trees. Machine learning, 81–106, 2006.

[38] Han, J. Data Mining. Concepts and Techniques. 548. Editorial Morgan Kaufmann. 2011.

[39] Kamber, J. H. Data Mining: Concepts and Techniques. Amsterdam Boston. 2006.

[40] Hand, D. J. Data Mining: Statistics and More. The American Statistician. 2007.

[41] Burroni, M. Melanoma Computer-Aided Diagnosis: Reliability and Feasibility Study. Clinical Cancer Research, 10(6), 1881–1886, 2004.

[42] Bertani, A., Cappello, A., Benedetti, M. G., Simoncini, L., & Catani, F. Flat foot functional evaluation using pattern recognition of ground reaction data. Clinical biomechanics, 14(7), 484–93, 1999.

[43] Gonzalez Avila, M. Aspectos éticos de la investigación cualitativa. Revista Iberoamericana de Educación - Número 29. 2002

[44] Hernández, S., Fernández, C., & Baptista, P. Metodología de la Investigación. México: Ed. Mc. Graw Hill. 2015.

[45] «www.ligacancercesar.org,» Liga contra el Cáncer-Seccional Cesar, [En línea]. Available: https://www.ligacancercesar.org/. [Último acceso: 20 10 2021].

[46] N. Baute Barrios, «El cáncer es la tercera causa de muerte en el Cesar,» EL PILÓN, 4 02 2021.

[47] «Boletín No. 02 Día mundial del cáncer,» SISPRO, 2021. [En línea]. Available: https://www.sispro.gov.co/observatorios/oncancer/Paginas/onc_boletin_02_cancer.aspx.

[48] JOHANNA-P-VILLERO-ALARCÓN, «INFORME DE GESTION 2020,» Liga Contra el Cáncer - Seccional Cesar, Valledupar, 2020.

[49] «INFORME DE GESTIÓN ACTIVIDADES RELEVANTES AÑO 2020,» LIGA COLOMBIANA CONTRA EL CÁNCER, 2020.

[50] O. L. Mangasarian y W. H. Wolberg, «Machine Learning for Cancer Diagnosis and Prognosis,» University of Wisconsin-Madison, 1990.

[51] Joaquín-Amat-Rodrigo, «Ciencia de Datos, Estadística, Machine Learning y Programación,» Octubre 2020. [En línea]. Available: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html.

[52] Joaquín-Amat-Rodrigo, «Ciencia de Datos, Estadística, Machine Learning y Programación,» Agosto 2016. [En línea]. Available: https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple.

[49] O. L. Mangasarian y W. H. Wolberg, «Machine Learning for Cancer Diagnosis and Prognosis,» University of Wisconsin-Madison, 1990.

[50] Joaquín-Amat-Rodrigo, «Ciencia de Datos, Estadística, Machine Learning y Programación,» Octubre 2020. [En línea]. Available: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html.

[51] Joaquín-Amat-Rodrigo, «Ciencia de Datos, Estadística, Machine Learning y Programación,» Agosto 2016. [En línea]. Available: https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple.

ANEXOS

Anexo A. *Carta del director del proyecto*

Valledupar, 04 de Agosto del 2021

Señores:

COMITÉ DE PROYECTOS DE GRADO

Facultad de Ingenierías y Tecnológicas

Programa Ingeniería de Sistemas

Universidad Popular Del Cesar

Cordial saludo

Yo **ÁLVARO OÑATE BOWEN**, identificado con la cédula de ciudadanía No. **77.170.220**, certifico que he revisado el documento correspondiente al proyecto que lleva por título **"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO DEL CÁNCER DE MAMA"**, presentado por los estudiantes **ANDRÉS CAMILO GONZÁLEZ OTERO** y **JUAN FRANCISCO ALMENARES ARAGÓN**, y, después de haberle realizado las respectivas correcciones, cuenta con mi aprobación para ser presentada ante el comité. Sugiero la aprobación por parte de ustedes.

LÍNEA DE INVESTIGACIÓN: Transformación Digital

SUBLÍNEA DE INVESTIGACIÓN: Big Data y Analytics

AREA: Data Mining

GRUPO: GISICO

Agradezco la atención prestada

Atentamente;

ÁLVARO OÑATE BOWEN

CC 77.170.220 de Valledupar

Docente Programa de Ingeniería de Sistemas

Anexo B. *Carta de los estudiantes*

Valledupar, 04 de Agosto del 2021

Señores:

COMITÉ DE PROYECTOS DE GRADO

Facultad de Ingenierías y Tecnológicas

Programa Ingeniería de Sistemas

Universidad Popular del Cesar

Cordial saludo,

Nosotros los abajo firmantes, estudiantes del programa de Ingeniería de sistemas, presentamos a ustedes el documento correspondiente al proyecto de grado denominado **"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO DEL CÁNCER DE MAMA"**.

Quedamos a la espera del concepto emitido por el comité respecto de la viabilidad y aceptación de dicha propuesta.

Agradecemos la atención prestada

Atentamente,

Andrés Camilo González Otero
CC. 1.122.414.594 de San Juan del Cesar

Juan Francisco Almenares Aragón
CC. 1.003.233.568 de Valledupar

Anexo C. *Evidencias de asesoría metodológica*

NOMBRE DEL PROYECTO: “APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO DEL CÁNCER DE MAMA”.

AUTORES: Andrés Camilo González Otero - Juan Francisco Almenares Aragón.

Director: Alvaro Oñate Bowen.

FECHA	REVISIÓN REALIZADA	FIRMA ASESOR
13-07-2021	**Revision de la propuesta**	
08-09-2021	**Revision del anteproyecto**	
04-02-2022	**Revision del proyecto final**	

Anexo D. *Carta de aval de entidad responsable*

LIGA CONTRA EL CÁNCER
Seccional - Cesar
NIT 892 300 937-1

Valledupar, 12 de mayo del 2021

L.C.C. 070 -2021

Señores: **COMITÉ DE PROYECTOS DE GRADO**
Facultad de Ingenierías y Tecnológicas
Programa de Ingeniería de Sistemas.
Universidad Popular Del Cesar.

Cordial Saludo respetados Ingenieros;

Me permito informarle que los estudiantes: **Andrés Camilo González Otero** y **Juan Francisco Almenares Aragón**, se encuentran autorizados por esta entidad para realizar su proyecto de grado, titulado "**APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO TEMPRANO DEL CÁNCER DE MAMA EN LA LIGA CONTRA EL CÁNCER SECCIONAL CESAR.**"

Aclaro de antemano que el desarrollo del proyecto no genera ningún vínculo laboral con la entidad.

Atentamente,

MARIA VILMA GARCIA DE SOTO
Presidenta

LUIS ENRIQUE MONTERO ARIAS
Archivo de Historias Clínicas

JOHANNA P.VILLERO ALARCON
Directora Ejecutiva y de Calidad

Personería Jurídica N.º 02025 de marzo 11/ 81 – Cra. 19 n.º 16–191. Teléfono 5713567 Fax 5703200 –Cel.: 3106411437
Afiliada a la Liga Colombiana Contra el Cáncer
www.ligacancercesar.org e-mail:ligacancer@hotmail.com
Valledupar - Cesar

Anexo E. *Carta declaración antifraude*

UNIVERSIDAD POPULAR DEL CESAR
FACULTAD DE INGENIERIAS Y TECNOLOGIAS
PRESENTACIÓN DE PROPUESTA DE PROYECTO O TESIS DE GRADO HOJA DE DECLARACION ANTI FRAUDE

SEMESTRE	**10° Semestre**
FECHA (aaaa/mm/dd)	**2021/08/04**

PROYECTO DE GRADO PARA OPTAR EL TÍTULO DE: INGENIEROS DE SISTEMAS

NOMBRES Y APELLIDOS DEL ESTUDIANTE:
ANDRÉS CAMILO GONZÁLEZ OTERO - JUAN FRANCISCO ALMENARES ARAGÓN
CÓDIGO:
1.122.414.594 --- 1.003.233.568

TÍTULO DE LA TESIS O PROYECTO:
"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO DEL CÁNCER DE MAMA".

DECLARACIÓN:
1 - Soy consciente que cualquier tipo de fraude en este proyecto es considerado como una falta grave en la Universidad. Al firmar, entregar y presentar esta propuesta de Proyecto de Grado, doy expreso testimonio de que esta propuesta fue desarrollada de acuerdo con las normas establecidas por la Universidad. Del mismo modo, aseguro que no participé en ningún tipo de fraude y que en el trabajo se expresan debidamente los conceptos o ideas que son tomadas de otras fuentes.

2- Soy consciente de que el trabajo que realizaré incluirá ideas y conceptos del autor y del director y/o Asesor y podrá incluir material de cursos o trabajos anteriores realizados en la Universidad y, por lo tanto, daré el crédito correspondiente y utilizaré este material de acuerdo con las normas de derechos de autor. Así mismo, no haré publicaciones, informes, artículos o presentaciones en congresos, seminarios o conferencias sin la revisión o autorización expresa del Asesor, quien representará en este caso a la Universidad.

NOMBRES Y APELLIDOS	NOMBRES Y APELLIDOS
ANDRÉS CAMILO GONZÁLEZ OTERO	JUAN FRANCISCO ALMENARES ARAGÓN
CC: 1.122.414.594	**CC:** 1.003.233.568

Anexo F. *Carta Derechos de autor*

UNIVERSIDAD POPULAR DEL CESAR
FACULTAD DE INGENIERIAS Y TECNOLOGIAS
PRESENTACIÓN DE PROPUESTA DE PROYECTO O TESIS DE GRADO HOJA DE DERECHOS DE AUTOR

SEMESTRE	**10° Semestre**
FECHA (aaaa/mm/dd)	**FECHA (2021/08/04)**

PROYECTO DE GRADO PARA OPTAR EL TÍTULO DE: INGENIERO DE SISTEMAS

NOMBRES Y APELLIDOS DEL ESTUDIANTE:
ANDRÉS CAMILO GONZÁLEZ OTERO - JUAN FRANCISCO ALMENARES ARAGÓN

CÓDIGO:
1.122.414.594 --- 1.003.233.568

TÍTULO DE LA TESIS O PROYECTO:
"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO DEL CÁNCER DE MAMA".

AUTORIZACIÓN DE SU USO A FAVOR DE LA UNIVERSIDAD:

Autorizo a LA UNIVERSIDAD POPULAR DEL CESAR, para que en los términos establecidos en la Ley 23 de 1982, Ley 44 de 1993, Decisión Andina 351 de 1993, Decreto 460 de 1995 y demás normas generales sobre la materia, utilice y use en todas sus formas, los derechos patrimoniales de reproducción, comunicación pública, transformación y distribución (alquiler, préstamo público e importación) que me corresponden como creador de la obra objeto del presente documento. PARÁGRAFO: La presente autorización se hace extensiva no sólo a las facultades y derechos de uso sobre la obra en formato o soporte material, sino también para formato virtual, electrónico, digital, óptico, usos en red, internet, extranet, intranet, etc., y en general para cualquier formato conocido o por conocer.

NOMBRES Y APELLIDOS	NOMBRES Y APELLIDOS
ANDRÉS CAMILO GONZÁLEZ OTERO	JUAN FRANCISCO ALMENARES ARAGÓN
CC: 1.122.414.594	**CC:** 1.003.233.568

Anexo G. *Carta de compromiso de realizar un artículo científico.*

Valledupar, 04 de Agosto del 2021

Señores:
COMITÉ DE PROYECTOS
Facultad de Ingenierías y Tecnologías
Programa de ingeniería de Sistemas
Universidad Popular Del Cesar

Cordial saludo,

Quiénes suscriben la presente carta se comprometen a desarrollar un artículo científico del presente proyecto de grado titulado: **"APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y EL DIAGNÓSTICO DEL CÁNCER DE MAMA",** la evidencia de la presentación del artículo para revisión a una revista será entregada en el documento final como Anexo S.

Agradecemos la atención prestada

Atentamente,

Andrés González

Juan Almenares

Andrés Camilo González Otero
CC. 1.122.414.594 de San Juan del Cesar

Juan Francisco Almenares Aragón
CC. 1.003.233.568 de Valledupar

Anexo H. *Carta de declaración de la Universidad*

Valledupar, 04 de Agosto del 2021

“La Universidad no se hace responsable de los conceptos emitidos por los estudiantes en su proyecto de grado, solo velará que no se publique nada contrario a la moral y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la justicia”

Agradecemos la atención prestada,

Atentamente,

Juan Almenares

Andrés González

Andrés Camilo González Otero	**Juan Francisco Almenares Aragón**
CC. 1.122.414.594 de San Juan del Cesar	**CC. 1.003.233.568 de Valledupar**

Anexo I. *Evidencias de la recolección de la información*

Valledupar, 16 de junio del 2021

La siguiente entrevista será realizada con el propósito de recopilar evidencias que sustenten la problemática que se planteará en la propuesta de proyecto de grado.

Entrevista realizada al ingeniero de sistemas y a la directora ejecutiva y de calidad de la Liga contra el cáncer – Seccional Cesar.

¿Qué problemáticas cree usted que se vienen presentando en la liga contra el cáncer?

Pregunta muy general porque podría abordar nuestras problemáticas desde diferentes aspectos, se más especifico.

Pero en particular, la problemática que hoy nosotros presentamos y creemos que otras instituciones o centros de salud radica en la actual pandemia del coronavirus debido a que los procesos son más lentos.

¿Ha habido un aumento significativo en los últimos años con respecto al cáncer de mama?

Si hay aumento porque lastimosamente los pacientes acuden a las consultas cuando la enfermedad ha avanzado ya sea por el desconocimiento que tienen las personas sobre esta enfermedad como los primeros síntomas o los primeros chequeos que la misma persona se puede realizar e incluso los hombres debido a que desconocen que también pueden sufrir de cáncer de mama, como también de la falta de conciencia y falta de cuidado en la salud propia.

¿Cómo se elabora el proceso para llevar a cabo una consulta por parte del usuario?

El usuario llama, solicita agenda, se le asigna cita de acuerdo con disponibilidad, cancela el valor de la cita, el usuario es atendido.

¿Con qué recursos cuentan para llevar a cabo una consulta o un examen relacionado con el cáncer de mama? (recursos humanos, físicos, software, entre otros).

Recurso humano médico cirugía general, Consulta de cirugía general, Software de historias clínicas, Examen físico por parte del médico, Exámenes apoyo diagnostico

¿Cuánto tiempo tardan en realizar un examen?

Lo que el médico indique, ya el paciente donde se lo realiza si es con su eps, o se los realiza con nuestros ips adscritos ya sea centro radiológicos o laboratorios.

¿Quiénes manejan o analizan la información?

El médico evalúa y direcciona al paciente.

La información de historias clínicas sirve para estadísticas.

¿Qué procesos se realizan con los datos obtenidos?

Reportes estadísticos a entes de control.

¿Utilizan algún software que permita realizar un diagnóstico?

No

¿Considera usted de utilidad un sistema de apoyo que permita a los especialistas obtener un primer diagnóstico?

Si claro

¿Considera usted de utilidad un aplicativo web que permita al usuario ingresar datos para obtener un primer diagnóstico?

Pues sería útil

¿Considera usted que es de beneficio para la liga contra el cáncer contar con desarrollo e implementación de este tipo de proyectos?

Si claro

¿Se ha implementado anteriormente un sistema parecido?

No

¿Cree usted que este tipo de proyectos es de utilidad para prevenir diferentes tipos de cáncer?

Si claro, es útil

¿Qué resultados se esperan obtener con la realización del proyecto?

Pues esperaría que me den más información de este proyecto.