Universidad Cenfotec

Maestría en Tecnologías de Bases de Datos

Documento final de Proyecto de Investigación Aplicada 2

Tema:

Desarrollo de una solución de asistencia médica computarizada para detección de cáncer de pulmón mediante imágenes de tomografía computarizada utilizando herramientas de Big Data

Estudiante:

Wu Feng, Greivin

Mayo, 2018

Declaratoria de derecho de autor

Declaro que el presente proyecto de investigación fue realizado en su totalidad por el autor Greivin Wu Feng, con base en indagación en Internet, literatura referente al tema y los conocimientos adquiridos de experiencias previas de proyectos similares al área de inteligencia de negocios, Big Data, aprendizaje de máquinas, minería de datos y un poco de ciencia de datos.

En caso de que se haya realizado referencias a definiciones específicas de diversos autores, se ha procedido a indicar las debidas referencias, a fin de no violentar los derechos de autor.

Dedicatoria

Mi tesis la dedico a mis padres, Zu Ling Feng y Rui Fu Wu, por su sacrificio y esfuerzo, por haberme inculcado todos los valores por los que me caracterizo hoy, por todas esas horas de trabajo que dieron los siete días de la semana, los trescientos sesenta y cinco días del año desde mi infancia para proveerme de la mejor educación posible. Esta educación culmina con el éxito de la obtención de esta maestría y del reconocimiento de summa cum laude por parte de la universidad. Sobre todo, por siempre brindarme comprensión, cariño y amor.

A todas aquellas personas que me alentaron a seguir cuando ya me sentía cansado y agobiado durante el desarrollo de la tesis, y especialmente durante la etapa de la carrera, sus palabras dieron como fruto la conclusión de esta maestría.

Agradecimientos

En primer lugar, a mi hermana Sujeili, por haberme ayudado en la defensa de la tesis.

Gracias al profesor Luis Naranjo por haber aceptado ser mi tutor y guía en el arduo camino del desarrollo de esta tesis, así como haber confiado en mi trabajo, y sobre todo por el apoyo que me dio para lograr mis objetivos.

Al profesor Marco Hernández, le doy las gracias no solo por haber aceptado ser mi lector primario, sino también por sus consejos, recomendaciones y demás contribuciones que permitieron que este trabajo fuera satisfactorio y exitoso para mi persona.

Gracias al profesor Ignacio Trejos y a María Eugenia Ucrós por sus recomendaciones, por sus palabras de apoyo para la finalización de este proyecto, así como el incansable y arduo trabajo que dan a la universidad, lo cual pude observar durante mi etapa de la carrera.

A mi gran amigo Mauricio Guzmán Obando, le agradezco sus aportes en materia de medicina que permitieron que este trabajo cumpliera con todos los objetivos que se deseaba cumplir.

Gracias a mi gran amiga Paola Granados por apoyarme desde el primer día que decidí el tema para este trabajo final de graduación, y por su ayuda durante la elaboración del anteproyecto.

A mi amigo Juan Pablo Fernández por ayudarme durante toda la etapa de la carrera.

Quiero también agradecer a todos aquellos que participaron directa o indirectamente en la elaboración de esta tesis, y, sobre todo, a todas aquellas personas que de alguna manera marcaron cada etapa de mi vida.

Epígrafe

“Investigar es ver lo que todo el mundo ha visto, y pensar lo que nadie más ha pensado”

Albert Szent-Györgyi

TRIBUNAL EXAMINADOR

Este proyecto fue aprobado por el Tribunal Examinador de la carrera: **Maestría en Tecnología de Bases de Datos,** requisito para optar por el título de grado de **Maestría,** para el estudiante: **Greivin Wu Feng.**

MBA. Luis Carlos Naranjo Z.
Tutor

MBD. Marco Hernández V.
Lector 1

M. Sc. Ignacio Trejos Zelaya
Lector 2

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Resumen Ejecutivo

El cáncer es una de las causas primarias de muertes a nivel mundial. Según datos de la Organización Mundial de la Salud (OMS), el cáncer de pulmón es uno de los que causa un mayor número de muertes anuales tanto en hombres como mujeres en comparación a otros tipos de cáncer. Los síntomas frecuentes del cáncer de pulmón son la dificultad de respirar, tos y en algunos casos esta viene acompañada con sangre, entre otras manifestaciones que pueden llevar a confundirlo con cualquier otra enfermedad respiratoria. De allí la importancia de una detección temprana, pues además posee una alta mortalidad.

El objetivo de este estudio es demostrar la detección de cáncer de pulmón en un enfoque de procesamiento de imágenes médicas de tomografía computarizada. El desarrollo del enfoque ofrecerá la aplicación de conceptos de procesamiento de imágenes, técnicas de segmentación de imágenes y finalmente aplicación de algoritmos de clasificación de minería de datos y aprendizaje de máquinas en un ambiente de Big Data.

Palabras Clave: Cáncer pulmonar, cáncer de pulmón, detección de cáncer, tomografía computarizada, minería de datos, machine learning, aprendizaje de máquinas, Big Data.

Capítulo 1. Introducción

1.1 Generalidades

Las imágenes que se utilizarán para el estudio son obtenidas de colecciones de imágenes de diversas bases de datos del consorcio de cáncer de pulmón, por tanto, son de acceso público, es decir, no se utiliza ni se utilizará imágenes de tomografía computarizada (TAC) de pacientes sin el debido consentimiento, que violentan su derecho a la privacidad médica. Se plantea la idea del uso de Apache Hadoop y sus API o framework complementarios como la tecnología de Big Data a aplicar.

1.2 Antecedentes del Problema

El cáncer de pulmón es una enfermedad con alta tasa de mortalidad en el mundo, se estima que anualmente se detectan 299 casos y un promedio de 309 personas fallecen en Costa Rica debido a la enfermedad. En países como Estados Unidos, según la Sociedad Americana Contra El Cáncer, para este 2017, se estima la detección de 222 500 nuevos casos, de los cuales más de la mitad, exactamente 155 870 personas tienen un alto porcentaje de mortalidad, debido a la alta complejidad del padecimiento.

Existen diversas propuestas en materia de asistencia computarizada en el área de la medicina para la detección de este padecimiento, desde publicaciones académicas hasta un sinfín de algoritmos, enfoques y estudios con el objetivo de proponer un modelo de detección de cáncer de pulmón que ayude al diagnóstico temprano de este padecimiento mediante imágenes médicas. Se menciona la palabra ayudar porque la finalidad es proveer una herramienta que permita asistir en la toma de decisión por parte del médico, y de esta manera proveer un diagnóstico más acertado al paciente.

Muchas de las soluciones expuestas involucran los conceptos de *Machine Vision* para el tratamiento de las imágenes y la minería de datos aplicada en algoritmos de clasificación con la finalidad de poder predecir con seguridad la ausencia o presencia de tumores, los cuales pueden ser señal de cáncer de pulmón, así como la aplicación de aprendizaje de máquinas (*machine learning* en inglés) y el uso de redes neuronales o máquinas de soporte vectorial, para enumerar algunos. Ninguna expone las posibilidades de una implementación o el desarrollo de una herramienta en función de

conceptos de Big Data con un ambiente de almacenamiento de datos distribuido y procesamiento en paralelo, que se pueda utilizar en ecosistemas médicos como clínicas, hospitales y consultorios, entre otros.

1.3 Definición y Descripción del Problema

La detección de cáncer de pulmón mediante procesamiento de imágenes es un problema habitual por resolver en el área de diagnósticos médicos asistidos por computadora, aprendizaje de máquinas y ciencias de datos, entre otras áreas de la computación.

Muchas propuestas se han publicado y expuesto de diversos investigadores y autores, pero a la fecha no existe una implementación real aplicable en el área de la medicina para la asistencia debido a la gran variedad de metodologías, desde soluciones que exponen únicamente el uso de técnicas de procesamiento de imágenes y detección de patrones hasta alternativas que combinan con aprendizaje de máquinas y minería de datos. Es por ello que el problema a resolver es poder desarrollar y demostrar una solución de detección de cáncer de pulmón mediante el uso de imágenes médicas, específicamente tomografía computarizada (TAC), para asistir en la detección por parte del médico con una probabilidad de acierto equivalente o mayor al 80%.

1.4 Justificación

En 2015, un estimado de 1,69 millones de personas murieron debido a cáncer de pulmón en el mundo según las estadísticas obtenidas de la Organización Mundial de la Salud, con lo cual se coloca como el tipo de neoplasia maligno número uno en causar defunciones, seguido por el cáncer hepático (788 000 defunciones) y colorrectal (774 000 defunciones). En 2016, durante el Congreso Mundial de Cáncer de Pulmón realizado en diciembre de dicho año en Viena, se expone el resultado prometedor de nuevos medicamentos que pueden funcionar en el tratamiento de pacientes en etapa avanzada sin la necesidad de requerir quimioterapia.

Todos estos números y logros son muestra de la gran cantidad de tiempo, recurso humano y financiero que se gasta en el estudio de esta enfermedad y la búsqueda de tratamientos que mitiguen sus efectos o curas definitivas al padecimiento. Dados los anteriores antecedentes, se justifica

plenamente la realización del presente trabajo con el fin de poder asistir en la detección temprana de la enfermedad, al proveer una herramienta adicional que permita fundamentar la toma de decisiones de los médicos especialistas en la materia y de esa forma determinar el tratamiento a seguir para cada paciente según su patología y diagnóstico.

La investigación e implementación es exponer una metodología que puede ser aplicable no solo en cáncer de pulmón sino en otros órganos, es decir, una metodología o formulación para detección de cáncer para diversos escenarios, así como diversos enfoques médicos, de manera que se puede disminuir los tiempos de atención y detección tanto para los médicos como para los pacientes.

1.5 Viabilidad

1.5.1 Punto de Vista Técnico. El presente autor, como único investigador y desarrollador en este proyecto, tiene las capacidades técnicas necesarias. Debido a que la solución implica ambientes y tecnologías de Big Data, como lo es Hadoop, se afirma que tiene más de 4 años de experiencia en Java, lo que le permite entender partes del ecosistema de Hadoop como YARN y MapReduce. Sumado a ello, tiene 4 meses de experiencia en desarrollo de aplicaciones en Hadoop, 3 meses de experiencia en Apache Spark y, aunque no tiene mucha experiencia profesional en aplicación de minería de datos o aprendizaje de máquinas, presenta los conocimientos en algoritmos de clasificación (redes neuronales, árboles de decisión y bosques aleatorios) que podrían ser utilizados como métodos evaluativos y predictivos del diagnóstico. En conclusión, él tiene la capacidad necesaria para lograr desarrollar este trabajo investigativo en las próximas 14 semanas.

1.5.2 Punto de Vista Operativo. Este proyecto investigativo es con finalidades evaluativas, es decir, se desea desarrollar una forma de detección de cáncer pulmonar y evaluar su potencial en métricas de exactitud y precisión del diagnóstico. Por ello, no se involucra empresa, cliente o patrocinador alguno, y por lo tanto es posible realizar la investigación sin alterar el funcionamiento normal de empresa alguna.

1.5.3 Punto de Vista Económico. La posibilidad de poder desarrollar este trabajo en términos económicos es factible, ya que el costo en software es

cero. Esto debido a que Hadoop, como herramienta de Big Data a utilizar, permite según la distribución que se utilice una versión gratuita de nodo único para virtualización en máquinas virtuales. Acompañado de ello, los demás frameworks a utilizar para complementar el entorno de Hadoop, como Spark, Scala, Java y Kafka, entre otros, presentan licencia de Apache, lo que significa que se pueden adquirir e instalar gratuitamente.

En cuanto a recurso humano, solo es una persona, por tanto, con base en la lista de salarios del Ministerio de Trabajo y Seguridad Social de Costa Rica, el salario mínimo para un programador de computación por jornada laboral de 8 horas es de 12,829.63 colones (Ministerio de Trabajo y Seguridad Social, 2017). Este valor representará el costo de esfuerzo y horas en el trabajo de este proyecto.

Se estima que el trabajo se completará en 14 semanas, donde por cada semana se dedicará 24 horas totales de trabajo en el proyecto investigativo, es decir un total de 336 horas aproximadamente. Por lo tanto, el costo total, también denominado “costo teórico”, equivale a 538,844.46 colones aproximadamente. Esto será asumido por el presente investigador como único recurso humano.

1.6 Objetivos

Para el presente trabajo investigativo, se decide utilizar la Taxonomía de Webb como taxonomía cognitiva, debido a que esta se centra en cuatro niveles de complejidad crecientes con la finalidad de poder demostrar las capacidades adquiridas en el proceso educativo. El proyecto se basa en esta idea, porque se desea poder aplicar todo el conocimiento adquirido en la Maestría de Tecnología de Bases de Datos impartida en la Universidad Cenfotec, para analizar, crear y evaluar un enfoque para apoyo en diagnósticos médicos en detección de cáncer de pulmón mediante procesamiento de imágenes.

1.6.1 Objetivo General

Desarrollar una solución de asistencia médica computarizada para detección de cáncer de pulmón mediante imágenes de tomografía computarizada (TAC), utilizando herramientas de Big Data que ayuden a la detección temprana del problema.

1.6.2 Objetivos Específicos

- Identificar diferentes enfoques, metodologías o soluciones de detección de cáncer mediante procesamiento de imágenes con un buen porcentaje de exactitud y precisión.
- Determinar los mejores algoritmos de preprocesamiento de imágenes, filtrado de imágenes, segmentación y detección de patrones en imágenes médicas.
- Determinar los algoritmos de minería de datos de tipo clasificación (árbol de decisión, bosque aleatorio y redes neuronales, entre otros) que complementarán el diagnóstico predictivo de la solución.
- Construir una metodología de detección adaptativa bajo la combinación de las mejores implementaciones en aplicación de los mejores algoritmos definidos.
- Seleccionar las diversas imágenes a utilizar como pruebas, a fin de tener las mejores muestras.
- Valorar la exactitud, además, si es posible, la sensibilidad y especificidad, de la nueva propuesta versus algunas soluciones publicadas.

1.7 Alcances y Limitaciones

1.7.1 Alcances. Los entregables son: documento escrito que conforma el trabajo final de graduación donde se definen todos los aspectos del desarrollo y evaluación del mismo cumpliendo con los objetivos generales y específicos. También se entregará un video donde se muestra a manera de DEMO el programa ejecutable con la solución al problema expuesto en ambiente de Hadoop con imágenes reales, y de ser requerido un artículo científico-académico.

La solución aplicará procesamiento paralelo que provea en cuanto a los aspectos evaluativos conclusivos de exactitud y precisión cercanas o mayores al 80% de probabilidad de diagnóstico acertado.

1.7.2 Limitaciones. No se realizará entrega del código fuente de toda la solución ni el ejecutable, esto debido a las pocas semanas para la realización de este trabajo final de graduación. El autor divisa un potencial de mejora funcional fuera de la entrega de la propuesta, así como una posibilidad de

negocio que requiere mayor tiempo e inversión financiera para la creación de una solución más robusta y completa para el bien común en materia de medicina y salud.

Tampoco se explicará el cómo se debe instalar la máquina virtual de la distribución de Hadoop a utilizar en este trabajo, o la configuración de la misma debido a que es sencillamente instalable, únicamente consiste en descargar los archivos para el software de virtualización a utilizar y ejecutarlo. De igual manera, sea cual sea la distribución que se utilice, la compañía propietaria tiene documentados estos pasos.

1.8 Estado de la Cuestión

1.8.1 Planificación de la revisión

1.8.1.1 Formulación de la pregunta. Identificar las mejores propuestas para detección de cáncer de pulmón mediante el procesamiento de imágenes, minería de datos y aprendizaje de máquinas con aplicabilidad en tecnologías de Big Data.

1.8.1.2 Amplitud y calidad de la pregunta.

- Problema: El cáncer de pulmón es el tipo de neoplasia maligno con mayores muertes a nivel mundial. No existe una manera de detección efectiva debido a la sintomatología del padecimiento, pero muchos han propuesto diversas soluciones computacionales, y de ello se desprende nuestro problema que es identificar las mejores soluciones propuestas y de fácil implementación computacional para un entorno de Big Data, específicamente Hadoop.
- Pregunta de investigación: ¿Cuáles soluciones, de las diversas propuestas para la detección de cáncer pulmonar mediante imágenes de tomografía computarizada, presentan una facilidad de implementación en software y un alto porcentaje de acierto en un diagnóstico de identificación de tumores neoplásicos?
- Palabras clave y sinónimos: A continuación, un cuadro de resumen de las palabras clave y conceptos relacionados que utilizaremos en esta revisión.

Palabra clave	Sinónimo	Traducción en inglés
Cáncer de pulmón	Cáncer pulmonar	Lung cancer
Aprendizaje por máquina	Machine learning, aprendizaje automático	Machine learning
Tomografía computarizada	TAC	Computerized tomography
Minería de datos		Data mining
Procesamiento de imagen		Image processing

- Intervención: En el contexto de la revisión sistemática planificada, se observará todas las posibles propuestas y enfoques existentes sobre detección de cáncer de pulmón mediante imágenes para analizar y encontrar las que mejor se ajusten tanto en resultados positivos como facilidad de implementación.
- Control: En la presente revisión sistemática se han observado algunos trabajos para poder definir las palabras clave y la idea general del trabajo investigativo, pero ninguno que clasifique como un conjunto de resultados derivado de los criterios definidos y que cumplen con el objetivo buscado.
- Resultado: Los resultados esperados de esta revisión sistemática son detectar y conocer aquellas propuestas en materia de detección de cáncer pulmonar, para posteriormente analizarlas y encontrar aquellas que mejor probabilidad de acierto presenten como facilidad de implementación computarizada.
- Medida de salida: En la medición de los resultados obtenidos, iniciaremos con obtener el número de propuestas, enfoques, algoritmos, metodologías identificadas para la realización de una comparación de

todas mediante la comparación del análisis de resultados en el valor de precisión de diagnóstico de cada uno.

- Población: La población sujeta a análisis está compuesta por las publicaciones presentes en los repositorios seleccionados y que se aportan al tema de este trabajo investigativo.
- Aplicación: Los beneficiarios de la revisión sistemática serán las personas relacionadas directamente con el estudio y aplicación de computación asistida en el área de la medicina, así como todas aquellas personas con interés en conocer los trabajos relevantes existentes en materia de detección de cáncer.
- Diseño experimental: No se aplicará un metaanálisis de la revisión sistemática.

1.8.2 Selección de fuentes

1.8.2.1 Definición del criterio de selección de fuentes. El criterio para la selección de las fuentes de búsqueda está basado en la opinión del autor de este trabajo, el cual, basándose en su experiencia profesional como desarrollador de soluciones de software, inteligencia de negocios y Big Data, sumado al conocimiento en matemáticas, probabilidades, estadística y minería de datos, sugerirá la lista de fuentes sobre las cuales realizar la revisión. Otros requisitos exigidos a las fuentes para su selección son accesibilidad vía web y la inclusión de motores de búsqueda que permitan la realización de consultas avanzadas.

1.8.2.2 Lenguaje de estudio. El lenguaje de los estudios primarios será en inglés, pero el informe de la revisión sistemática se realiza en español.

1.8.2.3 Identificación de fuentes.

- Método de selección de fuentes: La disponibilidad de revistas y artículos en la Web que cumplen los criterios de selección previamente establecidos, presentan motores de búsqueda para consultas complejas y que se encuentra calificadas un rango de confiabilidad aceptable en la plataforma métrica de revistas web Scimago Jr.
- Lista de fuentes: La lista de fuentes obtenida sobre la cual se ejecutará la revisión sistemática es:

- IEEE Xplore Digital Library
- Google Scholar
- International Journal of Biomedical Data Mining
- DBLP
- IOSR journals
- International Journal of Recent Innovation in Engineering & Research (IJRIER)

- Cadenas de búsqueda: Combinaciones de los operadores lógicos “Y” y “O” sobre las palabras clave, así como conceptos sinónimos citados anteriormente en la sección “Palabras clave y sinónimos”. De esta manera, establecemos la siguiente cadena de búsqueda a utilizar en la presente revisión:

(lung cancer and (image processing and (machine learning or data mining)) and (computerized tomography or CT))

1.8.2.4 Selección de fuentes después de la evaluación. Todas las fuentes seleccionadas han cumplido y satisfecho los criterios de selección. Pero se hará un refinamiento con la finalidad de agregar artículos provechosos que no se encuentran en las fuentes seleccionadas.

1.8.2.5 Comprobación de las fuentes. Todas las fuentes serán aprobadas al cumplir con el método de selección de fuentes estipulado.

1.8.3 Selección de los estudios

1.8.3.1 Procedimiento para la selección de los estudios.

El proceso de selección de los estudios primarios de las diversas fuentes sigue las etapas mostradas en el siguiente diagrama de flujos.

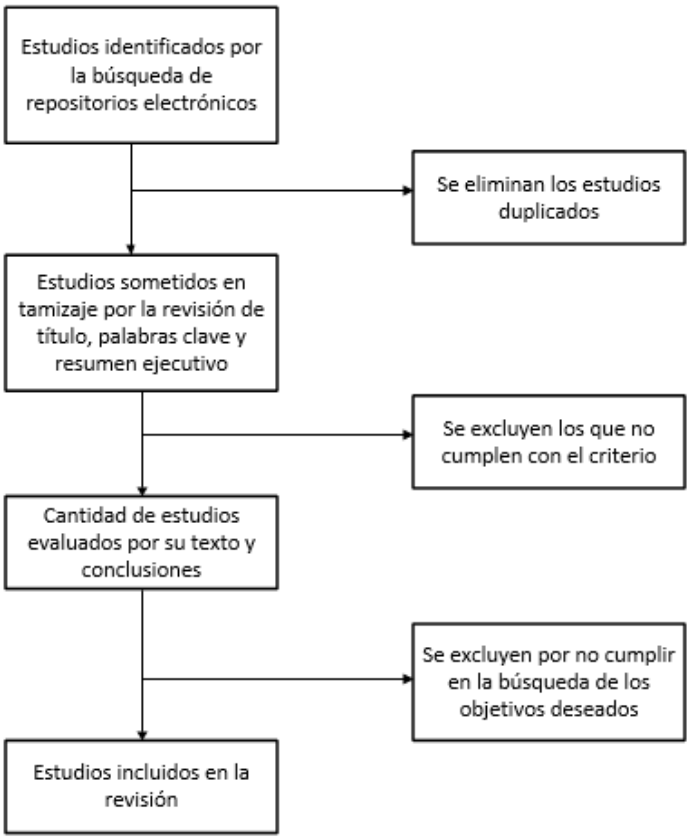


Figura 1: Procedimiento de selección de estudios primarios. Fuente: Elaboración propia.

1.8.3.2 Definición del criterio de inclusión y exclusión de estudios.

Como criterio de inclusión se efectúa principalmente un análisis sobre el título, las palabras claves y el resumen ejecutivo de cada documento, esto permitirá eliminar aquellos estudios que no aportan al tema e incluir aquellos que realizan aportaciones.

La lectura y el análisis detallado del resumen ejecutivo y las conclusiones concentrarán el criterio de exclusión. De esta manera se es capaz de filtrar detalladamente de la temática de cada documento, las ideas principales del mismo, su relevancia para la revisión sistemática y su aporte a los objetivos buscados.

1.8.3.3 Definición de tipos de estudio. Los estudios por utilizar como base serán aquellos adquiridos en las fuentes definidas en la sección 1.8.2.3 del presente documento, que cumplen con el criterio definido de selección de fuentes y que hayan pasado por el procedimiento definido en la sección 1.8.3.1.

1.8.4 Extracción de información

1.8.4.1 Definición del criterio de inclusión y exclusión de información. La información obtenida de los estudios debe incluir técnicas, metodologías, algoritmos y fórmulas que detallen la manera de cómo es posible detectar cáncer de pulmón en escaneos de tomografía computarizada y el cómo evaluar su desempeño en la identificación de la neoplasia de tipo maligna.

1.8.4.2 Formulario para la extracción de información. El formulario que se utiliza para documentar la extracción de información realizada sobre cada estudio primario consta de una primera parte donde, mediante la identificación del título y la publicación, se determina el estudio para luego poder tener una breve descripción general en la que se analiza la propuesta de detección de cáncer pulmonar. Además se hace el análisis de las conclusiones o resultados obtenidos en cada estudio primario. Por último, se incluye una parte en la que se documenta varios aspectos a destacar, los cuales el autor de este trabajo ha considerado importantes, conforme se realizó el análisis sistemático sobre cada estudio primario.

Identificación	Título
	Publicación
Descripción	Resumen
	Resultados
	Conclusiones
Aspectos a destacar	

Figura 2: Formulario de extracción de información. Fuente: Elaboración propia.

1.8.4.3 Extracción de resultados objetivos y subjetivos.

Esta sección tiene como objetivo registrar el proceso de selección de los estudios primarios, informando los estudios obtenidos y el resultado de su evaluación. Todo se puede apreciar desde la Tabla 1 hasta la Tabla 5.

Identificación	
Título	A Computer Aided Diagnosis System for Lung Cancer Detection\Using Support Vector Machine
Autores	Gomanthi M.; Thangaraj P.
Año	2010

Descripción	
Elementos tratados: • Utilización de imágenes de TAC. • El estudio se centra en disminuir los falsos positivos del método convencional de detección automática mediante otra propuesta.	
Aspectos por destacar	
El texto presenta suficiente fundamento matemático en cuanto a cada aspecto definitorio de la máquina de soporte vectorial y sus parámetros.	
Extracción de resultados objetivos	
Metodología del estudio	Los autores exponen la posibilidad del uso de máquinas de soporte vectorial en sistemas de asistencia computarizada.
Resultados del estudio	De 15 nódulos cancerígenos, solo se logra la detección de 9 correctamente. Pero, aunque es baja probabilidad, tiene fundamento matemático y algorítmico en materia de aprendizaje de máquinas suficiente para revisar y corregir en un nuevo enfoque o perspectiva.
Problemas del estudio	Dificultad de implementación y no se precisa la probabilidad de exactitud del mismo.
Extracción de resultados subjetivos	
Información a través de los autores	No fue solicitado.
Impresiones generales y abstracciones	Los autores afirman la necesidad siempre de aplicar técnicas de preprocesamiento de imágenes y segmentación para obtener el conjunto de datos a definir los hiperplanos del algoritmo de máquina de soporte vectorial.

Tabla 1: Primera fuente literaria investigativa. Información obtenida de Gomathi & Thangaraj (2011).

Fuente: Elaboración propia.

Identificación	
Título	Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques
Autores	Gajdhane A., Vijay; L. M., Deshpande
Año	2014
Descripción	
Elementos tratados: • Utilización de imágenes de TAC. • Técnicas de procesamiento de imágenes digitales. • Uso de máquinas de soporte vectorial como algoritmo de aprendizaje de máquinas de tipo clasificación.	
Aspectos por destacar	
El texto aplica el cálculo de los aspectos de área, perímetro y diámetro en función de los pixeles de la imagen luego de segmentada para comparar con una tabla de valores provista por un experto en medicina de la materia.	
Extracción de resultados objetivos	
Metodología del estudio	Los autores exponen la posibilidad del uso de máquinas de soporte vectorial de la mano de técnicas de procesamiento de imágenes.
Resultados del estudio	Los autores no definen qué tan exitosa fue su propuesta en métricas de exactitud, sensibilidad, especificidad, etc.
Problemas del estudio	No se encuentra ninguno.
Extracción de resultados subjetivos	
Información a través de los autores	No fue solicitado.
Impresiones generales y abstracciones	Posibilidad de poder aplicar a futuro no solo a imágenes TAC, sino también rayos X, resonancias magnéticas (MRI).

Tabla 2: Segunda fuente literaria investigativa. Información obtenida de Gajdhane & L.M. (2014). Fuente: Elaboración propia.

Identificación	
Título	Lung Cancer Detection with fusion of CT and MRI Images Using Image Processing
Autores	Deshpande S., Anuradha; Lokhande D., Dhanesh; Mundhe P., Rahul; Ghatole M., Juilee
Año	2015
Descripción	
Elementos tratados: • Utilización de imágenes de TAC y resonancia magnética. • Aplicación de técnicas de procesamiento de datos como segmentación. • Los autores definen que utilizaron MATLAB para el desarrollo de la solución.	
Aspectos por destacar	
El texto muestra la serie o conjunto de pasos a seguir para desarrollar la solución propuesta por los autores. Además, muestra un diagrama de flujo del proceso.	
Extracción de resultados objetivos	
Metodología del estudio	Un enfoque muy completo donde se expone el complemento e integración de las áreas de: procesamiento de imágenes y aprendizaje de máquinas.
Resultados del estudio	Comparación entre los distintos métodos convencionales de procesamiento de imágenes.
Problemas del estudio	No se encuentra ninguno.
Extracción de resultados subjetivos	
Información a través de los autores	No fue solicitado.
Impresiones generales y abstracciones	Posibilidad de desarrollar la solución en MATLAB.

Tabla 3: Tercera fuente literaria investigativa. Información obtenida de Deshpande S., Lokhande D., Mundhe P., & Ghatole M. (2015).

Fuente: Elaboración propia.

Identificación	
Título	Lung cancer detection using modified log-gabor filter based features
Autores	Rupali R. Mali
Año	2017
Descripción	
Elementos tratados: • Utilización de imágenes de TAC. • Técnicas de procesamiento de imágenes digitales. • Programación mediante MATLAB.	
Aspectos por destacar	
El texto muestra la serie o conjunto de pasos a seguir para desarrollar la solución propuesta por los autores. Además, muestra un diagrama de flujo del proceso.	
Extracción de resultados objetivos	
Metodología del estudio	Un enfoque de seis pasos simples a realizar.
Resultados del estudio	Probabilidad de éxito del enfoque de 89.56% mediante el uso de matrices de confusión.
Problemas del estudio	No se explica el cómo se implementa el filtro log de Gabor ni cuál es la metodología o técnica utilizada en la etapa de extracción de información de las imágenes luego de segmentadas.
Extracción de resultados subjetivos	
Información a través de los autores	No fue solicitado.
Impresiones generales y abstracciones	Un enfoque posible por utilizar, pero con deficiencias al no explicar muchos aspectos importantes como la forma de desarrollo o fórmula de relación utilizada en la aplicación del filtro de Gabor a la imagen, y cuál o cuáles algoritmos de clasificación se utilizan.

Tabla 4: Cuarta fuente literaria investigativa. Información obtenida de Mali (2017).

Fuente: Elaboración propia.

Identificación	
Título	Lung cancer classification using neural networks for CT images
Autores	Kuruvilla, Jinsa; Gunavathi, K.
Año	2013
Descripción	
Elementos tratados: • Utilización de imágenes de TAC. • Segmentación de imágenes. • Probabilidad de éxito de 93.3% mediante aplicación de redes neuronales.	
Aspectos por destacar	
Un texto muy completo donde el autor define mediante aplicación matemática cómo se calcula cada uno de los parámetros que definen la etapa de entrenamiento de la red neuronal y su aplicación en dos tipos de redes neuronales: propagación hacia adelante y propagación hacia atrás.	
Extracción de resultados objetivos	
Metodología del estudio	Aplicación de todos los conceptos disponibles para esta investigación. Es una metodología simple y que en general sigue los mismos pasos propuestos en los otros estudios seleccionados.
Resultados del estudio	Utilización de una matriz de confusión para obtener las métricas de especificidad, sensibilidad y exactitud. En conclusión, se tiene una solución con una probabilidad mayor a 90% bajo afirmaciones de los autores.
Problemas del estudio	No se encuentra ninguno.
Extracción de resultados subjetivos	
Información a través de los autores	No fue solicitado.
Impresiones generales y abstracciones	Un algoritmo con una alta probabilidad de exactitud.

Tabla 5: Quinta fuente literaria investigativa. Información obtenida de Kuruvilla & Gunavathi (2014).

Fuente: Elaboración propia.

1.8.4.4 Resolución de divergencias entre los revisores.

No hay divergencia alguna.

1.8.5 Análisis de resultados

1.8.5.1 Resultados cálculo estadístico. No se realizaron cálculos estadísticos.

1.8.5.2 Presentación de resultados.

La siguiente tabla de resultados muestra los tipos de estudios seleccionados en combinación con la idea de procesamiento o preprocesamiento de imágenes de tomografía computarizada, a la que sumamos también la posibilidad de enfoques con solo técnicas de procesamiento de imágenes para detección de nódulos cancerígenos.

	Aprendizaje por máquinas	Procesamiento de imágenes	Minería de datos	Procesamiento de imágenes y minería de datos	Procesamiento de imágenes y aprendizaje de máquinas	Combinación de los 3	Total
# estudios seleccionados	0	1	0	0	4	0	5

Figura 3: Cantidad de estudios por tipo de solución. Fuente: Elaboración propia.

Como criterio de extracción de los estudios y la información de los mismos, se utiliza el criterio de si el o los autores respectivos proveen detalles de cómo desarrollaron su implementación en software.

	Código de programación presente en algún lenguaje	Pseudo-código	Fórmulas matemáticas	Explicación textual	Total
# estudios encontrados	0	0	4	8	12
# estudios seleccionados	0	0	2	3	5

Figura 4: Resultados por desarrollo de implementación. Fuente: Elaboración propia.

Y finalmente, el criterio de resultados obtenidos de las conclusiones de cada estudio seleccionado.

Probabilidad de acierto	No definido	< 50%	50%-70%	70%-80%	80%-90%	> 90%	Total
# estudios seleccionados	1	0	0	1	1	1	5

Figura 5: Probabilidad exactitud obtenida en los estudios. Fuente: Elaboración propia.

1.8.5.3 Análisis de sensibilidad. No fue aplicado.

1.8.5.4 Gráficos. No fue aplicado.

1.8.5.5 Comentarios finales.

- Número de estudios: 15 estudios encontrados, 7 seleccionados.
- Sesgo de búsqueda, selección y extracción: Todos aquellos estudios donde sus conclusiones provean una probabilidad de exactitud menor al 75% fueron descartados.
- Sesgo de publicación: No fue definido.
- Variación entre revisores: No hay variación alguna.
- Aplicación de resultados: Los estudios seleccionados comparten de manera general el mismo método de procesamiento de imágenes, de ahí la posibilidad de combinar únicamente en la implementación las diversas técnicas de minería de datos y aprendizaje de máquina a fin de comparar el mejor resultado para diagnóstico o incluso aplicación de teoremas de estadística como el de Bayes de independencia entre probabilidades.
- Recomendaciones: Ninguna.

Capítulo 2. Marco Conceptual

A continuación se presenta el sustento conceptual generalizado de las literaturas seleccionadas en el capítulo anterior, tomado en consideración para la elaboración de la investigación sobre cáncer de pulmón y los modelos de detección del mismo mediante el procesamiento de imágenes computarizadas. Esta información ayudará a una mejor interpretación y aplicación del proyecto.

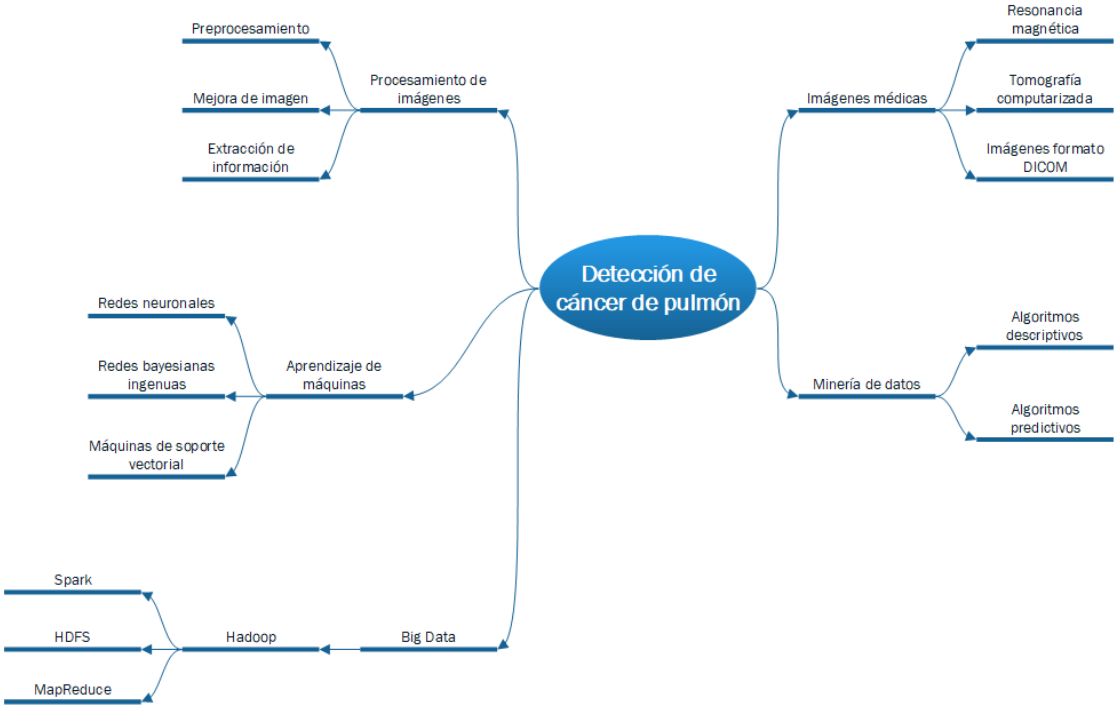


Figura 6: Diagrama conceptual. Fuente: Elaboración propia.

2.1 Tomografía computarizada (TAC)

Procedimiento computarizado de imágenes por rayos X en el que se proyecta un haz angosto de rayos X a un paciente y se gira rápidamente alrededor del cuerpo, produciendo señales que son procesadas por la computadora de la máquina para generar imágenes transversales -o “cortes”- del cuerpo. Estos cortes se llaman imágenes tomográficas y contienen información más detallada que los rayos X convencionales. Una vez que la computadora de la máquina recolecta varios cortes sucesivos, se pueden “apilar” digitalmente para formar una imagen tridimensional del paciente que permita más fácilmente la identificación y ubicación de las estructuras básicas, así como de posibles tumores o anormalidades (biomédicas, 2013).

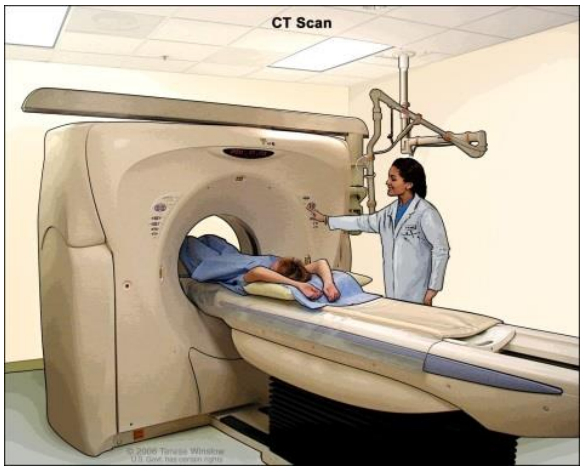


Figura 7: Ejemplo de TAC. Obtenida de Biomédicas (2013).

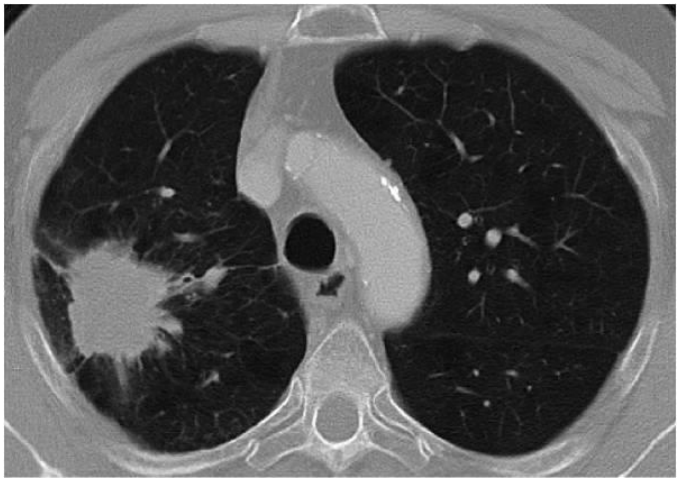


Figura 8: Ejemplo de imagen pulmonar de TAC. Obtenida de DocCheck Pictures.

2.2 Cáncer de pulmón

La medicina actual define neoplasia como un crecimiento descontrolado de células que no están bajo control fisiológico, es decir, la formación de tumores. Estos se pueden clasificar en benigno o maligno, donde el segundo se caracteriza por una tendencia al crecimiento rápido, de manera tal que invaden otros tejidos y se pueden diseminar o propagar a otras regiones del cuerpo, proceso llamado metástasis.

Es bajo la categoría de neoplasias malignas en la que definimos a un cáncer. Por tanto, cuando hablamos de cáncer de pulmón o cáncer pulmonar, nos referimos al tipo de neoplasia maligna que ocurre específicamente en estos órganos.

2.2.1 Causas

- La principal causa conocida es el tabaquismo. El fumar aumenta considerablemente el riesgo de padecimiento debido a la alta concentración de cancerígenos presente en los cigarros.
- Algunas enfermedades como la tuberculosis o ciertas neumonías pueden dejar cicatrices en los pulmones, lo que aumenta el riesgo en desarrollar cáncer de pulmón.
- La herencia genética es un factor en todos los seres vivos, ya que es la manera de poder transmitir características de un ser a otro afín con la finalidad primaria de supervivencia, de la mano de la evolución. La transmisión de genes por parte de los padres puede ser una de las causas, a esto se le llama predisposición genética.
- Cancerígenos como el amianto, uranio, arsénico y algunos productos derivados del petróleo.

2.2.2 Síntomas

- Cansancio.
- Pérdida de apetito.
- Tos seca o con flemas. Este es la manifestación más frecuente.
- Expulsión de sangre de las vías respiratorias (Hemoptisis o expectoración sanguinolenta).
- Sensación de dificultad para respirar conocida como disnea. El paciente ve complicada su capacidad para realizar esfuerzos físicos como subir escaleras.
- Dolor torácico.

2.2.3 Tipos de cáncer de pulmón

Existen dos tipos principales de cáncer de pulmón: cáncer de pulmón de células no pequeñas (NSCLC, del inglés Non-small cell lung cáncer) y cáncer de pulmón de células pequeñas (SCLC, del inglés Small cell lung cancer).

Los NSCLC son el tipo de cáncer pulmonar más común y “representan el 85% de los casos” (Afsaneh & Pennell, 2010). Crece y se propaga más lentamente que el cáncer de pulmón de células pequeñas. La enfermedad en etapa temprana se asocia con pocos síntomas específicos; por lo tanto,

aproximadamente el 70% de los casos no se diagnostican hasta que la enfermedad se encuentra en una etapa avanzada, cuando las posibilidades de curación o beneficio significativo para el paciente son limitadas (Schiller, y otros, 2002).

Por el otro lado, “alrededor del 13% al 15% de los cánceres de pulmón son SCLC alrededor del mundo” (Rosti, y otros, 2006). Es más agresivo, ya que es de rápido crecimiento y propagación a otras zonas del interior del cuerpo, lo cual provoca que la muerte de la persona sea incluso en pocas semanas.

2.2.4 Criterio médico de posible neoplasia maligna en imágenes médicas

Mediante escaneos de rayos X o tomografías computarizadas, los nódulos o crecimiento anormal de células en los pulmones se muestran generalmente como sombras blancas y redondas. Pero existe un sistema que se utiliza para describir tanto la cantidad de cáncer como su diseminación en el cuerpo de un paciente mediante el uso de las letras TNM, donde:

- La T se refiere al tamaño y extensión del tumor principal.
- La N se refiere a la extensión de cáncer que se ha diseminado a los ganglios (o nódulos) linfáticos cercanos.
- La M se refiere a si el cáncer ha tenido metástasis. Es decir, el cáncer se ha diseminado desde el tumor primario a otras partes del cuerpo.

El sistema de estadificación TNM es ampliamente utilizado en muchos hospitales y clínicas. En cáncer de pulmón, existen cuatro categorías de la letra T para describir el tamaño de un tumor en pulmones. El resto de las letras son equivalentes a la etapa del cáncer y su propagación en el cuerpo, sobre lo cual no se entrará en detalle para esta investigación.

La siguiente tabla representa el cómo se describe los tumores en pulmones que pueden definir la presencia de cáncer de pulmón en imágenes de TAC.

T-Descriptor	Dimensiones	Clasificación adicional	Ubicación
T1	Menor o igual 1 cm	T1a	
	Mayor a 1cm hasta 2cm	T1b	
	Mayor a 2cm hasta 3cm	T1c	
T2	Mayor a 3cm hasta 4cm	T2a	Tumores que invaden el bronquio principal
	Mayor a 4cm hasta 5cm	T2b	
T3	Mayor a 5cm hasta 7cm		Tumores que invaden el bronquio principal
T4	Mayor a los 7cm		Pleura mediastinal

Figura 9: TNM octava edición obtenida de Goldstraw (2015).

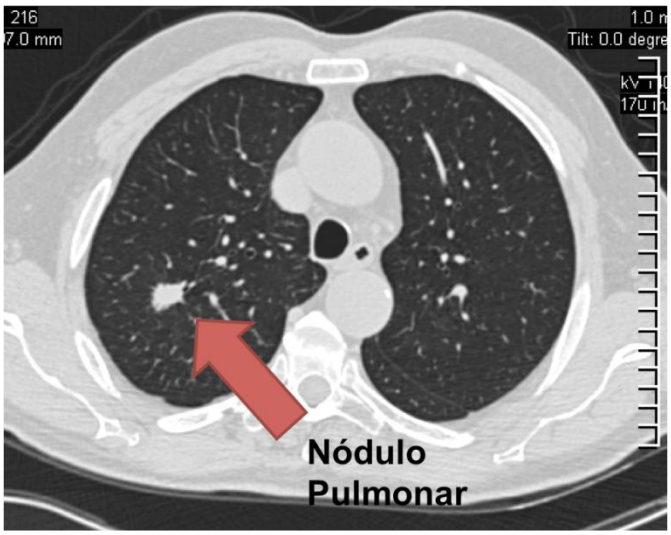


Figura 10: Ejemplo de un nódulo pulmonar en imagen de TAC. Obtenida de Defranchi.

2.3 Minería de datos

La minería de datos es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos (Maimon & Rokach, 2005). Su objetivo principal es la extracción de información para convertirla y transformarla en una estructura comprensible para su posterior uso, todo esto mediante algoritmos o métodos de tipo supervisado o no supervisado.

El primer tipo se fundamenta en la generalización de los datos para poder realizar inferencias y, con base en estas, predicciones. “Los algoritmos supervisados o predictivos predicen el valor de un atributo (*etiqueta*) de un conjunto de datos, conocidos otros atributos (*atributos descriptivos*). A partir de datos cuya etiqueta se conoce, se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en

datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como *aprendizaje supervisado* y se desarrolla en dos fases: Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos)" (García, Quintales, Peñalvo, & Martín, 2006).

Los métodos no supervisados, también llamados como de descubrimiento del conocimiento, se enfocan en encontrar patrones y tendencias en los datos actuales. Es decir, caracterizan las propiedades generales de los datos provistos. "El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficicio (científico o de negocio) de ellas" (García, Quintales, Peñalvo, & Martín, 2006).

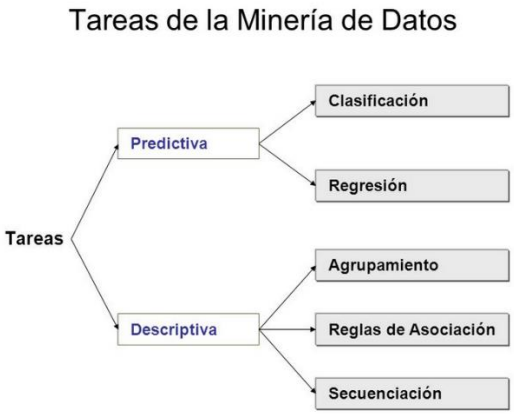


Figura 11: Obtenida de fuente Díaz (2005).

2.3.1 Técnicas principales de minería de datos

- Árboles de decisión: Son herramientas analíticas empleadas para el descubrimiento de reglas y relaciones mediante la ruptura y subdivisión sistemática de la información contenida en el conjunto de datos. El árbol de decisión se construye partiendo el conjunto de datos en dos (CART) o más (CHAID) subconjuntos de observaciones a partir de los valores que toman las variables predictoras. Cada uno de estos subconjuntos vuelve después a ser particionado utilizando el mismo algoritmo. Este proceso continúa hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta (Martínez, 2009).

- Agrupamiento o clustering: Permite la identificación de tipologías o grupos donde los elementos guardan similitud entre sí y diferencias con aquellos de otros grupos. Para alcanzar las distintas tipologías o grupos existentes en una base de datos, estas herramientas requieren, como entrada, información sobre el colectivo a segmentar. Como resultado del tratamiento de la información, estas herramientas presentan los distintos grupos detectados junto con los valores característicos de las variables (Martínez, 2009).
- Reglas de asociación: Técnicas que tienen como objetivo el encontrar asociaciones o correlaciones entre los elementos u objetos, y lograr "descubrir hechos que ocurren de manera común dentro de un determinado conjunto de datos" (Cedano, 2015).
- Secuenciación: Permite identificar cómo, en el tiempo, la ocurrencia de una acción desencadena otras posteriormente (Martínez, 2009).

Las redes neuronales también son una técnica de tipo predictiva-clasificación utilizada en el ámbito de la minería de datos. Pero son un paradigma de aprendizaje y procesamiento automático proveniente de la inteligencia artificial aplicable a la estadística; esto se detallará en secciones posteriores.

2.3.2 Proceso de minería de datos

Para poder asegurar un estudio o proyecto exitoso de minería de datos, se sugiere seguir un ciclo de vida de desarrollo. El enfoque más utilizado y aceptado es el modelo CRISP-DM (Cross Industry Standard Process for Data Mining). Este es un enfoque compuesto por seis fases que se observan en la figura 12, y será la metodología por utilizar en este trabajo.

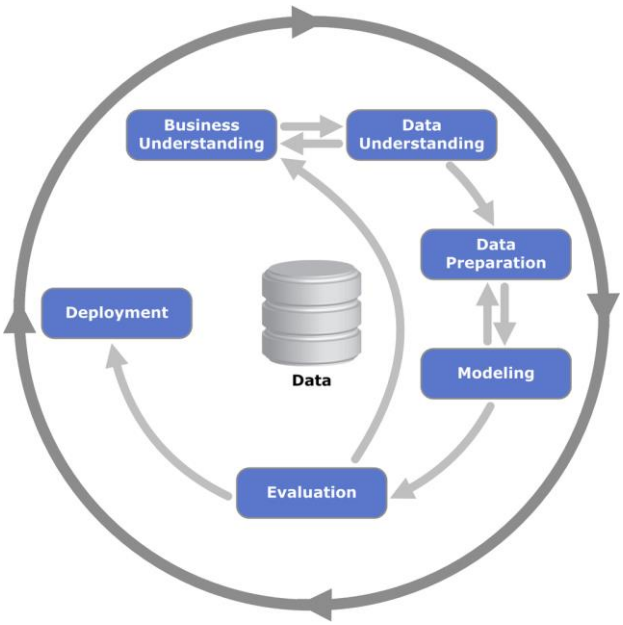


Figura 12: CRISP-DM Diagrama de proceso obtenido (Jensen, 2012)

La primera fase inicia con el entendimiento de los objetivos y requerimientos del proyecto, desde una perspectiva de negocio, para poder definir el problema de minería de datos que se desea resolver.

La fase dos es el entendimiento de la colección inicial de datos para poder proceder con las actividades que nos permitan encontrar problemas de calidad en los datos, las primeras pistas de información al problema y todo lo que nos permita formar hipótesis, para poder construir el conjunto de datos finales. En el caso de no cumplir con lo estipulado en la etapa de entendimiento del negocio, podemos ir hacia atrás y redefinir el alcance, objetivos, requerimientos, etc.

La etapa de preparación de datos consiste en tareas de transformación y limpieza de los datos, selección de atributos, tablas, registros y demás que nos permitan aplicar las técnicas de modelado en nuestra etapa siguiente de modelado. Durante esta fase, se seleccionan y aplican diversos modelos de datos con base en parámetros que nos permitan calibrar los valores óptimos y obtener los mayores detalles a nuestro modelo final. Como se observa en la figura 12, existe la posibilidad bidireccional, es decir, podemos redefinir cuantas veces sea necesario nuestro conjunto de datos a fin de poder encontrar los modelos y resultados aptos y adecuados a aplicar y obtener.

Finalmente se encuentra la etapa de evaluación de la información, para comprobar que respondemos a cada uno de los aspectos definidos en la primera etapa y se cumple con los objetivos del negocio. En la fase de despliegue la idea principal es dar visualización a nuestro modelo y los resultados obtenidos mediante reportes y gráficos, entre otros.

2.4 Aprendizaje de máquinas

El aprendizaje de máquinas o aprendizaje automático es una rama de la inteligencia artificial que “se encarga de estudiar y modelar computacionalmente los procesos de aprendizaje en sus diversas manifestaciones.” (Morales & González, 2013). Se entiende este como los procesos en los que las computadoras son capaces de aprender sin necesidad de programación, nos brindan dos escenarios en donde se puede definir el término clasificación de una manera contextualizada como el proceso de encontrar un modelo (o función) que describa y distinga clases de datos (o conceptos), con el fin de poder predecir la clase de objetos nuevos o desconocidos.

Es darle a las máquinas una de las capacidades distintivas de la inteligencia humana a fin de que puedan adquirir conocimiento, desarrollen habilidades a través de la inducción y la práctica en la resolución de problemas computacionales. Para ello la detección de patrones en los datos y el uso del descubrimiento de los mismos es fundamental.

En muchas ocasiones el campo de actuación del aprendizaje automático se solapa con el de Data Mining, ya que las dos disciplinas están enfocadas en el análisis de datos, sin embargo el aprendizaje automático se centra más en el estudio de la complejidad computacional de los problemas con la intención de hacerlos factibles desde el punto de vista práctico, no únicamente teórico (Introducción al Aprendizaje Automático, 2017). Sus aplicaciones se dan en el área de minería de datos, e inversamente la minería tiene sus aplicaciones en el área de aprendizaje computacional.

La principal diferencia entre aprendizaje de máquina y minería viene dado porque el primero busca que los modelos sean precisos, mientras que el segundo busca encontrar modelos interpretables de los datos en estudio. El complemento de ambas perspectivas en esta investigación será provechoso,

ya que abrirá el camino en la posibilidad de obtener un enfoque y modelo de detección de cáncer de pulmón que cumpla con los objetivos del trabajo investigativo.

Para Murphy (2012), existen tres tipos de aprendizaje computacional en función de las señales o retroalimentación recibida, del cual los dos primeros son los más comunes y comparten el mismo concepto de los dos tipos de algoritmos en minería de datos:

1. El enfoque por aprendizaje supervisado o predictivo.
2. Aprendizaje no supervisado o descriptivo.
3. Aprendizaje por reforzamiento donde mediante señales de castigo o premio se busca el cómo debe comportarse o actuar.

Las técnicas principales utilizadas en asistencia computarizada para la detección de enfermedades (por ejemplo, el cáncer en distintas zonas del cuerpo) tenemos:

- Redes neuronales
- Redes bayesianas
- Máquinas de soporte vectorial

2.4.1 Redes neuronales

Las redes neuronales son más que otra forma de emular ciertas características propias de los humanos, como la capacidad de memorizar y de asociar hechos. Si se examinan con atención aquellos problemas que no pueden expresarse a través de un algoritmo, se observará que todos ellos tienen una característica en común: la experiencia. El hombre es capaz de resolver estas situaciones acudiendo a la experiencia acumulada. Así, parece claro que una forma de aproximarse al problema consiste en la construcción de sistemas que sean capaces de reproducir esta característica humana. En definitiva, las redes neuronales no son más que un modelo artificial y simplificado del cerebro humano, que es el ejemplo más perfecto del que disponemos para un sistema que es capaz de adquirir conocimiento a través de la experiencia. Una red neuronal es “un nuevo sistema para el tratamiento de la información, cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona” (Matich, 2001).

Para entender el cómo una red neuronal artificial funciona, nos guiaremos por la figura 13.

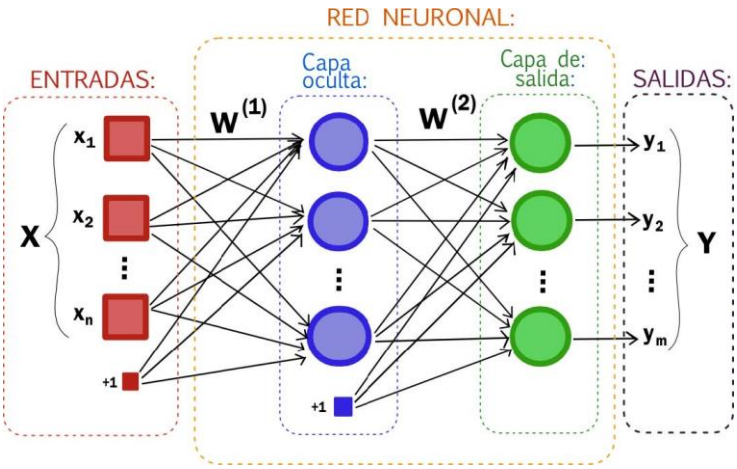


Figura 13: Estructura general de una red neuronal de una sola capa oculta con n entradas y m salidas. Obtenida de García V. G. (2010).

La red es una interconexión de nodos o neuronas de manera arreglada en capas; en la figura 13 observamos únicamente tres capas, pero es importante aclarar que podemos conformar la capa oculta por más de una única capa.

La capa de entrada es la encargada de recibir la información proveniente de las fuentes externas, con base en la imagen, recibe un conjunto X compuesto por $\{x_1, x_2, \dots, x_n\}$.

> La capa intermedia u oculta son internas a la red y no tienen contacto directo con el entorno exterior. El número de niveles ocultos puede estar entre cero y un número elevado. Las neuronas de las capas ocultas pueden estar interconectadas de distintas maneras, lo que determina, junto con su número, las distintas topologías de redes neuronales (Matich, 2001, pág. 19).

La capa de salida es encargada de entregar la información resultante al exterior, denotada como el conjunto Y compuesto por $\{y_1, y_2, \dots, y_n\}$.

La idea es la combinación de los distintos valores de entrada del conjunto X, y un conjunto de pesos sinápticos (representado por W) por cada valor de entrada, para luego mediante una función de entrada definir la importancia relativa de cada entrada, por ejemplo, la sumatoria de los pesos multiplicado por los valores entrada:

$$\sum_j (x_{ij} \times w_{ij})$$

Todo esto es pasado a las siguientes neuronas, las cuales reaccionarán según una función de activación, generalmente binaria de 0 y 1, donde la neurona se activará cuando se cumpla un cierto umbral para finalmente transferirlo y transmitirlo a la capa de salida.

2.4.2 Redes bayesianas ingenuas

Las redes bayesianas ingenuas o Naive Bayes, es una familia de algoritmos probabilísticos que aprovechan la teoría de probabilidades y el teorema de Bayes para predecir la categoría de una muestra. Son probabilísticas, lo que significa que calculan la probabilidad de cada categoría para una muestra dada, y luego producen la categoría con la más alta probabilidad. La forma en que obtienen estas probabilidades es utilizando el teorema de Bayes, que describe la probabilidad de una característica, basada en el conocimiento previo de las condiciones que podrían estar relacionadas con esa característica. El Teorema de Bayes nos dice que:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Donde

- $P(A|B)$ es la probabilidad de hipótesis *A* dado el dato B. Esto se llama la probabilidad posterior.
- $P(B|A)$ es la probabilidad de dato *B* dado que la hipótesis *A* es verdadera.
- $P(A)$ es la probabilidad de que la hipótesis *A* sea verdadera (independientemente de los datos). Esto se conoce como la probabilidad previa de *A*.
- $P(B)$ es la probabilidad de los datos (independientemente de la hipótesis).

Las redes bayesianas son modelos gráficos probabilísticos que nos permiten calcular la probabilidad de un evento, basándose en alguna evidencia dentro del grafo. Las redes bayesianas “naive” utilizan el mismo principio de las redes bayesianas, pero hacen fuertes suposiciones de independencia, es decir,

hipótesis de independencia condicional de las variables predictoras dada la clase dependiente.

Para poder aplicar una correcta clasificación, es importante tomar en cuenta los siguientes puntos:

1. Naive Bayes es un algoritmo de clasificación adecuado para la clasificación binaria y multi-clase. Para este estudio, hablamos de si es o no un posible caso de cáncer de pulmón que requiere mayores exámenes clínicos.
2. Las probabilidades de clase son simplemente la frecuencia de instancias que pertenecen a cada clase dividida por el número total de instancias.
3. Las probabilidades condicionales son la frecuencia de cada valor de atributo para un valor de clase dado, dividido por la frecuencia de instancias con ese valor de clase.
4. Gaussian Naive Bayes. Si las variables de entrada son de valor real, se supone una distribución gaussiana. En este caso el algoritmo funcionará mejor si las distribuciones de sus datos son gaussianas o casi gaussianas. Esto es tan simple como calcular la media (μ) y los valores de desviación estándar (σ) de cada variable de entrada.

2.4.3 Máquinas de soporte vectorial (SVM)

Algoritmo de aprendizaje supervisado que puede ser empleado tanto para la clasificación como para la regresión. SVM son más comúnmente utilizados en los problemas de clasificación y, como tales, en esto se centrará en esta sección.

Las máquinas de soporte vectorial se basan en la idea de encontrar un hiperplano que mejor divide un conjunto de datos en dos clases, donde un hiperplano se puede definir en forma sencilla como una línea recta que separa y clasifica el conjunto de datos, como se muestra en la figura 14.

SVM funciona correlacionando datos a un espacio de características de grandes dimensiones, de forma que los puntos de datos se puedan categorizar, incluso si los datos no se pueden separar linealmente de otro modo. Se detecta un separador entre las categorías, y los datos se transforman de manera que el separador se puede extraer como un hiperplano. Tras ello, las características

de los nuevos datos se pueden utilizar para predecir el grupo al que pertenece el nuevo registro (Funcionamiento de SVM, s.f.).

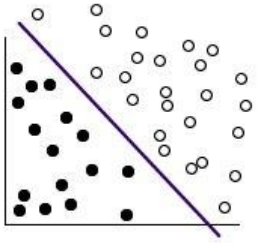


Figura 14: Ejemplo de SVM. Obtenido de Funcionamiento de SVM (s.f.).

2.5 Procesamiento imágenes

El Procesamiento de imágenes es un método para convertir una imagen en forma digital y realizar operaciones en él, con el fin de obtener una imagen mejorada o extraer alguna información útil de ella.

Existen dos tipos de procesamiento de imágenes, el analógico y el digital. Nos centraremos únicamente en este último. Cuando hablamos de procesamiento digital de imágenes (DIP, por sus siglas en inglés), nos referimos a manipular imágenes digitales mediante computadoras y el uso de algoritmos computacionales.

2.5.1 ¿Qué es una imagen digital?

Una imagen digital en un espacio discreto 2D es un conjunto de N filas y M columnas. La intersección de una fila y una columna se le llama píxel (Young, Gerbrands, & Vliet, 1995).

Computacionalmente, podemos interpretar una imagen como una especie de matriz de $N \times M$, donde a su vez $N \times M$ representan la resolución de la imagen digital e incluso el resultado aritmético de $N \times M = cantidad\ de\ pixeles\ en\ la\ imagen$. Cada valor $a[N, M]$ representa un valor de una función $C(x, y, t, \lambda)$ que describe el color, contraste, y otras características individuales de los pixeles que al final conforman toda la imagen.

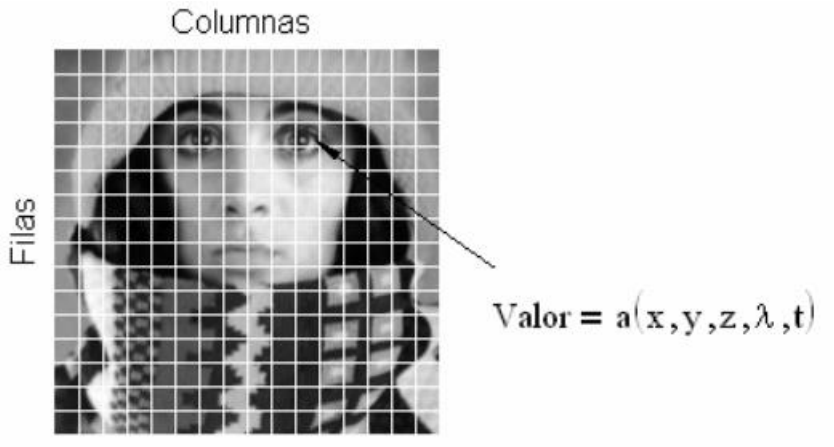


Figura 15: Digitalización de una imagen continua. Obtenido de Young, Gerbrands, & Vliet (1995)

En DPI, existen tres fases generales: preprocesamiento, mejora de imagen y extracción de información.

2.5.2 Preprocesamiento

El preprocesamiento de imágenes es la técnica de mejorar las imágenes de datos antes del procesamiento computacional, que consiste en eliminar el ruido de fondo de baja frecuencia, normalizar la intensidad de las imágenes de partículas individuales (transformaciones de color), eliminar las reflexiones y enmascarar porciones de imágenes (filtrado), entre otras.

2.5.2.1 Recorte de imágenes

Algunas partes irrelevantes de la imagen se pueden quitar y la región de la imagen de interés se centra (Miljković, 2006). El recortar una imagen extrae una región rectangular de interés de la imagen original, donde se enfoca la atención en una porción específica de la imagen, y se descartan ciertas áreas de la imagen que contienen menos información útil.

La técnica general consiste en definir un par de coordenadas (x, y) que determinan las esquinas de la nueva imagen recortada, y al realizar un recorrido en la matriz computacional de la imagen digital se extrae los valores que se encuentran dentro del rango de coordenadas definido anteriormente.

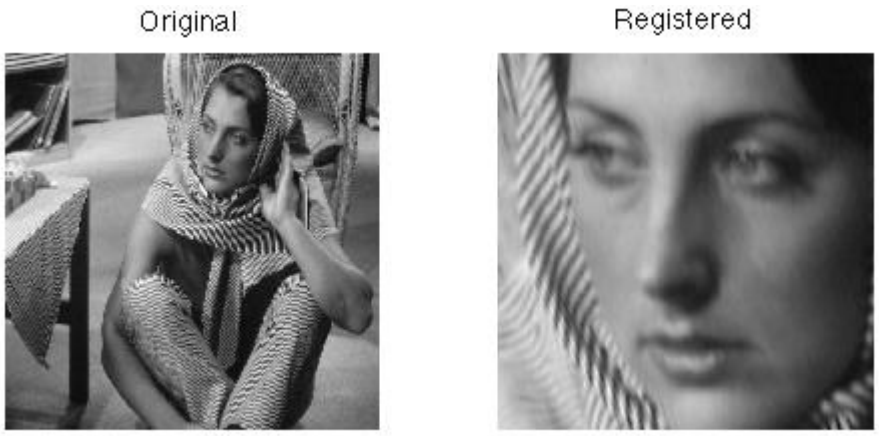


Figura 16: Obtenida de Toolox Image - A Toolbox for General Purpose Image Processing (2008).

2.5.2.2 Filtrado

Se define una matriz filtro o máscara compuesta de una serie de valores, los cuales identificaremos como pesos, y se aplican sobre cada pixel de la imagen. El proceso general consiste en mover la máscara de filtro de punto a punto en una imagen. En cada punto (x, y) de la imagen original, la respuesta de un filtro se calcula mediante una relación predefinida, como la función siguiente general:

$$g(x,y) = \sum_{\alpha=-(k-1)/2}^{(k-1)/2} \sum_{\beta=-(k-1)/2}^{(k-1)/2} f_i(\alpha,\beta)h(x-\alpha,y-\beta)$$

Donde

- k es nuestra matriz máscara (también llamada kernel)
- $g(x,y)$ representa la imagen salida
- x, y, α, β son números enteros

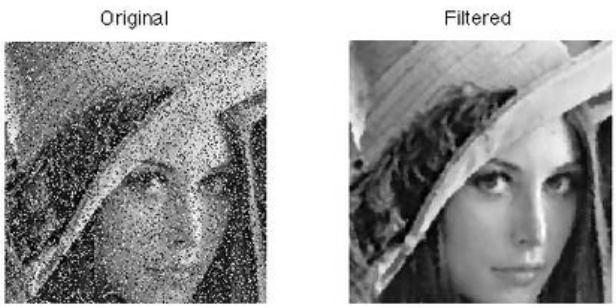


Figura 17: Ejemplo de filtrado. Obtenida de Toolox Image - A Toolbox for General Purpose Image Processing (2008).

2.5.2.3 Ajuste de intensidad

El contraste generalmente se refiere a la diferencia en los valores de luminancia o nivel de grises en una imagen, y es una característica importante. Se puede definir como la relación entre la intensidad máxima y la intensidad mínima sobre una imagen (Kumar, 2013).

Los colores son un formato generalmente de 24bit divido en tres bloques de 8bits cada uno, los cuales representan los colores primarios (generalmente se utiliza el modelo RGB, es decir rojo-verde-azul) en específico y la cantidad o proporción de cada una define un color en específico. Por ejemplo, sabemos tenemos que con base en el modelo RGB:

- El conjunto de bloques de bytes (0, 0, 0) representa el color negro
- (255,255,255) es blanco
- (255,0,0) es rojo
- (0,255,0) es verde
- (0,0,255) es azul
- (128,128,128) es equivalente al gris

La idea para lograr convertir una imagen a escala gris es detectar el color que contribuye más a la imagen para disminuir dicha contribución y aumentar proporcionalmente el otro conjunto de colores. Algorítmicamente sería:

$$Nueva\ imagen\ escala\ gris\ =\ (\ (W_R \times R)\ +\ (W_G \times G)\ +\ (W_B \times B)\ )$$

Donde:

- W_R, W_G, W_B representan el peso a disminuir de la contribución.
- R, G, B representan el color al que se debe multiplicar por contribución.

2.5.3 Mejora de imagen

Conjunto de técnicas para mejorar la calidad de las imágenes para una percepción del ser humano. Al igual que en la etapa de preprocesamiento, algunas de las técnicas en esta fase son filtrado y ajuste de intensidad.

Uno de los procesos utilizados es la segmentación, en el que se subdivide una imagen en un número de regiones uniformemente homogéneas.

Cada región homogénea es una parte u objeto constituyente en toda la escena. En otras palabras, la segmentación de una imagen está definida por un conjunto de regiones que están conectadas y no se superponen, de manera que cada píxel de un segmento de la imagen adquiere una etiqueta de región única que indica la región a la que pertenece. La segmentación es uno de los elementos más importantes en el análisis automatizado de imágenes, principalmente porque en este paso se extraen los objetos u otras entidades de interés de una imagen para su posterior procesamiento, como la descripción y el reconocimiento.

Por ejemplo, en el caso de una imagen aérea que contenga el océano y la tierra, el problema es segmentar la imagen inicialmente en dos partes: el segmento terrestre y el segmento de agua o el océano. A partir de entonces, los objetos en la parte de la tierra de la escena necesitan ser segmentados apropiadamente y posteriormente clasificados (Acharya & Ray, 2005).

Fig. 7.18 (a) An original image, (b) interframe difference image, (c) segmented human object.

Figura 18: Ejemplo de segmentación. Obtenido de Acharya & Ray (2005).

2.5.4 Extracción de información

La etapa de extracción de información consiste en que, mediante los procesos de mejora de imagen y aplicando segmentación, se obtiene una característica o rasgo como el observado en la última imagen de la figura 18. Luego se aplican algoritmos de aprendizaje de máquinas, tanto supervisados como no supervisados, para la toma de decisión y clasificación en las imágenes, donde el conjunto de datos entrada puede ser la aplicación de una función $f(x, y)$, por ejemplo, puede ser la sumatoria de todos los pixeles en el segmento obtenido.

2.6 Big Data

Podemos definir Big Data como un gran volumen, alta velocidad y/o gran variedad de información que requieren nuevas formas de procesamiento, ya que esta información no puede ser procesada o analizada utilizando procesos o herramientas tradicionales, para permitir la toma de decisiones mejores, optimización de procesos, etc.

La cantidad de información suministrada hoy en día ha llegado a cantidades inesperadas. Las organizaciones están topando con retos relacionados a Big Data, se tiene riqueza de información, pero no se sabe cómo sacarle el valor simplemente porque se encuentra en crudo, en formato parcialmente estructurado o no estructurado del todo, y por ende no se sabe qué hacer con esa información.

Entonces cuando hablamos de Big Data se tiene tres características: volumen, variedad y velocidad. Volumen es la cantidad de datos generados y almacenados, el tamaño de estos datos puede determinar si realmente pueden considerarse big data o no. Por el otro lado, variedad es el tipo y la naturaleza de los datos que ayudan a los analistas para obtener una visión resultante efectiva. Velocidad es la rapidez a la que se generan los datos y se procesa para satisfacer las demandas y desafíos que se encuentran en el camino del crecimiento y el desarrollo.

Big Data está disponible a nuestro alrededor en varias formas y tamaños. La comprensión de la relevancia de cada uno de estos conjuntos de datos a las necesidades de negocio es un aspecto clave para tener éxito con las iniciativas de big data. Las siguientes categorías de grandes volúmenes de datos disponibles en la actualidad son:

- Los datos estructurados. Los datos se organizan en filas y columnas y el modelo de metadatos es "nativo" por la estructura. Estas fuentes de datos pueden proporcionar una estructura lógica a través de los metadatos obtenidos fácilmente. Ejemplos son datos de los sensores, datos de la máquina, modelos actuariales, modelos financieros, modelos de riesgo y otros productos de modelos matemáticos.
- Los datos no estructurados incluyen texto, vídeos, audio e imágenes.
- Semiestructurada de datos incluye correo electrónico, informes de ganancias, hojas de cálculo y módulos de software.

2.6.1 Apache Hadoop

Apache Hadoop es un framework que permite el procesamiento distribuido de grandes conjuntos de datos a través de clústeres de computadoras usando modelos de programación sencillos. Está diseñado para escalar de servidores individuales a miles de máquinas, cada una ofreciendo computación y almacenamiento local. En lugar de confiar en el hardware para ofrecer alta disponibilidad, la propia biblioteca está diseñada para detectar y manejar fallos en la capa de aplicación, por lo que ofrece un servicio altamente disponible encima de un grupo de equipos, cada uno de los cuales puede ser propenso a fallas (Foundation, Welcome to Apache™ Hadoop®!, 2014).

Dentro de Hadoop existe un módulo importante: el Hadoop Distributed File System (HDFS). Este es un sistema de archivos distribuido diseñado para ejecutarse en hardware de características básicas. Tiene muchas similitudes con los sistemas de archivos distribuidos existentes. Sin embargo, las diferencias con otros sistemas de archivos distribuidos son significativas. HDFS es altamente tolerante a fallos y está diseñado para ser desplegado en hardware de bajo costo. HDFS proporciona un alto rendimiento de acceso a los datos de aplicación y es adecuado para aplicaciones que tienen grandes conjuntos de datos (Borthakur, 2008).

Un clúster HDFS consta de un único nodo NameNode, un servidor maestro que gestiona el espacio de nombres del sistema de archivos y regula el acceso a los archivos por parte de los clientes. Además, hay un número de DataNodes, por lo general uno por nodo en el clúster, que gestionan el almacenamiento adjunto a los nodos en los que se ejecutan. HDFS expone un espacio de nombres de sistema de archivos y permite almacenar datos de usuario en archivos. Internamente, un archivo se divide en uno o más bloques y estos bloques se almacenan en un conjunto de DataNodes.

El NameNode ejecuta las operaciones de espacio de nombre como abrir, cerrar y cambiar el nombre de archivos y directorios. También determina la asignación de bloques a DataNodes. Los DataNodes son responsables de atender las solicitudes de lectura y escritura de los clientes del sistema de archivos. Los DataNodes también realizan la creación, eliminación y replicación de bloques a partir de la instrucción del NameNode.

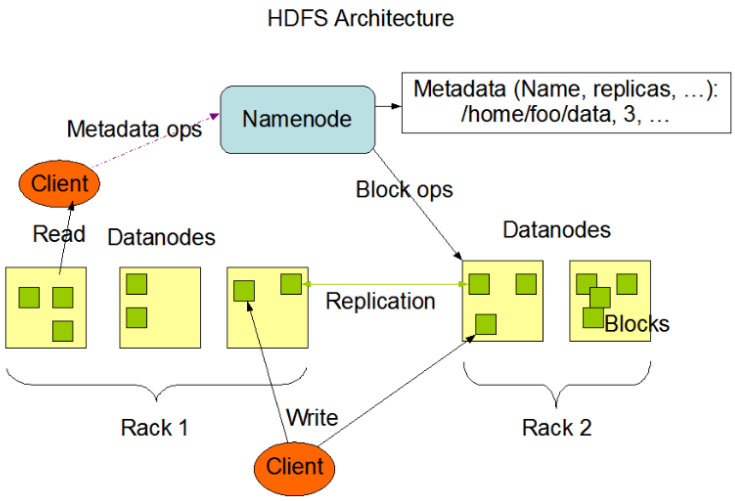


Figura 19: Arquitectura del HDFS. Fuente: Borthakur (2008).

Existen diversas distribuciones comerciales de Hadoop, por ejemplo, tenemos Cloudera, Hortonworks y HDInsights de Microsoft. Cada una utiliza el Apache Hadoop, pero configurado a manera de poder simplificar muchas tareas de los usuarios, desde la administración del ecosistema de Hadoop (creación de usuarios, operaciones de lectura y escritura sobre HDFS, monitoreo de los clústeres, entre otras) hasta integración con otras herramientas propias de la empresa, como en el caso de Microsoft con la combinación de Azure y sus demás tecnologías propietarias.

2.6.2 MapReduce

Hadoop MapReduce es un framework de software para escribir fácilmente aplicaciones que procesan grandes cantidades de datos (conjuntos de datos de varios terabytes) en paralelo en grandes clústeres (miles de nodos) de hardware común y corriente y tolerante a fallos (Foundation, MapReduce Tutorial, 2008).

MapReduce trabaja junto con el HDFS y otro módulo llamado YARN (*Yet Another Resource Manager*, por sus siglas en inglés), el cual no es importante para esta investigación, para poder acceder a los archivos y proveer procesamiento de manera distribuida mediante el uso de lenguaje de programación Java. Está compuesto por las siguientes fases:

- Map: Los datos son divididos en pequeñas piezas. De cada división se pasan a una función de asignación o mapeo para

producir valores de salida conformados en tuplas de una llave K y un valor V.

- Shuffle: Consume lo obtenido de la fase anterior, consolidando todas las salidas en función de la llave K.
- Reduce: La etapa final, consiste en la agregación de todos los valores de salida de la fase anterior en función de la llave K.

Para ejemplificar a manera gráfica cómo funciona, tomemos la figura 20, la cual ejemplifica el ejercicio clásico del contador de palabras. Este es un ejercicio donde se da como entrada un texto o conjunto de palabras para luego devolver como salida la cantidad de apariciones de cada una de las palabras en el texto entrada:

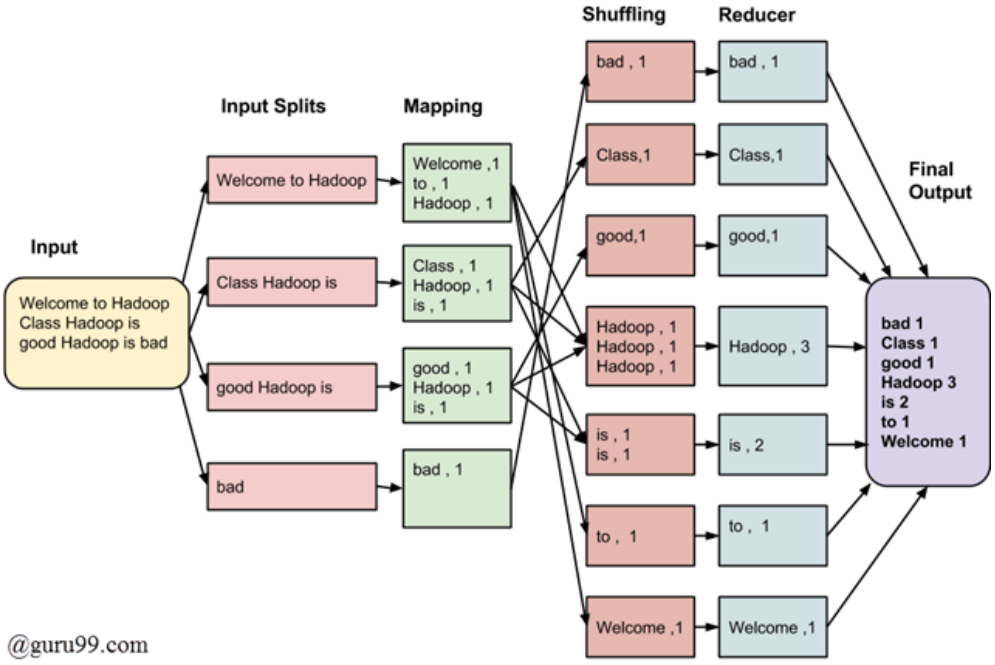


Figura 20: Obtenida de Guru99 (s.f.).

Observamos que:

- Nuestra entrada (Input) es un texto.
- Dividimos nuestro texto en cuatro pequeñas divisiones (Input Splits) para iniciar la primera fase, la de Map, donde tenemos cuatro bloques de mapeo independientes.

- La fase de Map tiene como función el agrupar todas las palabras de entrada y contar todas sus apariciones en los diversos bloques de mapeo. Podemos ver, por ejemplo, que para el primer bloque de Map, cada una de las palabras encontradas solo aparecen una única vez. Además, se conforma la tupla llave <palabra> con el valor <cantidad de veces la palabra aparece>.
- La fase de Shuffle tiene como objetivo el consolidar todos aquellos resultados obtenidos de los diversos bloques de mapeo por la llave <palabra>. Entonces podemos observar que la palabra “Hadoop” tiene 3 apariciones, la palabra “is” únicamente 2, mientras que el resto aparecen una sola vez.
- La fase final de Reduce se encarga de sumar todos los valores de las tuplas consolidadas en la fase anterior para luego proveer el resultado final: la palabra y cuántas veces apareció en el texto.

2.6.3 Apache Spark

Apache Spark es una plataforma de computación en clúster diseñada para ser rápida y de uso general (Karau, Kowinski, & Zaharia, 2015).

Es una solución que se puede ejecutar junto con ambientes de Hadoop o como autónomo, y que viene a solventar las debilidades del paradigma MapReduce en:

- Procesamiento de flujo, es decir, el continuo procesamiento de datos recibidos. Por ejemplo, procesar los tweets de Twitter referentes a la organización.
- Biblioteca para aprendizaje de máquinas llamada MLib. Es posible programar algoritmos complejos de aprendizaje de máquinas.
- Integración con otros frameworks del ecosistema de Hadoop, como Hive e Impala, que son dos soluciones para “simular una base de datos.”
- Posibilidad de realizar consultas SQL en vez de código de programación.
- Nos permite utilizar código Java, Scala o Python.

- La capacidad de ejecutar computación en memoria, a diferencia del MapReduce tradicional que es computación en disco. Esto hace que Spark sea más rápido.

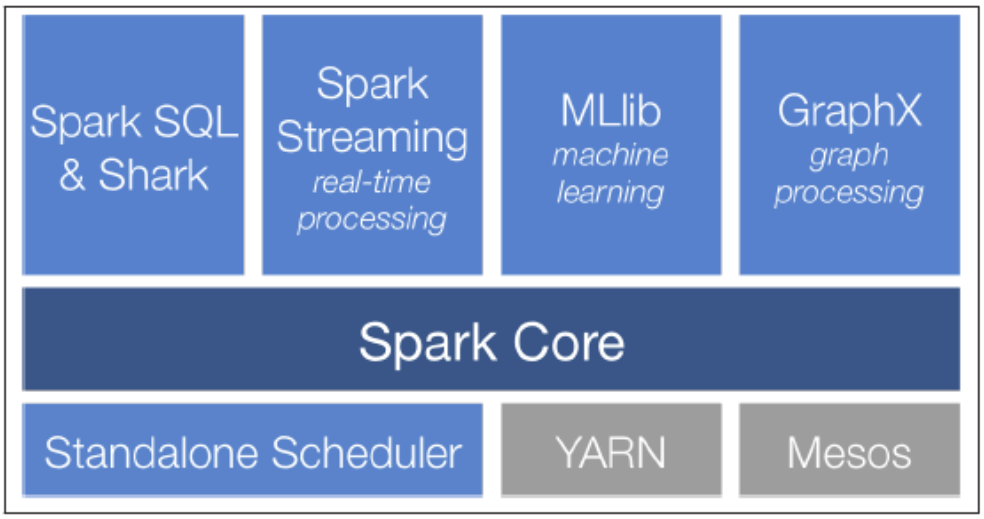


Figura 21: El conjunto de Spark. Imagen obtenida de Karau, Kowinski, & Zaharia (2015).

Capítulo 3. Marco Metodológico

En el presente capítulo se detalla la metodología empleada en la investigación, abarcando temas como: el diseño y el enfoque de la investigación, detalles de la investigación formal realizada a través de las fuentes de información utilizadas y las técnicas de recopilación de datos, entre otros puntos importantes a recalcar para el éxito del trabajo investigativo.

3.1 Tipo de Investigación

El tipo de investigación que se utiliza para el proyecto es evaluativo porque se ajusta a las características del proyecto, debido a que no se atiende necesidades de clientes particulares. Sí se pretende, a través de la investigación, emitir un criterio comparativo entre la solución propuesta de detección de cáncer de pulmón mediante Big Data versus otras soluciones expuestas en los diversos estudios seleccionados en el capítulo 1.

La investigación evaluativa es “la recolección sistemática de información acerca de actividades, características y resultados de programas, para realizar juicios acerca del programa, mejorar su efectividad, o informar la futura toma de decisiones.” (Patton, 1996).

> La evaluación es el proceso de identificar, obtener y proporcionar información útil y descriptiva acerca del valor y el mérito de las metas, la planificación, la realización y el impacto de un objeto determinado con el fin de servir de guía para la toma de decisiones, solucionar los problemas de responsabilidad y promover la compresión de los fenómenos implicados. (Stufflebeam & Shinkfield, 1987, pág. 183).

Debido a que nuestro trabajo está centrado en el área de salud, según Ayres (2004), la evaluación en salud es “un conjunto de procedimientos sistemáticos que buscan hacer visible lo que se hace, con referencia a lo que se pretende hacer; respecto a intereses, efectividad, operatividad y calidad de las acciones, tecnologías, servicios o programas de salud.” (pág. 585)

La investigación evaluativa es de gran ayuda para el presente proyecto ya que se busca exponer un enfoque en comparación a otras ideas expuestas, que permite asistir en la detección de posibles casos de cáncer de pulmón de una forma masiva. La finalidad es proveer a los centros médicos la posibilidad

de disminuir las listas de espera en resultados, disminuir los costos financieros tanto para el paciente o los pacientes y el centro médico en la necesidad de realizar exámenes más detallados cuando no es necesario, en infraestructura tecnológica que permita alta disponibilidad de los datos y a un bajo costo mediante sistemas de cómputo comunes.

3.2 Alcance Investigativo

El alcance investigativo será principalmente de tipo exploratorio, ya que la literatura indica que existen varias teorías que aplican al tema de estudio respecto a asistencia computarizada en detección de cáncer, pero ninguna pensada en aplicación de tecnologías de Big Data como lo es Hadoop.

> Los estudios exploratorios se realizan cuando el objetivo es examinar un tema o problema de investigación poco estudiado, del cual se tienen muchas dudas o no se ha abordado antes. Es decir, cuando la revisión de la literatura reveló que tan sólo hay guías no investigadas e ideas vagamente relacionadas con el problema de estudio, o bien, si deseamos indagar sobre temas y áreas desde nuevas perspectivas. (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014, pág. 91).

También se plantea un alcance investigativo descriptivo, ya que se selecciona un conjunto de casos de cáncer pulmón y se mide o recolecta información sobre ellos, para mostrar con precisión las dimensiones de una solución de detección.

> Con los estudios descriptivos se busca especificar las propiedades, las características y los perfiles de personas, grupos, comunidades, procesos, objetos o cualquier otro fenómeno que se someta a un análisis. Es decir, únicamente pretenden medir o recoger información de manera independiente o conjunta sobre los conceptos o las variables a las que se refieren, esto es, su objetivo no es indicar cómo se relacionan éstas. (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014, pág. 92).

3.3 Enfoque

El tipo de enfoque adoptado en el proyecto es mixto, es decir, una combinación del enfoque cuantitativo y cualitativo.

El enfoque cuantitativo consiste en “la recolección de datos para probar hipótesis, con base en la medición numérica y el análisis estadístico, para establecer patrones de comportamiento y probar teorías.” (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014). La recolección de las imágenes médicas, y la comparación estadística entre los resultados estadísticos (precisión, sensibilidad y especificidad, entre otros) obtenidos de este trabajo, en comparación con los expuestos por otros autores, principalmente los mencionados en nuestra sección de selección de estudios, son aspectos de nuestro enfoque cuantitativo.

> El enfoque cualitativo en lugar de que la claridad sobre las preguntas de investigación e hipótesis preceda a la recolección y el análisis de los datos (como en la mayoría de los estudios cuantitativos), los estudios cualitativos pueden desarrollar preguntas e hipótesis antes, durante o después de la recolección y el análisis de los datos. (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 2014, pág. 7).

Se va a utilizar el enfoque cualitativo para realizar un proceso inductivo, esto consiste explorar y describir qué tan acertada es esta propuesta de investigación para la ayuda en asistencia médica computarizada.

3.4 Diseño

El diseño será de tipo experimental para poder medir nuestra propuesta en comparación con otros.

En primera instancia es necesaria la recolección de un grupo de imágenes de TAC que muestren cáncer de pulmón, principalmente de células no pequeñas, e imágenes sin cáncer de pulmón para el análisis de datos que se van evaluar.

Una vez que se ha identificado el conjunto de datos de entrada, se procesa mediante técnicas de preprocesamiento de imágenes, para luego realizar técnicas de procesamiento de imágenes a las cuales al final se les aplica técnicas de clasificación y aprendizaje de máquinas para la toma de

decisiones. Así se podrá predecir si se está ante un caso de cáncer pulmonar o no.

Con toda la información recolectada, se analizará las métricas obtenidas de una matriz de confusión respecto a los resultados y poder comparar qué tan efectiva es la propuesta ante las expuestas en los estudios seleccionados.

3.5 Población y Muestreo

El conjunto de datos es discrecional; corresponderá a imágenes de cáncer pulmonar confirmado, casos posibles e imágenes donde no hay presencia alguna de cáncer, para poder tener una población completa de estudio.

3.6 Instrumentos de Recolección de Datos

Las imágenes por utilizar serán obtenidas del sitio de acceso público “The Cancer Imaging Archive”, el cual es una fuente de base de datos del consorcio de imágenes médicas. Allí podemos encontrar más de 2GB de muestras, suficientes para nuestra implementación con Big Data.

La selección de las imágenes será con aquellas de la categoría CT en el sitio web, es decir, tomografía computarizada por sus siglas en inglés, y que correspondan con casos de cáncer de pulmón. Se buscará otro conjunto de imágenes que correspondan a casos sin cáncer de pulmón, para poder tener una población completa entre posibles casos y casos detectados verídicos para el estudio.

3.7 Técnicas de Análisis de Información

Se utilizará el diagrama Ishikawa para mostrar las reglas básicas de cómo se ha generalizado las ideas de asistencia computarizada de detección de cáncer de pulmón expuestas en las literaturas seleccionadas en el primer capítulo. Con esta técnica se trata de utilizar una serie de conceptos, y se identifica cómo podemos generar una predicción errónea de un posible cáncer de pulmón.

En la figura 22 (Esquema de causa y efecto) se muestra el flujo para la evaluación. Las causas son el medio del por qué no se están realizando clasificaciones correctas o qué se dejó de hacer; por medio de estas causas se

logra el efecto adecuado, el efecto es lo que se debe mejorar para acertar en la detección con un porcentaje de predicción alto, es decir, mayor a 80%.

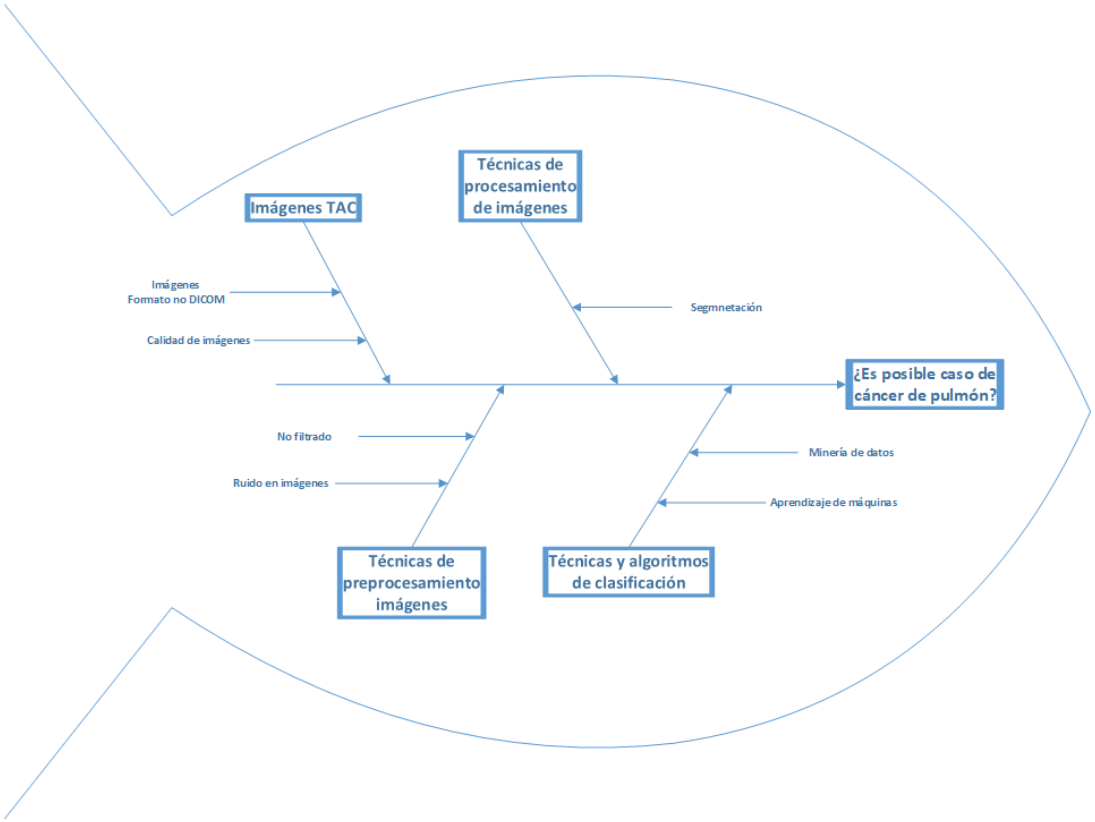


Figura 22: Esquema de causa y efecto.

Fuente: Elaboración propia.

Capítulo 4. Análisis del Diagnóstico

A continuación se realizará un análisis de los datos obtenidos en la sección 3.6. En este caso específico, las imágenes médicas de pulmón obtenidas a través de una tomografía computarizada y que se encuentran en el formato DICOM, el cual es la extensión de archivos utilizado en imágenes médicas obtenidas por cualquier instrumento de estudio. Dicho estudio de las imágenes consiste en la aplicación del diagrama de expuesto en la última sección del capítulo tres.

Se obtiene un alto volumen de imágenes mayor a 1GB de memoria, y en aplicación de los siguientes factores se determina el cumplimiento y validez de la información recolectada y que será utilizada en el desarrollo de la herramienta de software.

4.1 Formato de las imágenes

"Mediante la utilización del programa gratuito MicroDicom se comprueba que cada una de las imágenes se encuentra en formato DICOM ya que son posibles de abrir en dicho programa" (Visualizador de imágenes médicas [Software], 2017). Además, MicroDicom nos permite observar los atributos encriptados en el archivo, con lo cual podemos comprobar que:

- La imagen corresponde a una obtenida mediante TAC.
- La cantidad de bites por pixel equivale a 16, lo que significa una alta calidad de la imagen.
- La no presencia de información personal específica de un paciente almacenada y encriptada en la imagen DICOM, que pueda incurrir en algún grado de violación de privacidad por la no autorización de las imágenes por parte de la página o del investigador del presente trabajo.
- Las imágenes están en escala de gris, cualquier otro estilo como el RGB no es aceptado. Esto permitirá la aplicación de las técnicas de procesamiento de imágenes correcta para el estudio, ya que con imágenes que no sean escala de gris hay un aumento de la complejidad computacional en la transformación de las imágenes a la escala correcta. Esto puede derivar en una pérdida de calidad a nivel de bytes en la imagen y a su vez impactar la implementación de la herramienta

para detección de cáncer pulmonar, debido a mejores y más complejos algoritmos para la segmentación de las imágenes.

4.2 Imágenes de pulmones

Gracias a la ayuda del estudiante de medicina de la Universidad Internacional de las Américas (UIA), Mauricio Guzmán Obando (Obando, 2017), se comprueba que todas las imágenes corresponden a tomografías realizadas a los pulmones.

4.3 Validación de los casos de neoplasias

El conjunto de datos obtenido consiste en imágenes con posibles casos de cáncer de pulmón y casos donde se tiene pulmones sanos, todas obtenidas del mismo sitio especificado en secciones anteriores. Pero para corroborarlo, se recurre de nuevo a la ayuda del estudiante Mauricio Obando (Obando, 2017), su cooperación permite identificar los posibles tumores o nódulos cancerígenos en los pulmones. Esto ayudará a poder definir el conjunto de datos de entrenamiento y de igual manera el algoritmo o los algoritmos de aprendizaje de máquinas y/o minería de datos a utilizar para una alta probabilidad de detección de cáncer de pulmón en imágenes TAC que cumpla con los objetivos estipulados en este trabajo investigativo.

Capítulo 5. Propuesta de Solución

5.1 Lenguaje de programación y bibliotecas a utilizar

Debido a que nuestra herramienta gira en torno a los frameworks de Apache Spark y Apache Hadoop, el lenguaje de programación a utilizar es Java, ya que, en comparación a las otras posibilidades de Python y Scala, el autor tiene mayor experiencia.

Otra razón de la selección de Java tiene origen en la necesidad de poder leer imágenes médicas en formato DICOM, para lo cual se propone la utilización de ImageJ, la cual es una biblioteca gratuita para Java, disponible en extensión jar, para leer diversos formatos de imágenes (jpeg, png, tiff y DICOM, entre muchas otras) y a la vez provee operaciones básicas a nivel de pixeles en imágenes. Sumado a ello, es ampliamente utilizado en soluciones visuales como Fiji, por lo tanto para muchos de los algoritmos de procesamiento de datos a implementar ya se encuentran y solamente es necesario adaptarlo a las necesidades del proyecto, cumpliendo con los derechos de autor y las reglas estipuladas en GNU General Lesser Public License. ImageJ puede ser descargada en ImageJ - Image Processing and Analysis in Java [Software] (s.f.).

5.2 Metodología de detección a aplicar

La Figura 23 muestra la propuesta del autor para la detección de cáncer de pulmón en imágenes de TAC.

Figura 23: Enfoque de detección.

Fuente: Elaboración propia.

5.2.1 Preprocesamiento de la imagen

5.2.1.1 Filtro de la mediana

Para poder eliminar el ruido que podemos encontrar en las imágenes, se escoge la aplicación del filtro de la mediana con base en un matriz de tamaño 3x3. Esta es una técnica de procesamiento donde, a cada conjunto de tamaño nxn, en este caso como hemos mencionado va a ser de 3x3, en la imagen se calcula la mediana de todos los valores vecinos al píxel central de la matriz para luego reemplazar dicho centro con el valor obtenido. De esta manera se elimina ruido impulsional (sal y pimienta), pero con la desventaja de que podemos perder un poco de detalles en la imagen.

Las Figuras 24 y 25 muestran un ejemplo de cómo funciona el filtrado en una imagen.

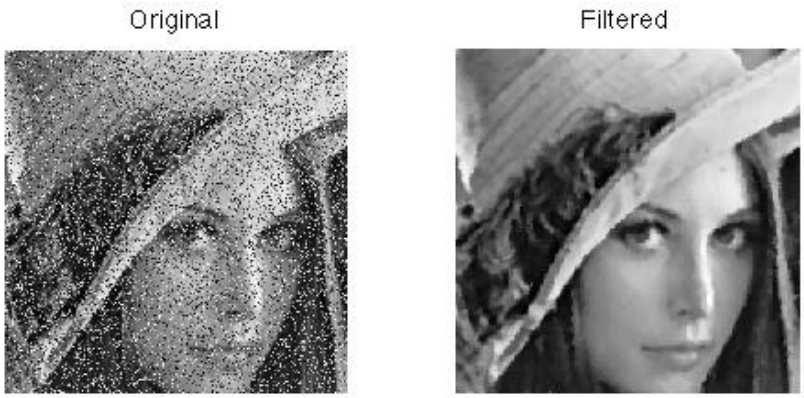


Figura 24: Filtrado de mediana. Obtenido de median filter in AWK (2015).

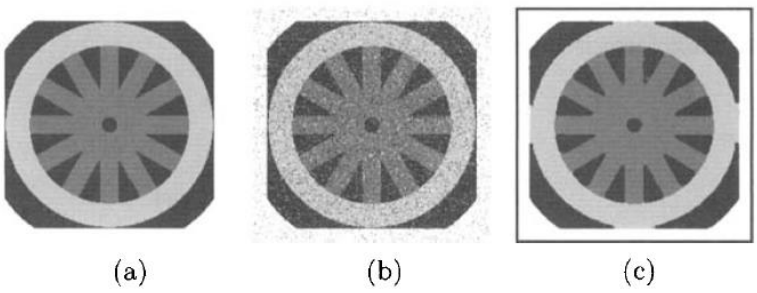


Fig. 6.2 Result of median filtering: (a) original image, (b) salt and pepper noisy image, (c) result of median filtering.

Figura 25: Filtrado de mediana. Obtenido de Acharya & Ray (2005).

5.2.1.2 Filtro de Gabor

En las literaturas seleccionadas y que se detallaron en la sección Estado de la Cuestión del capítulo uno, todos mencionan la necesidad e

implementación del filtro de Gabor, razón por la cual se hace el estudio pertinente y se decide incorporar a la solución que se busca desarrollar en el presente trabajo.

Definimos el filtro de Gabor como un filtro de tipo lineal utilizado para detección de texturas y de bordes en las imágenes que sigue la función sinusoidal mostrada en la Figura 26, en el caso específico 2-D, debido a que ese es el tipo de entrada para la generación y aplicación en busca de mejorar nuestra imagen médica a analizar. Es importante mencionar que la matriz del filtro o kernel generado debe ser de tamaño 3x3.

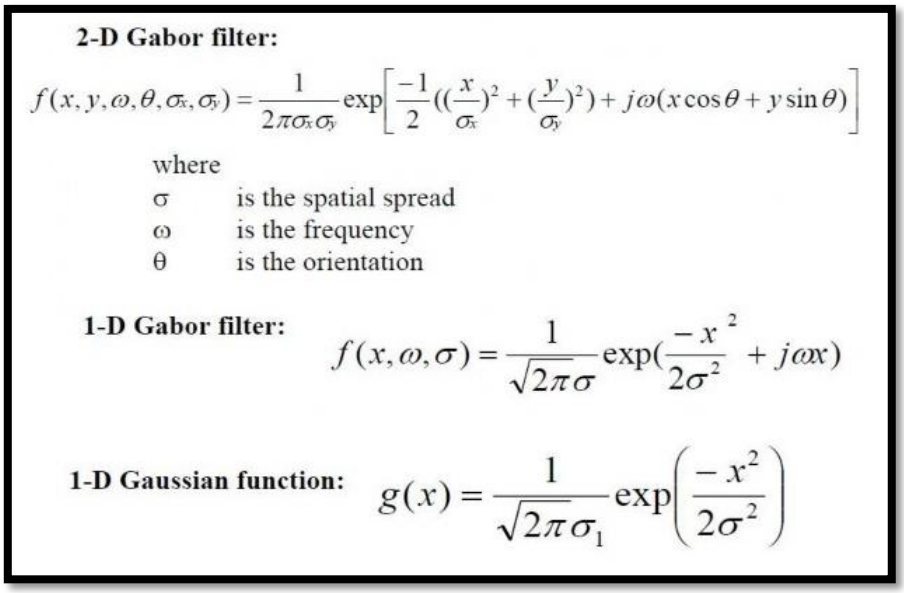


Figura 26: Fórmula para generación de filtro de Gabor. Obtenida de Image Processing Toolkit (s.f.).

Se define cada uno de los parámetros de la función con los siguientes valores con base en que se realiza un estudio de histograma de pixeles en las imágenes DICOM obtenidas a fin de encontrar los mejores valores en un patrón repetible:

x, y: $representación\ de\ la\ coordenada\ del\ pixel\ en\ la\ imagen\ original$

$\omega = 1$

θ: $se\ aplican\ varias\ orientaciones, 0, 45, 90\ y\ 135\ grados\ respectivamente$

$\sigma = 0.5$

La Figura 27 muestra la aplicación de la fórmula de Gabor definida anteriormente en una de las imágenes obtenidas y el resultado del filtro de

tamaño 3x3 en dicha imagen, donde se muestra claramente la resolución y mejora en bordes y texturas.

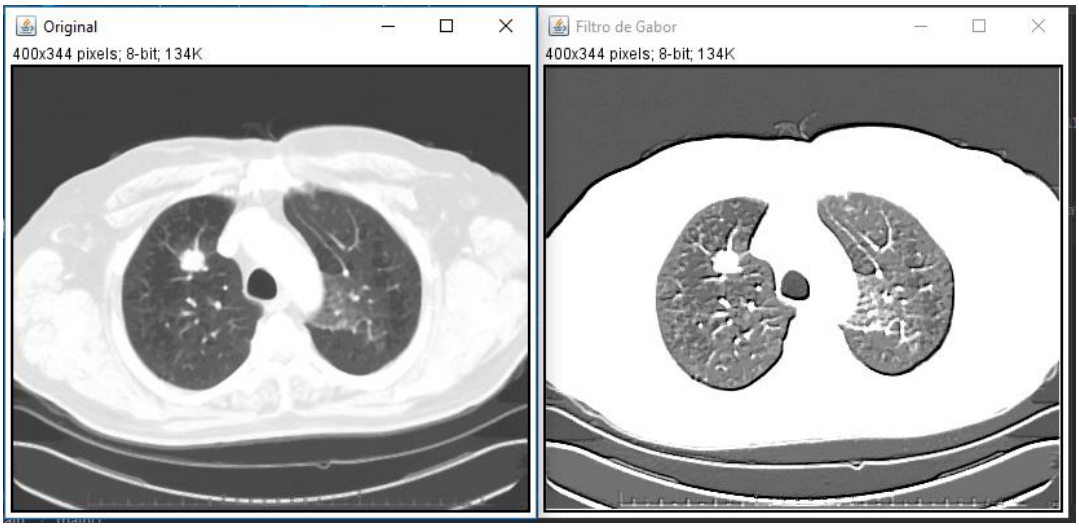


Figura 27: Aplicación de filtro de Gabor.

Fuente: Elaboración propia.

5.2.2 Segmentación

5.2.2.1 Segmentación por método del valor de umbral

Conocido en inglés como “Threshold”.

> Las técnicas de umbral de nivel de gris son métodos computacionalmente baratos para dividir una imagen digital en regiones mutuamente exclusivas y exhaustivas. La operación de umbral implica la identificación de un conjunto de umbrales de optimidad, basados en los cuales la imagen se divide en varias regiones significativas. (Acharya & Ray, 2005, pág. 143).

Es decir, cada pixel es transformado de un valor a otro a fin de facilitar la partición en áreas y grupos con significados que ayuden en la definición de un criterio de homogeneidad en detección de regiones de interés en la imagen.

Debido a que nuestras imágenes se encuentran en escala de grises, es necesario poder convertirlas en binario, es decir, en pixeles de valores de 0 y 1. Además, debido a la necesidad computacional que requerimos para la aplicación de procesamiento en paralelo en almacenamiento distribuido, las técnicas de valor de umbral nos permiten convertir no solo las imágenes en escalas binarias sino también disminuir cada 16 bits de los pixeles en la imagen

a simplemente 8 bits, de manera que es mucho más sencillo el tratamiento en la imagen de TAC.

Pero no solo se escogen por las razones anteriores estas técnicas como método de segmentación, sino también que, debido a la regionalización significativa, podemos eliminar de la imagen todos aquellos atributos no requeridos como lo son las venas y arterias en la imagen o pequeños nódulos respiratorios. De este modo se obtiene una imagen mucho más limpia a procesar únicamente con la presencia de los posibles tumores neoplásicos.

Se realiza un exhaustiva prueba y análisis de diferentes métodos de segmentación, la Figura 28 es muestra de los resultados obtenidos al aplicar cada una de las variantes a una imagen.

Figura 28: Aplicación de diversos métodos de segmentación por valor umbral

Fuente: Elaboración propia.

Para el trabajo investigativo, se decide utilizar el método de umbralización propuesto por Nobuyuki Otsu, conocido como el método del

valor de umbral de Otsu (Otsu's Threshold en inglés), debido a que es donde se observan mejores resultados en el proceso de segmentación de la imagen.

> Se basa en el principio de que el nivel de grises para el cual la varianza entre clases es máxima se selecciona como umbral. Para un nivel de gris k todo el valor de grises 5k formará una clase (Co) y todos los demás formarán una clase diferente (Cl). Seleccione k como umbral para el cual la varianza de clase V(k) es máxima. El criterio propuesto por Otsu maximiza la varianza entre clases de la intensidad de píxeles. (Acharya & Ray, 2005, pág. 144).

La Figura 29 muestra cómo la herramienta convierte una imagen TAC correspondiente a pulmones en niveles de grises a una imagen binaria mediante el método de Otsu.

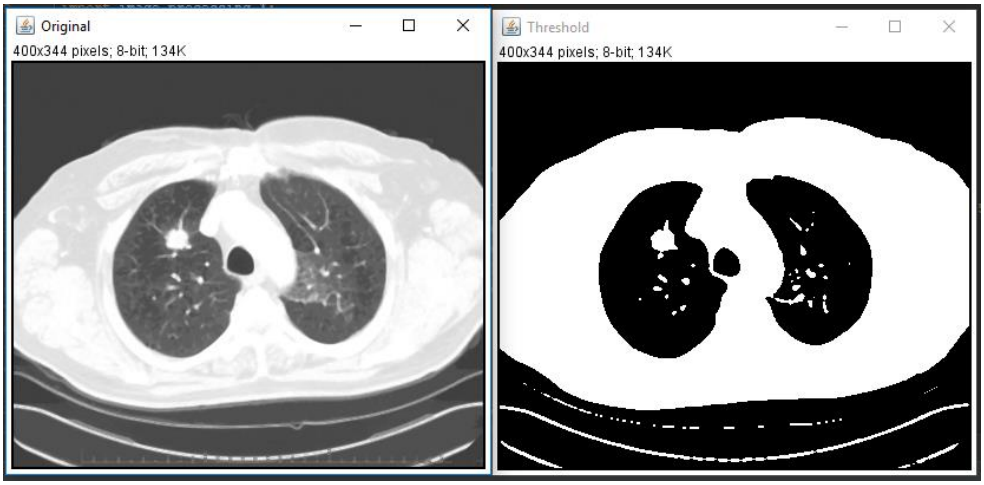


Figura 29: Aplicación del método Otsu.
Fuente: Elaboración propia.

5.2.2.2 Operación morfológica de apertura

Las operaciones morfológicas son una herramienta de extracción de componentes en imágenes "que se utiliza para la representación y descripción de regiones." (Montes & Castillo, 1996).

> Las imágenes binarias pueden contener numerosas imperfecciones. En particular, las regiones binarias producidas por la umbralización son distorsionadas por el ruido y la textura. El procesamiento de imágenes morfológicas persigue los objetivos de eliminar estas imperfecciones, explicando la forma y

estructura de la imagen. Estas técnicas pueden extenderse a imágenes en escala de grises. (Morphological Image Processing, s.f., parr. 1-2).

Dentro de dichas operaciones se tiene la apertura, con la cual mediante la erosión de la imagen y dilatación de imágenes se “tiene un efecto suavizante sobre la forma inicial X, cortando las prolongaciones estrechas y suprimiendo las partes pequeñas aisladas. Todo ello al precio de perder detalles que poseía el conjunto original” (Montes & Castillo, 1996). De esta manera podemos suprimir aquellas pequeñas imperfecciones en la imagen, y se obtiene una mayor limpieza en la imagen a procesar en el proceso de detección de cáncer en pulmones.

La Figura 30 muestra el resultado de la aplicación de la operación de apertura en la metodología de detección propuesta, luego de haber aplicado los pasos de preprocesamiento mencionados anteriormente en la imagen original de la izquierda.

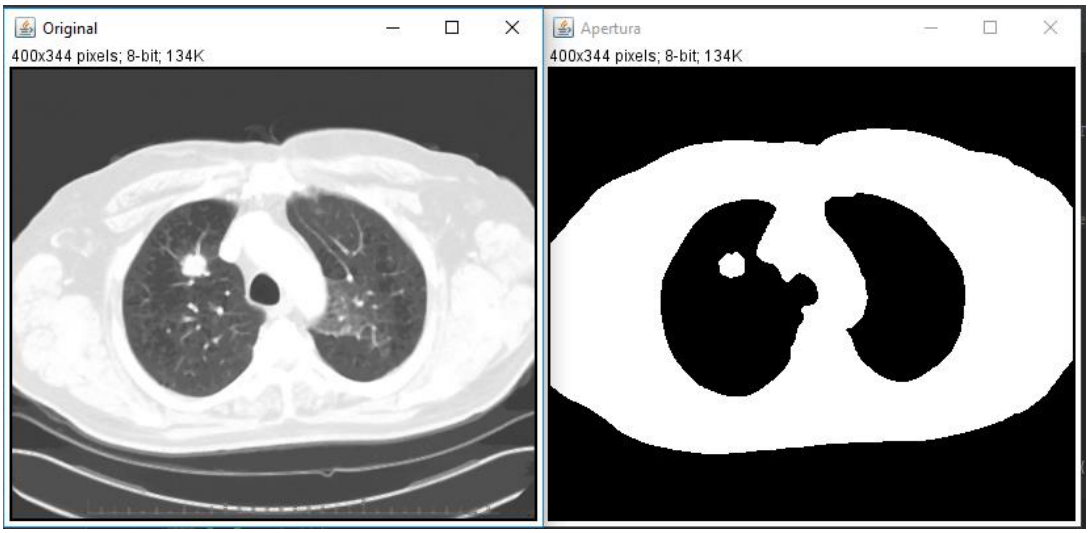


Figura 30: Operación de apertura.

Fuente: Elaboración propia.

5.2.2.3 Eliminación de bordes

La eliminación de bordes es otra operación morfológica que se aplica con la finalidad de poder obtener únicamente todos aquellos puntos blancos dentro de los pulmones que son posibles casos de nódulos cancerígenos.

La Figura 31 muestra la implementación de la eliminación de bordes en comparación a una imagen original, luego de la aplicación de los pasos predecesores. Como se puede observar, el círculo blanco resultante de la

operación morfológica corresponde a un nódulo de cáncer pulmonar en la imagen.

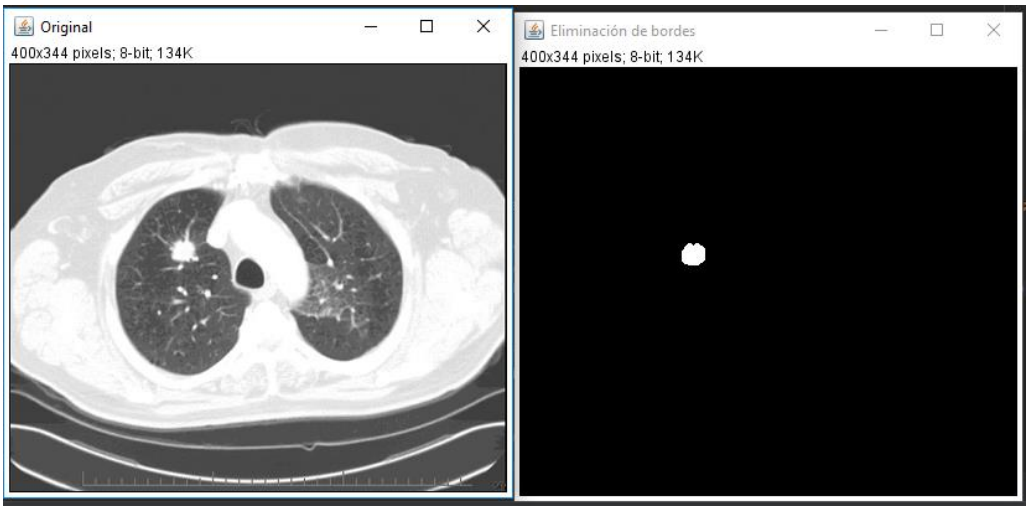


Figura 31: Eliminación de bordes.

Fuente: Elaboración propia.

5.2.3 Detección de cáncer mediante Inteligencia Artificial (IA)

En este punto, se considera importante retomar la idea expresada anteriormente de que se trata de una clasificación binaria, se determina si es o no un posible caso de cáncer de pulmón que requiere mayores exámenes clínicos. Por ello, se utilizará el método de redes bayesianas como la técnica de clasificación, encargado de verificar cada uno de los puntos blancos que se obtendrán luego de aplicar la operación morfológica de eliminación de bordes, y de asignar si se está ante un posible caso de tumor pulmonar o no.

Para poder satisfacer los objetivos que se busca lograr con el análisis de los datos, se sigue la metodología de CRISP-DM, especificada en el capítulo dos del marco conceptual del presente trabajo. La metodología de CRISP-DM permite estructurar no solo en el proceso de análisis y post-análisis, sino también los procesos anteriores a estos, como lo es la definición del conjunto de datos de entrenamiento.

Cabe resaltar que la implementación tanto del modelo de minería de datos sugerido como algún otro aplicable debe ser utilizando la biblioteca MLib, provista en Spark. De esta manera se evita el desarrollo del mismo, pero a la vez se asegura procesamiento paralelo en grandes volúmenes de datos, lo cual es una de las finalidades de este trabajo investigativo.

5.2.3.1 Comprensión del negocio

Para este punto de la metodología no existe un cliente específico ya que no es una investigación aplicativa, por lo que el término “negocio” se usa libremente desde el punto de vista de identificación de objetivos, para encajar dentro del modelo más amplio tomado de referencia.

Se tiene un conjunto de datos obtenidos del procesamiento de imágenes de TAC referentes a pulmones donde se registra las características de posibles tumores cancerígenos. El problema es: mediante todas las muestras obtenidas, realizar un estudio que nos permita detectar aquellas que presentan signos de tumores de tipo benigno o maligno que nos dan señales de cáncer.

5.2.3.2 Comprensión de los datos

De cada resultado obtenido luego de la eliminación de bordes en las imágenes procesadas, se extrae las siguientes variables numéricas continuas, equivalentes a características de los tumores, utilizando la biblioteca gratuita “IJBlob”, explicada con mayor detalle en Wagner & Lipinski (2013):

- Área: Generalmente definido como “el número total de píxeles en un área extraída.” (Gajdhane & L.M., 2014).
- Perímetro: “El perímetro del contorno exterior de un objeto” (Wagner & Lipinski, 2013), es decir, “el largo de contorno de la región de interés.” (Gajdhane & L.M., 2014).
- Área del casco convexo (*Area Convex Hull* en inglés).
- Casco convexo del perímetro (*Perimeter Convex Hull* en inglés).
- Circularidad: Sinónimo del atributo excentricidad mencionado en la literatura fuente definida en capítulos anteriores como “la forma o circularidad de un objeto.” (Gajdhane & L.M., 2014). Se formula como el perímetro al cuadrado divido por el área.
- Alargamiento.
- Convexidad.

5.2.3.3 Preparación de los datos

Mediante el conjunto de datos, se prosigue a la clasificación de todas las muestras agrupadas entre las que tienen presencia de cáncer y las que no.

El definir los diferentes parámetros de un modelo estadístico requiere de un grado de análisis detallado según sea el algoritmo de aprendizaje de máquina a utilizar. Dos de los parámetros a definir cuando se realizan algoritmos supervisados son los conjuntos de datos de entrenamiento y pruebas, la incógnita a resolver siempre es qué tan grande o cuál es el tamaño correcto que asegura un buen resultado estadístico, una mayor imparcialidad de los datos.

En proyectos de inteligencia artificial, la mejor práctica para definir los conjuntos de datos entrenamiento y pruebas es la técnica de validación cruzada (cross-validation en inglés). Esta permite eliminar el sobreajuste, que es cuando los datos de entrenamiento son muy pequeños o se está ante un modelo con un gran número de parámetros a definir. Además, “la validación cruzada es una manera de predecir el ajuste de un modelo a un hipotético conjunto de datos de prueba cuando no disponemos del conjunto explícito de datos de prueba” (Refaeilzadeh, Tang, & Lui, 2008).

El método de validación 70:30 es la forma más convencional y sencilla de hacer la tarea de definir los conjuntos de datos entrenamiento y pruebas. Simplemente se define de manera aleatoria el conjunto de entrenamiento compuesto por el 70% de los datos y el restante 30% conforma el conjunto de pruebas.

Para Andrew (2012), existe una razón para utilizar una proporción 70:30.

> Muy a menudo, la proporción elegida es del 70% para el conjunto de entrenamiento y del 30% para la prueba. La idea es que una mayor cantidad de datos de capacitación es algo bueno porque hace que el modelo de clasificación sea mejor, mientras que más datos de prueba hacen que el error estimado sea más preciso (Andrew, 2012, parr. 4-6).

Una de las razones principales para utilizar la validación cruzada en lugar de usar la validación convencional 70:30 la explica Grossman, Seni, Elder, Agarwal, & Liu (2010).

> El error en el conjunto de entrenamiento no es un estimador útil del rendimiento del modelo. Lo que se necesita es una forma de estimar el riesgo de predicción, también llamado riesgo de prueba o riesgo futuro. Si no hay suficientes datos disponibles para dividirlos en conjuntos de entrenamiento y prueba por separado, se puede usar la poderosa técnica general de validación cruzada. (Grossman, Seni, Elder, Agarwal, & Liu, 2010, pág. 26).

Se comprende que el método de validación cruzada debe ser la práctica correcta por seguir al no tener un conjunto explícito de datos prueba identificado en el modelo de clasificador bayesiano ingenuo a aplicar. Pero la biblioteca MLib de Apache Spark no permite definir una validación cruzada en conjunto con el modelo clasificador seleccionado, únicamente permite utilizar el método convencional. Es por esta limitante que se utiliza el método de validación 70:30 como la técnica a aplicar en el presente proyecto investigativo.

Se desarrolla un pequeño programa específicamente para la tarea de validación, mostrada en la Figura 32, el cual procesa el conjunto de imágenes de entrenamiento y muestra en el lado derecho de la figura todos los posibles casos de tumores con base en los criterios definidos en el capítulo cuatro. Mientras tanto, del lado izquierdo se muestra la imagen original luego de la etapa de procesamiento de imagen.

Se va definiendo de manera manual cuáles corresponden a tumores y cuáles no, luego el programa almacenará todas las variables numéricas definidas anteriormente junto con una variable binaria numérica discreta (sugerencia valores a utilizar 0 y 1) que representa si es un caso de tumor o no, en un archivo de texto en formato libsvm. Este archivo será almacenado ya sea local o directamente en el HDFS del ambiente Hadoop, para poder ser procesado por el algoritmo clasificador bayesiano ingenuo como el conjunto de entrenamiento que requiere para poder clasificar correctamente el conjunto de pruebas.

El formato libsvm es un formato de datos utilizado ampliamente en aprendizaje de máquinas. Sencillamente consiste en que no existe un encabezado en el archivo, y el archivo se encuentra delimitado por espacios. El primer valor de cada fila debe ser el valor de la variable dependiente, seguido

de los atributos o características que deben seguir el patrón “n:atributo”, donde n representa el valor ordinal de la columna empezando con 1 y atributo equivale al valor en dicha columna.

La importancia de este programa radica en que facilita el definir los datos y variables necesarios para que el modelo bayesiano ingenuo clasifique de forma precisa con base en el conjunto de datos entrenamiento.

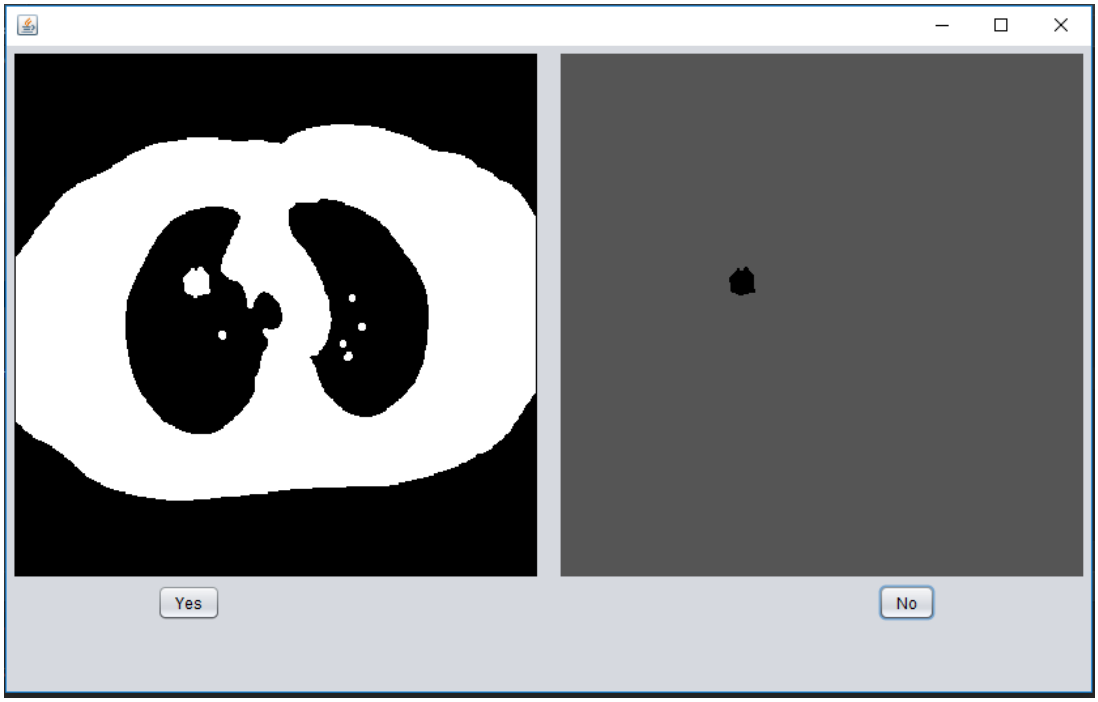
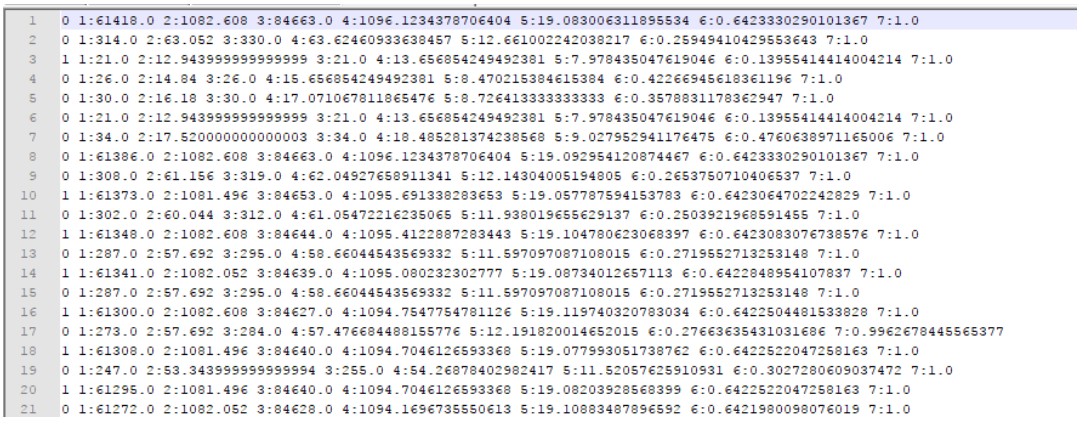


Figura 32: Herramienta para obtener datos de entrenamiento.

Fuente: Elaboración propia.

La Figura 33 muestra un ejemplo de los datos presentes en el archivo de entrenamiento obtenido para el aprendizaje del modelo de minería de datos y aprendizaje de máquinas. Es importante notar que la última columna, “class”, representa la variable binaria que indica si es un caso de tumor pulmonar, equivalente al valor numérico 1, o si no lo es, este último equivalente a 0.



```
0 1:61418.0 2:1082.608 3:84663.0 4:1096.1234378706404 5:19.083006311895534 6:0.6423330290101367 7:1.0
0 1:314.0 2:63.052 3:330.0 4:63.62460933638457 5:12.661002242038217 6:0.25949410429553643 7:1.0
1 1:21.0 2:12.943999999999999 3:21.0 4:13.656854249492381 5:7.978435047619046 6:0.13955414414004214 7:1.0
0 1:26.0 2:14.84 3:26.0 4:15.656854249492381 5:8.470215384615384 6:0.42266945618361196 7:1.0
0 1:30.0 2:16.18 3:30.0 4:17.071067811865476 5:8.726413333333333 6:0.3578831178362947 7:1.0
0 1:21.0 2:12.943999999999999 3:21.0 4:13.656854249492381 5:7.978435047619046 6:0.13955414414004214 7:1.0
0 1:34.0 2:17.520000000000003 3:34.0 4:18.485281374238568 5:9.027952941176475 6:0.4760638971165006 7:1.0
0 1:61386.0 2:1082.608 3:84663.0 4:1096.1234378706404 5:19.092954120874467 6:0.6423330290101367 7:1.0
0 1:308.0 2:61.156 3:319.0 4:62.04927658911341 5:12.14304005194805 6:0.2653750710406537 7:1.0
1 1:61373.0 2:1081.496 3:84653.0 4:1095.691338283653 5:19.057787594153783 6:0.6423064702242829 7:1.0
0 1:302.0 2:60.044 3:312.0 4:61.05472216235065 5:11.938019655629137 6:0.2503921968591455 7:1.0
1 1:61348.0 2:1082.608 3:84644.0 4:1095.4122887283443 5:19.104780623068397 6:0.6423083076738576 7:1.0
0 1:287.0 2:57.692 3:295.0 4:58.66044543569332 5:11.597097087108015 6:0.2719552713253148 7:1.0
1 1:61341.0 2:1082.052 3:84639.0 4:1095.080232302777 5:19.08734012657113 6:0.6422848954107837 7:1.0
0 1:287.0 2:57.692 3:295.0 4:58.66044543569332 5:11.597097087108015 6:0.2719552713253148 7:1.0
1 1:61300.0 2:1082.608 3:84627.0 4:1094.7547754781126 5:19.119740320783034 6:0.6422504481533828 7:1.0
0 1:273.0 2:57.692 3:284.0 4:57.476684488155776 5:12.191820014652015 6:0.27663635431031686 7:0.9962678445565377
1 1:61308.0 2:1081.496 3:84640.0 4:1094.7046126593368 5:19.077993051738762 6:0.6422522047258163 7:1.0
0 1:247.0 2:53.343999999999994 3:255.0 4:54.26878402982417 5:11.52057625910931 6:0.3027280609037472 7:1.0
1 1:61295.0 2:1081.496 3:84640.0 4:1094.7046126593368 5:19.08203928568399 6:0.6422522047258163 7:1.0
0 1:61272.0 2:1082.052 3:84628.0 4:1094.1696735550613 5:19.10883487896592 6:0.6421980098076019 7:1.0
```

Figura 33: Ejemplo archivo de conjunto de datos entrenamiento.

Fuente: Elaboración propia.

En un formato no libsvm, como el delimitado por comas, los datos en el archivo de entrenamiento se observarían como se muestra en la Figura 34.

	A	B	C	D	E	F	G	H
1	Area	Perimeter	Area_conv	perimeter	Circularity	Elongation	Convexity	class
2	61418	1082.608	84663	1096.123	19.08301	0.642333	1	0
3	314	63.052	330	63.62461	12.661	0.259494	1	0
4	21	12.944	21	13.65685	7.978435	0.139554	1	1
5	26	14.84	26	15.65685	8.470215	0.422669	1	0
6	30	16.18	30	17.07107	8.726413	0.357883	1	0
7	21	12.944	21	13.65685	7.978435	0.139554	1	0
8	34	17.52	34	18.48528	9.027953	0.476064	1	0
9	61386	1082.608	84663	1096.123	19.09295	0.642333	1	0
10	308	61.156	319	62.04928	12.14304	0.265375	1	0
11	61373	1081.496	84653	1095.691	19.05779	0.642306	1	1
12	302	60.044	312	61.05472	11.93802	0.250392	1	0
13	61348	1082.608	84644	1095.412	19.10478	0.642308	1	1
14	287	57.692	295	58.66045	11.5971	0.271955	1	0
15	61341	1082.052	84639	1095.08	19.08734	0.642285	1	1
16	287	57.692	295	58.66045	11.5971	0.271955	1	0
17	61300	1082.608	84627	1094.755	19.11974	0.64225	1	1
18	273	57.692	284	57.47668	12.19182	0.276636	0.996268	0

Figura 34: Conjunto entrenamiento en formato csv.

Fuente: Elaboración propia.

5.2.3.4 Modelado

Como se define en puntos anteriores, el modelo de minería de datos a aplicar es de un clasificador bayesiano ingenuo, pero en esta etapa es posible redefinir a otro modelo o técnica en el caso de que las métricas y resultados obtenidos del modelo sugerido no cumplan con los objetivos definidos en este trabajo investigativo.

5.2.3.5 Evaluación

Se inicia al comparar las distribuciones obtenidas de nuestra población total contra el conjunto de observaciones prueba, el cual representa el 30% de los datos de la población total. La idea es poder evaluar los datos, obtener el porcentaje de acierto y desacierto respecto a los resultados obtenidos con el grupo de observación analizado y además poder decir si el modelo utilizado fue el correcto mediante un clasificador bayesiano ingenuo, todo mediante la evaluación de una matriz de confusión.

En la sección 5.4 posterior se dará mayores detalles de esta sección en función de los resultados obtenidos de toda la propuesta de solución definida.

5.2.3.6 Explotación

En esta fase de CRISP-DM, como la teoría nos dice, se considera las tareas de planeo de despliegue, monitoreo y mantenimiento, reporte final y revisión con retroalimentación del proyecto, de las cuales únicamente se toma en cuenta la última a fin de determinar qué salió bien y qué salió mal, y encontrar qué se puede mejorar. Además, se determina si el algoritmo clasificatorio escogido no es el adecuado y es necesario replantear otra solución como redes neuronales, redes bayesianas, máquinas de soporte vectorial, árboles de decisión y regresiones logarítmicas, entre otros.

5.3 Funcionamiento en Spark y Apache Hadoop

La Figura 35 muestra un diagrama de la funcionalidad desde una perspectiva de alto nivel sobre cómo la herramienta aplica los conceptos de almacenamiento distribuido, procesamiento paralelo, y cómo el modelo de programación MapReduce en el conjunto de imágenes TAC en formato DICOM a utilizar en detección de cáncer de pulmón.

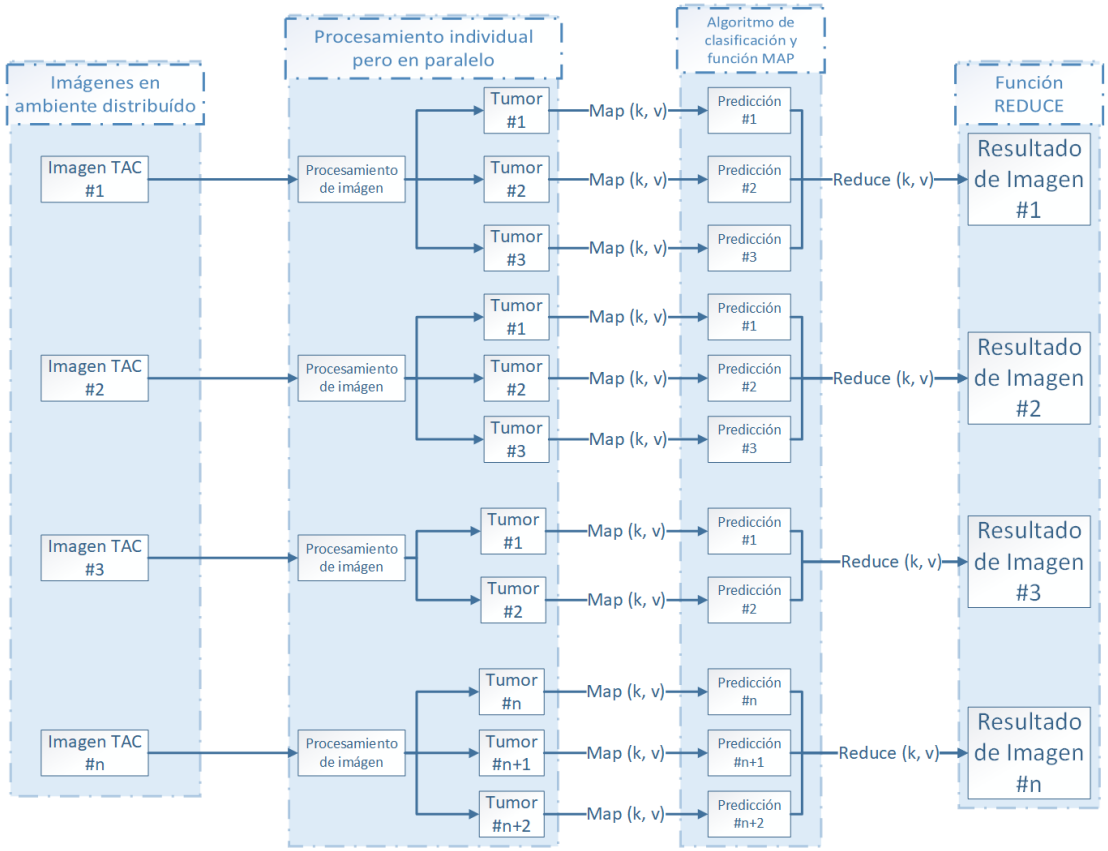


Figura 35: Spark aplicado en la herramienta.

Fuente: Elaboración propia.

En la Figura 35, de izquierda a derecha, se observa inicialmente que se tiene el conjunto de imágenes a estudiar, las cuales son alimentadas en el procesamiento de imágenes definido en la sección 5.2 del presente documento investigativo y que se encuentran almacenadas en el HDFS de Hadoop. Para dicho procesamiento se define un conjunto llave-valor, conformado por la llave como la etiqueta “Patient ID” de la imagen DICOM y el valor como la combinación de todos los atributos definidos en la sección 5.2.3.2 de este trabajo investigativo; esto permite ejecutar la función MAP en cada una de las imágenes a fin de obtener todas las posibles regiones de interés que muestran posibles tumores en los pulmones.

Seguidamente, se aplica el modelo de minería de datos, en este caso el sugerido y el que se utiliza es redes bayesianas de manera que obtenemos de nuevo otro conjunto llave que sigue siendo la misma definida en la función de MAP anterior, pero el binomio valor en este caso debe ser una variable binaria ya sea numérica o categórica que representa el resultado del algoritmo en

predecir si fue un tumor o no lo fue. Esto para, mediante la función REDUCE, agrupar todo aquel set de resultados ya clasificados con base en la llave y deducir si la imagen original presenta un caso de neoplasia pulmonar en el paciente.

5.4 Resultados de la propuesta en la herramienta aplicada con Big Data

En el presente apartado se analiza e interpreta la información de los resultados obtenidos después del desarrollo de la solución y la aplicación de este al conjunto de imágenes TAC obtenidas para este proyecto investigativo.

5.4.1 La herramienta aplicativa del enfoque

La herramienta desarrollada permite la detección de casos de cáncer pulmonar en imágenes en formato médico DICOM, pero también acepta formatos tradicionales como jpg o gif, entre otros.

La solución desarrollada inicia con una primera ventana compuesta de tres botones a seleccionar:

- Detección individual: Ejecución de la propuesta metodológica de detección de cáncer de pulmón mediante procesamiento paralelo en una única imagen.
- Detección múltiple: Ejecución de la propuesta metodológica de detección de cáncer de pulmón mediante procesamiento paralelo con un alto volumen de imágenes.
- Cerrar programa: Cierra el programa y termina el proceso de ejecución en Spark.

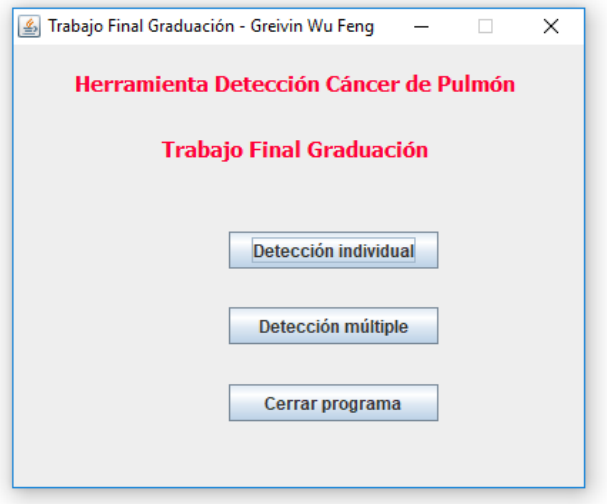


Figura 36: Ventana inicial de la solución.

Fuente: Elaboración propia.

5.4.1.1 Detección individual

Se diseña un módulo de la herramienta para procesar una imagen que puede localizarse tanto localmente en la máquina como en una ruta remota de HFDS, y que además acepta formato DICOM, así como los tradicionales. Este módulo se desarrolla con la idea de poder visualizar el comportamiento de las etapas de procesamiento de imágenes por parte de la metodología de detección de cáncer definida por el presente autor, de ahí el gran impacto que proporciona al proyecto principalmente para la etapa de pruebas.

La Figura 37 muestra el aspecto gráfico de la aplicación, los diferentes componentes de la ventana y su interacción con el usuario.

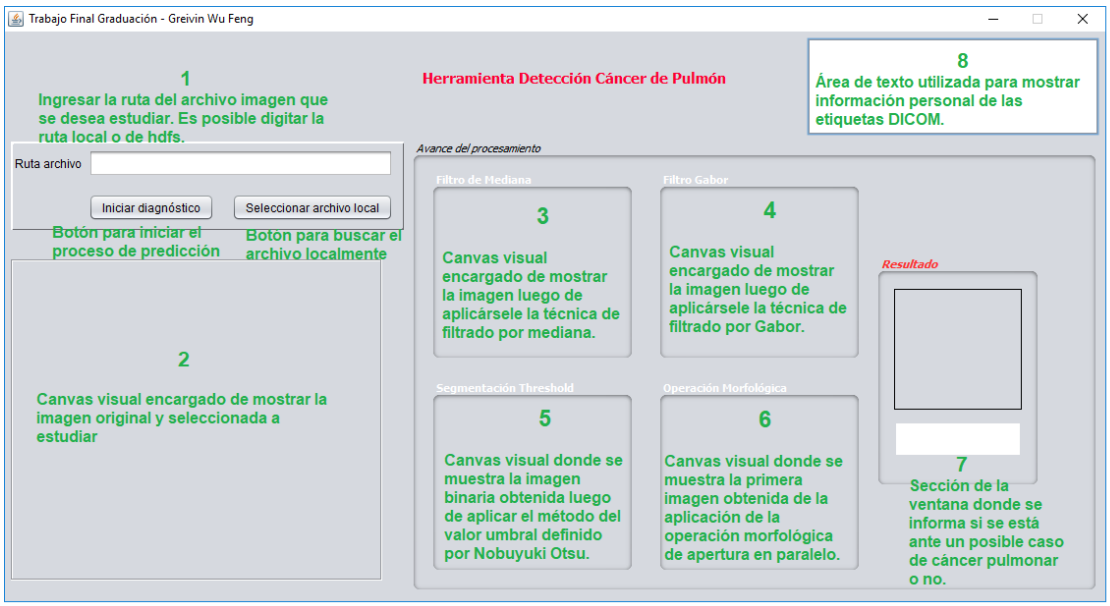


Figura 37: Ventana de procesamiento individual.

Fuente: Elaboración propia.

La Figuras 38 y 39 son ejemplos de cómo se observaría la ejecución de este módulo de detección individual ante una imagen con posible tumor y otra sin presencia alguna de posible cáncer pulmonar respectivamente.

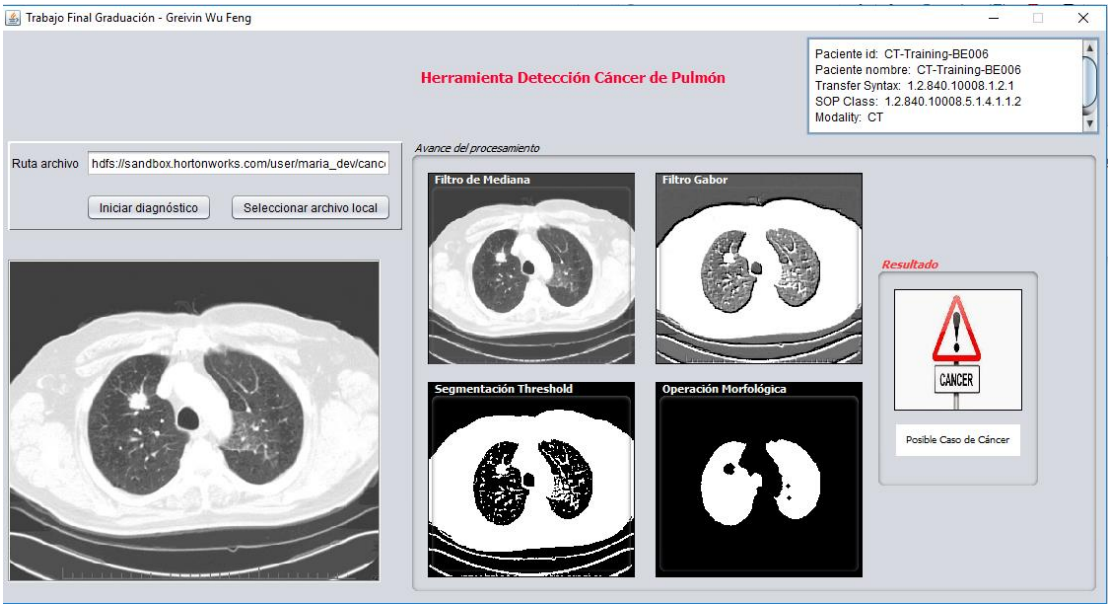


Figura 38: Ejemplo de caso de cáncer de pulmón detectado por la herramienta.

Fuente: Elaboración propia.

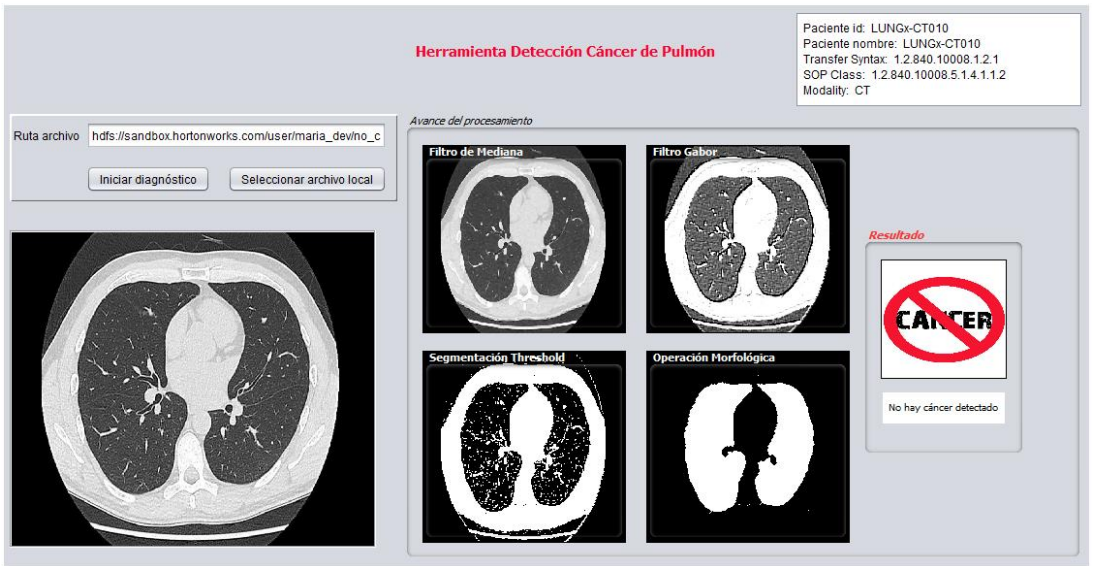


Figura 39: Ejemplo de caso sin cáncer de pulmón detectado por la herramienta.

Fuente: Elaboración propia.

5.4.1.2 Detección múltiple

Se diseña un módulo de la herramienta para procesar varias imágenes a la vez que pueden localizarse tanto en un folder localmente en la máquina como en una ruta de folder remota de HFDS, y que además acepta formato DICOM, así como los tradicionales. Este módulo es la parte principal del proyecto investigativo.

Recordemos que la meta es una solución que permita detección de cáncer en procesamiento paralelo y almacenamiento distribuido, y esta ventana, la cual se observa en la Figura 40, se encarga de cumplir con dichos requerimientos. Es esta sección de la herramienta donde:

1. Se digita la ruta del folder que contiene las imágenes a analizar.
2. Se muestra las predicciones de cada una de las imágenes de los pacientes en el cuadro blanco inferior de la ventana.

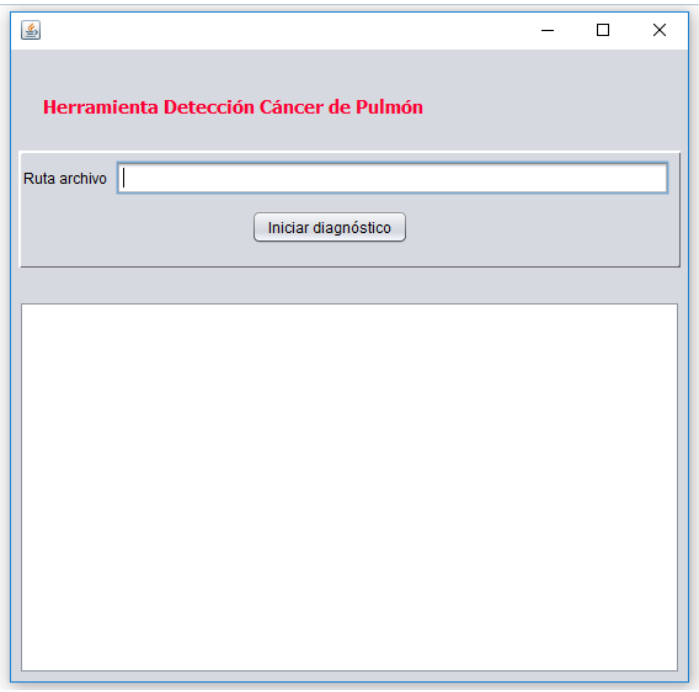


Figura 40: Ventana de procesamiento múltiple.

Fuente: Elaboración propia.

La Figura 41 muestra los archivos de prueba, utilizados en la ejecución de la Figura 42, almacenados en una instancia de Hadoop Hortonworks. La Figura 42 es un ejemplo de cómo se observaría la ejecución de este módulo de detección múltiple.

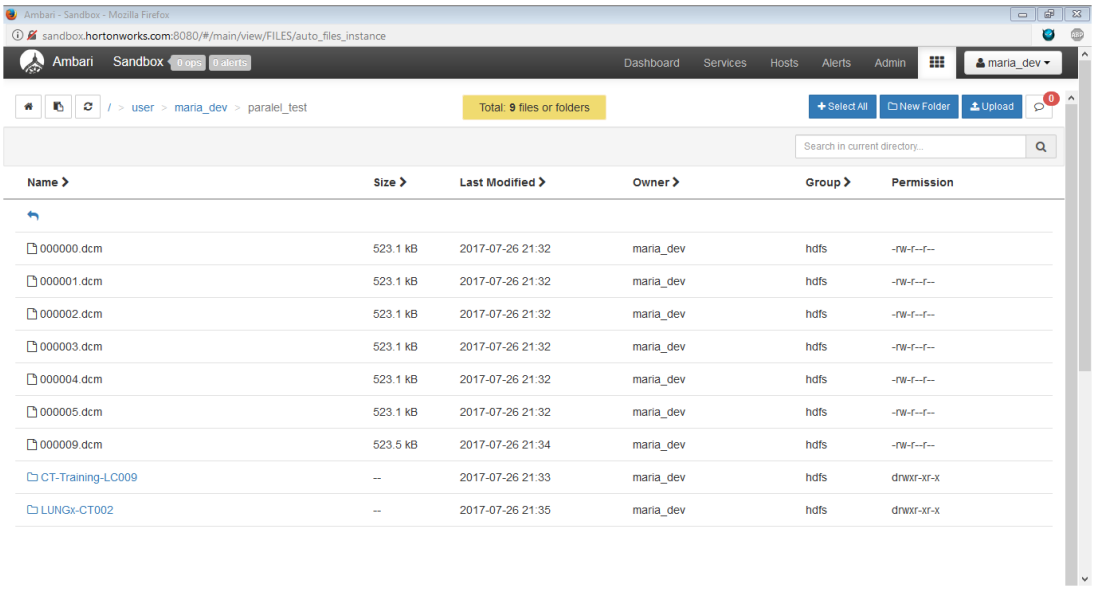


Figura 41: Archivos en Hadoop distribución Hortonworks.

Fuente: Elaboración propia.

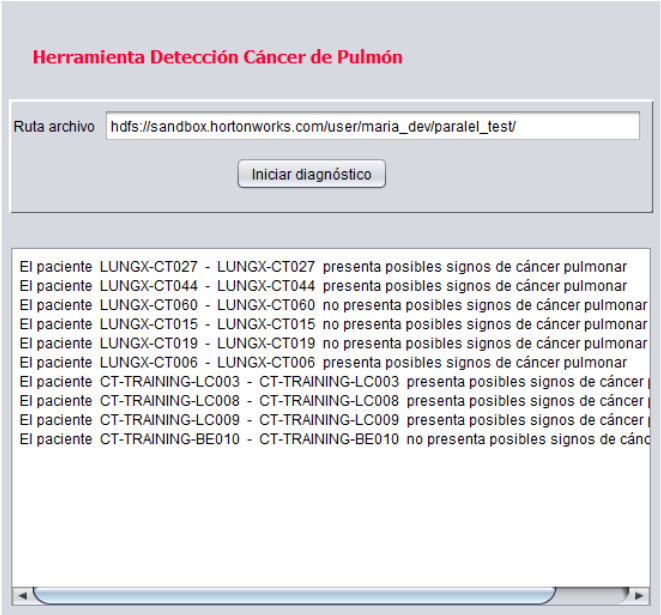


Figura 42: Ejemplo de procesamiento múltiple en la herramienta.

Fuente: Elaboración propia.

5.4.2 Matriz de confusión

Se tiene aproximadamente 12GB de imágenes DICOM referentes a imágenes TAC de pulmones, equivalentes a un total de 22 489 imágenes, donde:

- 5 468 son imágenes con cáncer de pulmón.
- 17 021 son imágenes sin cáncer de pulmón.

Para la primera prueba que se realiza, se seleccionaron al azar un total de 15 742 imágenes para generar nuestro archivo de datos del conjunto de entrenamiento y un total de 6 747 como el conjunto de imágenes de prueba.

Para las pruebas en un ambiente empresarial, se consigue una infraestructura con distribución de Cloudera empresarial compuesta por 8 nodos de puerta de enlace, 4 nodos de utilidad y 120 nodos para datos, todo en hardware HP. No se dan mayores detalles de quién presta el ambiente y cómo es el ambiente Hadoop, por motivos de confidencialidad.

La Figura 43 ilustra la matriz de confusión obtenida de la clasificación de las 6 747 imágenes que componen el conjunto de pruebas. Claramente se observa que las métricas obtenidas son muy buenas, porque son cercanas o mayor a la probabilidad de 90.

		Clasificación		
		Positivo	Negativo	
Valor actual	Positivo	2011	238	2249
	Negativo	158	4340	4498
		2169	4578	

Total	6747

Exactitud	94.13%
Precisión	89.42%
Especificidad	94.80%
Sensibilidad	92.72%

Figura 43: Matriz de confusión.
Fuente: Elaboración propia.

De la figura 43 se concluye que en el orden respectivo de los valores de las métricas calculadas y obtenidas con base en la matriz de confusión:

- La herramienta tiene una frecuencia buena y aceptable de 94.13% en clasificar correctamente los casos, esto es lo que se define como la exactitud.
- Tener una precisión de 89% aproximadamente significa que la propuesta de solución, en gran parte de las ocasiones, cuando detecta casos de cáncer de pulmón, acierta y en términos generales es un buen número a criterio del presente autor.
- La especificidad es la capacidad de una prueba para excluir correctamente a las personas que no tienen cáncer de pulmón. Se demuestra que es aproximadamente 95% específica, en palabras

simples, se tiene una herramienta que el 95% de las ocasiones logra separar correctamente los casos sin presencia de cáncer de los que sí presentan la enfermedad, lo que sugiere un buen valor de métrica de detección y clasificación.

- Por otro lado, sensibilidad es lo opuesto a la especificidad, es la capacidad de una prueba para identificar correctamente a las personas que tienen la enfermedad. Los resultados muestran que es aproximadamente 93% sensible, es decir, en dicho porcentaje los casos se detectan con cáncer correctamente, lo que sugiere un buen valor de métrica.

El realizar una comparación de los porcentajes obtenidos de exactitud y precisión contra las afirmaciones de otras propuestas desarrolladas por diversos autores facilita tener una mejor perspectiva del nivel de implicación de los datos obtenidos. Rupali Mali propone la utilización de diversos algoritmos de Gabor de manera secuencial para la detección de cáncer de pulmón en imágenes en su artículo “Lung cancer detection using modified log-gabor filter based features” (Mali, 2017); en su trabajo se afirma la obtención de métricas mejores a otras propuestas, de ello se obtiene la comparación de metodologías que se muestra en la Tabla 6.

	Metodología por autor			
	Greivin Wu	**Bhagyashri G. Patil y el Profesor Sanjeev N. Jain**	**Anita Chaudhary**	**Rupali Mali**
Exactitud	94.13%	84.56%	79.44%	89.67%
Precisión	89.42%	83.21%	81.83%	90.20%

Tabla 6: Comparativa de diferentes propuestas versus la propuesta del presente trabajo investigativo

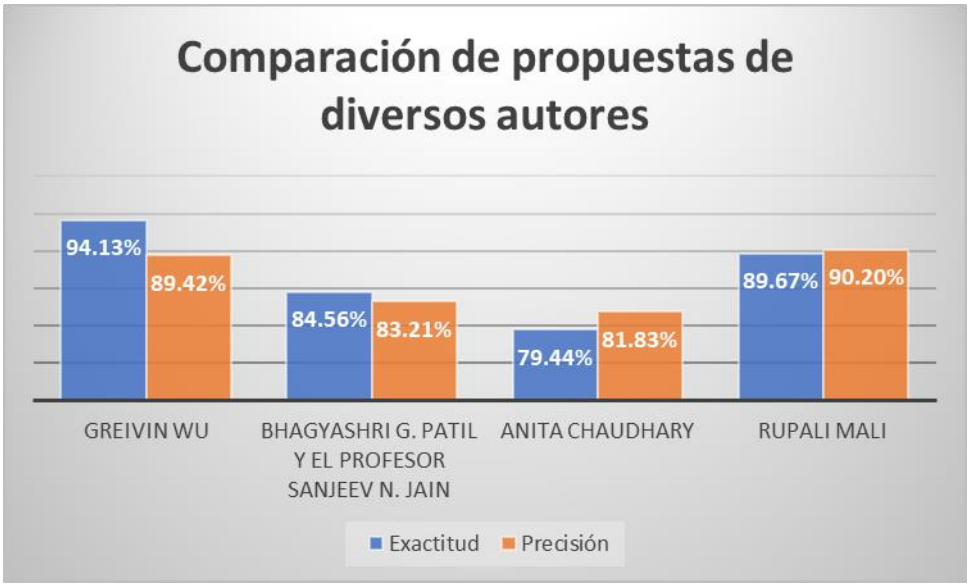


Figura 44: Gráfico de barras en base a la Tabla 6. Fuente: Elaboración propia

5.4.3 Tiempos de ejecución

Sumado a las pruebas anteriores realizadas, se registra los tiempos de ejecución de la solución desarrollada por la cantidad de imágenes a procesar. En el ambiente de Cloudera detallado se promedia 5 segundos de duración por cada 10 imágenes.

Se tiene además una versión standalone de distribución Hortonworks para máquina virtual descargada de la propia página de Hortonworks, la cual tiene como principal requerimiento 8GB de memoria RAM y 4 procesadores de cuatro núcleos. Esta máquina virtual se encuentra corriendo en una computadora personal de escritorio con las siguientes características:

- Memoria: 16GB de RAM.
- Procesador Intel i7-4790k de cuatro núcleos en una frecuencia base de 4GHz.
- Sistema operativo Windows 10 x64bits.

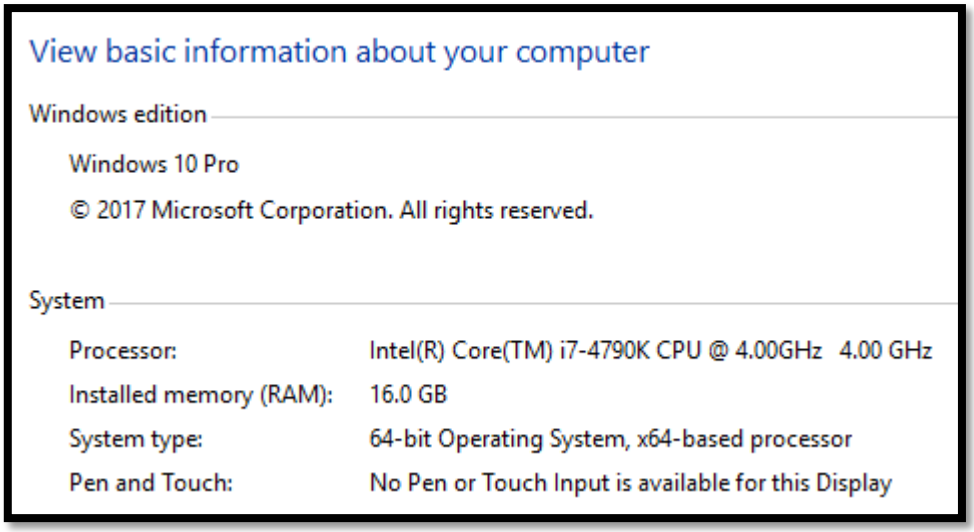


Figura 45: Características de la máquina standalone.

Fuente: Elaboración propia.

En dicho ambiente de un único nodo se obtiene un promedio de 16 segundos por imagen en comparación a un ambiente Hadoop Cloudera multi-nodo empresarial. Con base en dichos resultados, la Figura 47 muestra un gráfico de línea, tal que se observa la correlación lineal positiva entre la cantidad de imágenes a procesar y el tiempo requerido en el procesamiento de las mismas; es decir, a mayor cantidad de imágenes, mayor es el tiempo de procesamiento requerido por la herramienta.

Cantidad de imágenes	Tiempo procesamiento standalone	Tiempo procesamiento multinode
10	16	5
20	32	10
55	88	27.5
100	160	50
271	433.6	135.5
1000	1600	500
1022	1635.2	511
1303	2084.8	651.5
10000	16000	5000
1000000	1600000	500000

Figura 46: Registros de tiempo de ejecución según infraestructura Hadoop.

Fuente: Elaboración propia.

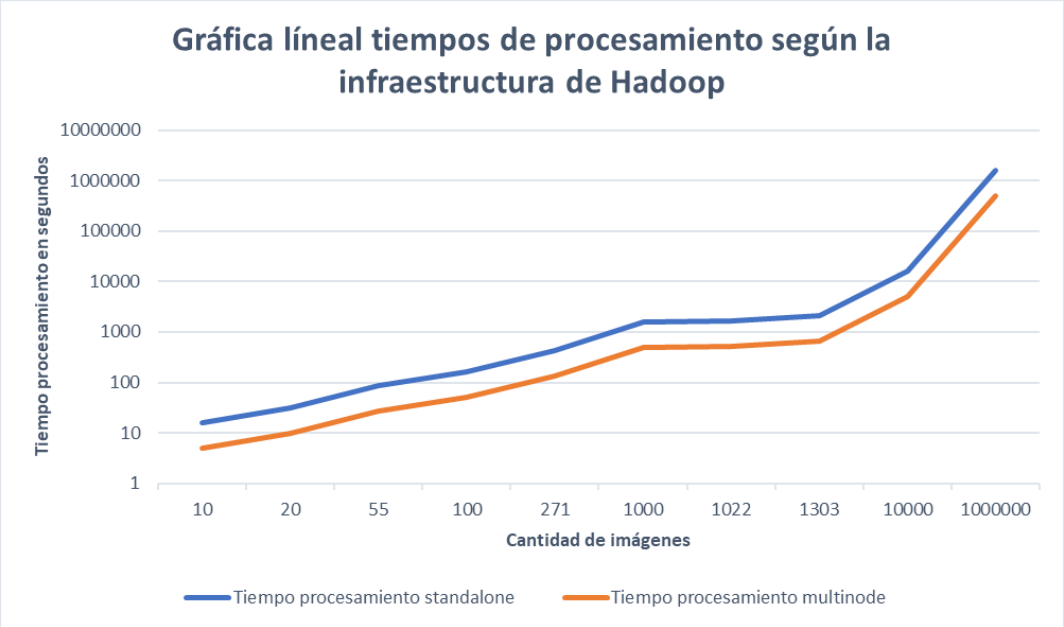


Figura 47: Gráfica lineal de los resultados de la figura 37.

Fuente: Elaboración propia.

5.4.4 Implicaciones de los resultados obtenidos

En general, los resultados del modelo son considerablemente buenos, a pesar de ser un modelo relativamente sencillo mediante redes bayesianas. Del total de observaciones clasificadas, el 94% se hizo de manera apropiada.

Con respecto al tipo de decisiones que se pueden tomar en este caso, la decisión que puede ser más dañina para un paciente es que tenga un tumor o nódulo maligno que representa un posible caso de cáncer pulmonar y erróneamente se diga que no hay presencia de tumor alguno, pues el paciente podría no recibir tratamiento alguno y eventualmente fallecer. Debido a que la especificidad del modelo es del 95%, se puede afirmar que el modelo solo tiene alrededor de un 5% de posibilidades de cometer este error.

Alternativamente, se le podría decir al paciente que tiene un tumor maligno cuando en realidad no existe presencia alguna. En este caso, el paciente podría pasar un mal rato emocionalmente y puede ser sujeto de más pruebas que eventualmente mostrarían la verdad. La sensibilidad del modelo es del 93%, lo cual quiere decir que el modelo tiene cerca de un 7% probabilidad de cometer este error.

La decisión final con respecto a estos números se compartió con Mauricio Guzmán Obando (Obando, 2017), quien considera que dichos

márgenes de error son buenos y aceptables para una versión inicial de herramienta para detección de cáncer, pero es sugerible que se necesitaría un modelo más complejo y a la vez más exacto que el desarrollado aquí para obtener mejores valores, como por ejemplo redes neuronales convolucionales de aprendizaje profundo (Deep learning, en inglés).

Capítulo 6. Conclusiones y Recomendaciones

A continuación se plantea en este capítulo, en primer término, las conclusiones de la presente investigación como la "certificación" del cumplimiento de los objetivos, tanto general como específicos definidos con anterioridad, es decir, las conjeturas a las que llega el presente autor de este trabajo.

Posteriormente, se exponen el conjunto de recomendaciones que son útiles para mejorar el proceso investigativo llevado a cabo, es decir, todas las condiciones que hubieran favorecido un proceso con los menores tropiezos posibles, conducente a un trabajo de investigación de la más alta calidad en términos de estándares, uso del tiempo y costo financiero.

6.1 Conclusiones

Sobre el objetivo específico: "Identificar diferentes enfoques, metodologías o soluciones de detección de cáncer mediante procesamiento de imágenes con un buen porcentaje de exactitud y precisión", se concluye:

- Existe una gran variedad de trabajos referentes a diversas maneras de lograr detección de cáncer mediante el procesamiento de imágenes, desde cómo detectar cáncer de piel hasta cáncer de mamas y por supuesto de pulmón, que es el tema del trabajo investigativo. Pero la mayoría de las soluciones propuestas no logran un nivel de exactitud y precisión cercano o mayor al 85% de probabilidad, que para este autor es el nivel de aceptación de un buen enfoque o metodología.
- Apenas se logra identificar una pequeña cantidad de trabajos con un grado de probabilidad mayor igual al 80% de exactitud, los cuales son los seleccionados y definidos en el capítulo 1.8.
- Muchos de los enfoques, metodologías y soluciones expuestas y encontradas presentan un grado alto de falta de definición por parte de cada uno de los investigadores y autores de cada propuesta, en cómo ellos detallan, especifican y puntualizan cada uno de sus pasos, desde las técnicas de procesamiento de imágenes hasta los parámetros utilizados em el algoritmo de aprendizaje de máquina aplicado, principalmente redes neuronales y máquinas de soporte vectorial.

- No se encuentra propuesta alguna que aproveche la teoría y las ideas de procesamiento en paralelo y distribuido para la predicción de cáncer de pulmón, así como tampoco la idea de utilizar plataformas como Hadoop y su ecosistema.

Sobre el objetivo específico: “Determinar los mejores algoritmos de preprocesamiento de imágenes, filtrado de imágenes, segmentación y detección de patrones en imágenes médicas”, se concluye:

- Respecto a técnicas de preprocesamiento de imágenes para eliminación de ruido y aumento de la calidad de los pixeles en las imágenes, las mejores propuestas son el filtrado por mediana y filtros de Gabor, ya que son ampliamente utilizados, además de ser los sugeridos, en la gran mayoría de los estudios seleccionados en el capítulo 1.8 del presente documento.
- En cuanto a algoritmos de procesamiento de imágenes y extracción de información de las mismas, el método del valor umbral definido por Otsu es la mejor solución, ya que es sencilla de desarrollar, aplicar y se demostró que funciona mejor en comparación a otros métodos de valor umbral.
- La aplicación de operaciones morfológicas es una solución expuesta únicamente en uno de los estudios seleccionados, pero genera el mismo nivel de resultados que otras propuestas mucho más complejas como el método de segmentación extracción mediante la transformación divisoria. Tiene sin embargo menor necesidad de recursos de la máquina por el tipo de algoritmo y un rendimiento computacional similar a las demás complejas propuestas. La elección se basa también en los resultados que se observaron en complemento con las demás técnicas de procesamiento de imágenes al lograr diferenciar el color negro y blanco, así como la detección de figuras redondeadas en las imágenes TAC que representan posibles nódulos cancerígenos.

Sobre el objetivo específico: “Determinar los algoritmos de minería de datos de tipo clasificación (árbol de decisión, bosque aleatorio, redes

neuronales, entre otros) que complementarán el diagnóstico predictivo de la solución", se concluye:

- Los estudios seleccionados como referencias para esta investigación, así como aquellos que fueron descartados, definen principalmente los algoritmos de redes neuronales y máquinas de soporte vectorial para la predicción de posibles casos de cáncer de pulmón, pero no detallan la parametrización necesaria en la implementación de los mismos, además de que son métodos con alta complejidad de desarrollo. Por ello el autor demuestra que se pueden utilizar algoritmos sencillos, en este caso el de redes bayesianas ingenuas, que generan el mismo nivel de confiabilidad en la detección de posibles tumores de neoplasia pulmonar, siempre y cuando se defina un buen conjunto de datos de entrenamiento.
- Aparte de las redes bayesianas, se probó la utilización de árboles de decisión y bosques aleatorios como otros algoritmos clasificatorios, con resultados tanto de precisión como exactitud menor igual al 50% de probabilidad. Esto debido a que la cantidad de datos entrenamiento clasificados como no cáncer excedía a su contraparte, lo cual ocasiona muchas predicciones erróneas, y la naturaleza propia de dichos algoritmos requiere un conjunto de entrenamiento con etiquetas clasificatorias lo más equitativo posible.

Sobre el objetivo específico: "Construir una metodología de detección adaptativo bajo la combinación de las mejores implementaciones en aplicación de los mejores algoritmos definidos", se concluye:

- La posibilidad de poder crear un enfoque y/o metodología de detección de cáncer de pulmón únicamente mediante la implementación de técnicas de procesamiento de imágenes, así como la combinación de estas con técnicas de minería de datos y aprendizaje de máquinas.
- La herramienta desarrollada en este trabajo investigativo puede ser utilizada para detección de otros tipos de neoplasia como por ejemplo el cáncer de mama, siempre y cuando se provea el conjunto de datos de entrenamiento adecuado, y el tipo de cáncer a detectar y estudiar

cumpla con la idea de posibles tumores en imágenes de tomografía computarizada (TAC).

Sobre el objetivo específico: “Seleccionar las diversas imágenes a utilizar como pruebas, a fin de tener las mejores muestras”, se concluye:

- La página de consorcio que se utiliza para obtener las imágenes presenta más de 50GB de imágenes TAC referentes a cáncer de pulmón a utilizar.
- La obtención exitosa de un gran volumen de imágenes TAC referentes a pulmones permitió que el estudio investigativo fuera satisfactorio.
- La selección correcta de las imágenes adecuadas del conjunto de imágenes obtenidas y la identificación de los posibles tumores pulmonares en ellas mediante la ayuda de una persona del área de medicina facilitó el desarrollo de la herramienta aplicativa.

Sobre el objetivo específico: “Valorar la exactitud, además de si es posible, la sensibilidad y especificidad, de la nueva propuesta versus algunas soluciones publicadas”, se concluye que en cuanto a exactitud y precisión refiere, la investigación arroja mejores números en contraste a propuestas de otros investigadores, como se detalla en el capítulo 5.4.2 de este documento.

Finalmente, respecto al objetivo general: “Desarrollar una solución de asistencia médica computarizada para detección de cáncer de pulmón mediante imágenes de tomografía computarizada (TAC) en Big Data”, se concluye:

- La completitud del desarrollo de la herramienta satisface las restricciones impuestas en el objetivo general y de los objetivos específicos.
- La aplicación de un enfoque para detección de cáncer de pulmón que difiere a los expuestos en los estudios seleccionados.

6.2 Recomendaciones

Después de un análisis exhaustivo de los resultados de las conclusiones, los problemas y observaciones producto del proceso

investigativo, se plantean las siguientes recomendaciones y acciones estratégicas dirigidas a todos aquellos interesados en las posibilidades médicas, científicas, tecnológicas y computacionales que este trabajo investigativo puede dar.

Como primera recomendación es importante la utilización del framework Apache Spark versión igual o superior a 2.0, debido a la introducción de las estructuras de Dataset y Dataframes, que permiten la creación de instancias de Spark con mayor facilidad y en menor consumo de memoria, además de la incorporación de SparkSQL a diferencia de versiones anteriores. Esta recomendación surge a raíz de que las actuales distribuciones de Hadoop, principalmente Cloudera, presentan la versión Apache Spark 1.6 como la predeterminada, y el tratar de utilizar una versión superior implica un grado de complejidad en la configuración necesaria para poder incorporarlo al ecosistema.

Respecto al uso y obtención de las imágenes, se recomienda que sean validadas con un profesional en el área médica. Esto porque es la persona correcta para ayudar a identificar el conjunto de posibles tumores y casos de cáncer pulmonar, de manera que se da credibilidad a los resultados que se obtendrán y de la calidad de los datos utilizados tanto para la generación del conjunto de datos entrenamiento, así como el de prueba.

Utilizar una distribución de Hadoop simplifica los tiempos de desarrollo, aunque puede implicar un aumento en el costo financiero, pero el beneficio supera a los riesgos. Las distribuciones de Hadoop ofrecidas en el mercado ya vienen diseñadas e incorporadas con todas las herramientas necesarias para desarrollo, administración y producción; si se deseara no recurrir a una distribución, sería necesario obtener Apache Hadoop de la página Apache, así como todo el conjunto de frameworks que conforman el ecosistema. La configuración e integración de todos ellos implica un alto grado de complejidad traducido en tiempo de instalación, configuración y pruebas, entre otros pasos no necesarios a realizar cuando se recurre a una distribución como solución completa.

El uso de estrategias de desarrollo de software es importante. Para este proyecto se aplica la metodología Agile como forma de poder organizar las

cargas de trabajo y poder completar el trabajo a tiempo sin mayores inconvenientes.

Otra de las recomendaciones es no obviar la metodología CRISP-DM definida en este proyecto investigativo. Al igual que la metodología Agile, CRISP-DM es importante de aplicar, porque mientras la primera se centra en la parte del desarrollo de la herramienta, la segunda se enfoca en la analítica necesaria a utilizar en el módulo de predicción de la solución final de este proyecto investigativo.

Finalmente, como última recomendación, es importante el estudio de diversos métodos y algoritmos de aprendizaje de máquina posibles a utilizar. Se probó, aparte de las redes bayesianas, el método de bosque aleatorio, el primer método con mejores resultados que el segundo. El análisis exhaustivo de diversas técnicas de clasificación permitirá una herramienta más confiable en cuanto a los resultados de predicción de cáncer que genere.

Capítulo 7. Reflexiones Finales

Gran parte de la temática que gira respecto a ciencias de datos a través de los años se ha enfocado y se ha basado principalmente en sistemas de almacenamiento de datos de manera estructurada. Es por ello que el surgimiento de nuevos tipos de datos, los no estructurados como las imágenes que se utilizan en esta investigación, implica que las teorías, conceptos y dimensiones de calidad de datos propuestas y aplicadas actualmente deben ser rediseñadas de manera que consideren los datos no estructurados, los grandes volúmenes de datos y las nuevas formas de almacenamiento de datos que surgen, y las nuevas formas de procesamiento distribuido y en paralelo.

El surgimiento de conceptos como Internet de las cosas (IoT) genera una masificación de nuevos productos y servicios que representará un incremento del Big Data, porque será tal el volumen de datos que no podrá ser analizado mediante herramientas tradicionales. El poder obtener información mediante un dispositivo conectado a la red y almacenar todos esos datos en soluciones de Big Data representará una forma para poder descubrir patrones de conducta de los usuarios y mejorar muchos aspectos de la vida, e incluso el famoso concepto de ciudad inteligente.

Es posible afirmar que el gran desafío no es la tecnología misma, que evoluciona a pasos agigantados, sino el cómo asegurarse de que se tiene las suficientes habilidades para hacer el uso efectivo de la tecnología a disposición y darle sentido a los datos recolectados. Para ello se requiere resolver los problemas legales que giran en torno a los derechos de propiedad intelectual, privacidad e integridad de los datos, ciberseguridad y un código de conducto sobre Big Data.

Sumado a todo lo anterior, en este trabajo investigativo es claro que se ha combinado muchas áreas de la ingeniería en computación y de la ingeniería eléctrica. El aprendizaje en materia de procesamiento de imágenes por parte de este autor no fue sencillo, ya que fue un tema nuevo, pero fue para él una satisfacción el proceso de aprendizaje que queda como resultado. En esta idea y sentimiento se desea dirigir que los futuros esfuerzos en materia de procesamiento de audio, videos y cualquier otro tipo de datos no estructurados proporcionará altas posibilidades de beneficios analíticos.

Una plataforma de análisis adecuado debe basarse en tres parámetros: rendimiento, infraestructura de tamaño correcto, y el crecimiento futuro. Para obtener un rendimiento, un servidor de aplicaciones, un servidor físico de varios inquilinos dedicado a un solo cliente es el mejor ajuste. Para la infraestructura y el crecimiento futuro, la idea de un híbrido es el mejor enfoque. Con implementaciones híbridas, que consisten en la nube, hosting gestionado, colocación y alojamiento dedicado, se combinan las mejores características de múltiples plataformas en un único entorno óptimo.

Finalmente, el hecho de que se pueda almacenar datos no estructurados y semiestructurados obviando los datos estructurados, no significa que no va ser necesario tomar en cuenta estos últimos. En el capítulo siguiente se hablará de las posibilidades futuras de este trabajo investigativo mediante la integración de todos los tipos de datos, abriendo así nuevas concepciones de generación de información con base en datos obtenidos de las imágenes.

Capítulo 8. Trabajos a Futuro

Este capítulo vislumbra posibles trabajos a futuros, derivados de la propia investigación, de forma que se hará un gran favor a la comunidad científica (compañeros que aún no hacen el TFG y demás personas), al sugerir posibles ampliaciones o cambios de enfoque en la solución ofrecida.

8.1 Posibilidades analíticas

La herramienta propuesta en el trabajo no contempla el almacenar el resultado de los estudios tanto de los casos positivos y falsos de detección de cáncer de pulmón en las imágenes de TAC. Esto es importante a futuro porque permitirá no solo salvaguardar la información de los pacientes para futuras necesidades médicas, sino abre un margen grande en cuanto a estudio demográfico o médico, entre otros, referente a los datos obtenidos.

Una propuesta es la posibilidad de poder almacenar toda la información que se obtiene en las imágenes DICOM y el resultado del procesamiento del mismo en el HDFS, y que este sea accesible mediante el framework de Impala, como parte del ecosistema de Hadoop. Se propone Impala debido a su eficiencia para devolver resultados de consultas en SQL con respecto a Hive.

La posibilidad de poder incorporar Impala permite entonces que se abra el abanico de las siguientes posibilidades:

1. Una solución en Spark que lee los datos de las tablas en Impala mediante SparkSQL y en conjunto con la biblioteca MLib. Se puede realizar análisis de datos más profundos como encontrar la relación lineal entre diferentes variables vinculadas a atributos que se obtienen de las imágenes DICOM o de las obtenidas luego del procesamiento de los datos (área, perímetro y otros del nódulo cancerígeno) mediante el uso de regresiones lineales simples y múltiples.
2. Existen diversas soluciones de reporteo, desde Excel como la herramienta más simple y fácil de adquirir, hasta soluciones más completas y empresariales por mencionar algunas, Oracle Business Enterprise Edition (OBIEE) de Oracle, TIBCO Spotfire o Tableau, que posibilitan la conexión a Impala. Así se puede realizar una visualización de los datos como reportes, dashboards con base en la información, de manera tal que se podría estudiar demografía en donde ocurren más

casos de detección de cáncer de pulmón, o qué tipo de cáncer de pulmón es mayormente detectado, entre muchas incógnitas posibles a responder mediante el estudio visual gráfico de los datos.

8.2 Herramienta de intranet, internet y extranet

La solución propuesta es ejecutable en un ambiente Hadoop local, siempre y cuando sea posible acceder a Spark desde la consola propia del ambiente Hadoop que se utiliza, es decir, no es posible tratar de conectarse al sistema desde fuera de dicha máquina. Por tanto, se sugiere la incorporación de soluciones de servicio web (webservice en inglés), como lo es Livy para Spark. Esta es una solución que nos permite comunicarnos a Spark mediante un servicio web de REST.

La incorporación de Livy permitiría ejecutar, desde otra máquina propia de la intranet del centro médico e incluso fuera de él, la solución de detección de cáncer pulmonar luego de que se haya subido las imágenes a HDFS o algún medio intermedio de acceso, por ejemplo. Esto permitiría, también desde la perspectiva de tecnologías de información, tener un lugar centralizado de donde se accede a la información y se solicita el análisis de los datos, para dar mayor privacidad de la información, seguridad de la misma y que se cumpla con estándares en materia de Gobernanza de los datos, además de una disminución en costos de equipos computacionales distribuidos.

Adicionalmente, debido al auge de los dispositivos inteligentes como lo son las tabletas o los teléfonos inteligentes, es factible utilizar la herramienta como una aplicación móvil que accede al sistema.

8.3 Posibilidades en el estudio de otros cánceres

La herramienta ciertamente está dirigida a neoplasias de tipo pulmonar, pero debido a la metodología general que el presente autor propone, es posible aplicarla en el diagnóstico preventivo de otros tipos de cáncer, como lo es el cáncer de mama o cáncer cerebral, entre otros.

Aquellos cánceres que comparten la existencia de tumores con formas irregulares redondeadas pero que son similares a las que se encuentran en los pulmones, similares a aquellas observadas en las imágenes TAC obtenidas para este trabajo investigativo, son posibles de detectar con el enfoque propuesto. Para ello, el único cambio que se debe realizar a la solución es en el

conjunto de datos entrenamiento que se le debe proveer al algoritmo clasificatorio de aprendizaje de máquina, de manera tal que este acierte en la detección del cáncer que se desea estudiar.

También se abre la posibilidad del estudio de cáncer de piel como otra idea, pero la metodología debe cambiar ligeramente para poder adaptarse. Para ello, es necesario ya no tratar imágenes tipo grises a imágenes binarias únicamente y analizar las mismas propiedades en tumores con los lunares, sino también se debe realizar el estudio de las variaciones de colores en los lunares, porque esto también representa la posibilidad de cáncer de piel. Por ello a la solución se debe integrar el procesamiento de colores RGB, o cualquier otro estándar de colores, en imágenes que ya no necesariamente deben ser de DICOM; para este caso, es posible hablar de otros formatos como JPEG o PNG, entre otros.

8.4 Aplicación móvil

Mediante la combinación de las ideas de los dos puntos anteriores, una de las ideas que surge es la posibilidad de una aplicación móvil en dispositivos inteligentes, donde se puede tomar la foto de un lunar utilizando un celular con cámara o una tableta y mediante la aplicación solicitar el procesamiento de la misma para la detección de un posible cáncer de piel.

La sugerencia de este autor permitiría que ya la herramienta no dependa de una computadora, porque en hospitales de Costa Rica en las zonas alejadas como Talamanca, por ejemplo, no es tan sencillo tener una computadora a mano, pero la facilidad de portar dispositivos inteligentes es viable. Por supuesto se debe tomar en cuenta cuestiones como la red telefónica, entre otros, pero se lanza la idea al aire como una posibilidad viable a realizar.

Referencias

(s.f.). Obtenido de ImageJ - Image Processing and Analysis in Java [Software]: https://imagej.nih.gov/ij/download.html

(s.f.). Obtenido de DocCheck Pictures: http://pictures.doccheck.com/es/photo/23703-tac-de-torax-carcinoma-pulmonar

Acharya, T., & Ray, A. K. (2005). *Image Processing Principles and Applications.* Tucson, Arizona, Estados Unidos de América: John Wiley & Sons, Inc.

Afsaneh, B., & Pennell, N. (2010). *Targeting Angiogenesis in Non-Small Cell Lung Cancer: Agents in Practice and Clinical Development.*

Andrew. (21 de Julio de 2012). *Why split data in the ratio 70:30?* Obtenido de Information Gain Ltd: http://information-gain.blogspot.com/2012/07/why-split-data-in-ratio-7030.html

Ayres, J. R. (2004). *Norma e formação: horizontes filosóficos para as práticas de avaliação no contexto da promoção da saúde.* Ciênc. saúde coletiva. Recuperado el 1 de Agosto de 2017, de http://dx.doi.org/10.1590/S1413-81232004000300011

Biolchini, J., Mian, P. G., Natali, A. C., & Travassos, G. H. (2005). *Systematic Review in Software Engineering.* Systems Engineering and Computer Science Department, Rio de Janeiro. Recuperado el 03 de 07 de 2017

biomédicas, I. N. (2013). *Tomografía Computarizada (TC).* Recuperado el 22 de Julio de 2017, de https://www.nibib.nih.gov/sites/default/files/Tomograf%C3%ADa%20Computarizada%20%28TC%29.pdf

Blanco, C., Lasheras, J., Valencia, R., Fernández, E., Toval, A., & Piattini, M. (2008). *Ontologías de Seguridad: Revisión sistemática y comparativa.* Universidad de Castilla-La Mancha, Departamento de Tecnologías y Sistemas de Información. Universidad de Castilla-La Mancha. Recuperado el 03 de 07 de 2017

Borthakur, D. (2008). *HDFS Architecture Guide.* The Apache Software Foundation.

Cedano, J. Á. (2015). *Modelo de minería de datos para identificación de patrones que influyen en el aprovechamiento académico.* Instituto Tecnológico de la Paz, La Paz, Baja California Sur, México.

Defranchi, D. S. (s.f.). *Cirugía Torácica.* Obtenido de ¿Qué es un nódulo pulmonar?: http://www.thoracicsurgeryblog.com/que-es-un-nodulo-pulmonar/

Deshpande S., A., Lokhande D., D., Mundhe P., R., & Ghatole M., J. (Marzo de 2015). Lung Cancer Detection with fusion of CT and MRI Images Using Image Processing. *International Journal of Advanced Research in Computer Engineering & Technology, 4*(3), 763-767.

Díaz, S. O. (2005). *Introducción a Técnicas de Minería de Datos.* Obtenido de http://images.slideplayer.es/1/92234/slides/slide_37.jpg

Foundation, T. A. (2008). *MapReduce Tutorial.* The Apache Software Foundation.

Foundation, T. A. (2014). *Welcome to Apache™ Hadoop®!* The Apache Software Foundation.

Funcionamiento de SVM. (s.f.). Recuperado el 23 de Julio de 2017, de IBM Knowledge Center:

https://www.ibm.com/support/knowledgecenter/es/SS3RA7_17.1.0/modeler_mainhelp_client_ddita/clementine/svm_howwork.html

Gajdhane, V. A., & L.M., D. (2014). Detection of Lung Cancer Stages on CT scan Images by Using Various Image Processing Techniques. *IOSR Journal of Computer Engineering, 16*(5), págs. 28-35.

García, M. N., Quintales, L. A., Peñalvo, F. J., & Martín, M. J. (2006). *Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requesitos de software.* Universidad de Salamanca, Departamento de Informática y Automática. Recuperado el 7 de Julio de 2017

García, V. G. (2010). *Estimación y clasificación de daños en materiales utilizando modelos AR y redes neuronales para la evaluación no destructiva con ultrasonidos.* Universidad de Granada, Departamento de Teoría de la Señal, Telemática y Comunicaciones, Granada.

Goldstraw, P. (2015). The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer.

Gomathi, M., & Thangaraj, P. (2011). A computer aided diagnosis system for lung cancer detection using machine learning technique. *European Journal of Scientific Research*, 260-275.

Grossman, R., Seni, G., Elder, J., Agarwal, N., & Liu, H. (2010). *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions.* Morgan & Claypool.

Guru99. (s.f.). *What is MapReduce? How it Works - Hadoop MapReduce Tutorial*. Obtenido de Guru99: https://www.guru99.com/introduction-to-mapreduce.html

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la Investigación* (6ta ed.). MCGRAW-HILL.

Image Processing Toolkit. (s.f.). Obtenido de VirtualLabs: http://bmsip-iitr.vlabs.ac.in/exp3/Theory.html?domain=Biotechnology&lab=Bio-Medical%20Signal%20and%20Image%20Processing%20Lab

Introducción al Aprendizaje Automático. (23 de Setiembre de 2017). Obtenido de Fernando Sancho Caparrini: http://www.cs.us.es/~fsancho/?e=75

Introducción al Aprendizaje Automático. (23 de Setiembre de 2017). Obtenido de Fernando Sancho Caparrini: http://www.cs.us.es/~fsancho/?e=75

Jensen, K. A. (9 de Febrero de 2012). Obtenido de Google Sites: https://sites.google.com/site/kennethagregaardjensen/crisp-dm

Karau, H., Kowinski, A., & Zaharia, M. (2015). *Learning Spark: Lightning-Fast Big Data Analysis.* O'Reilly Media.

Kumar, M. (2013). *DIGITAL IMAGE PROCESSING.* Indian Institute of Remote Sensing, Dehra Dun, Photogrammetry and Remote Sensing Division. Obtenido de http://www.wamis.org/agm/pubs/agm8/Paper-5.pdf

Kuruvilla, J., & Gunavathi, K. (Enero de 2014). Lung cancer classification using neural networks for CT images. *In Computer Methods and Programs in Biomedicine, 113*(1), 202-209. doi:https://doi.org/10.1016/j.cmpb.2013.10.011

Mad, E. (2005). *Informáticos Generalitat Valenciana Grupos a y B. Temario Bloque Específico Volumen 1.* MAD.

Maimon, O., & Rokach, L. (2005). *Data Mining and Knowledge Discovery Handbook.* Springer.

Mali, R. R. (5 de Mayo de 2017). Lung cancer detection using modified log-gabor filter based features. *International Journal of Recent Innovation in Engineering and Research, 2*, 65-70. Obtenido de https://ijrier.com/published-papers/volume-2/issue-5/lung-cancer-cell-detection-using-modified-log-gabor-filter-based-features.pdf

Martínez, B. B. (2009). *Minería de Datos.* Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación, México.

Matich, D. J. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones.* Universidad Tecnológica Nacional, Departamento de Ingeniería Química, Rosario. Obtenido de https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf

median filter in AWK. (24 de Junio de 2015). Obtenido de One Tip Per Day: http://onetipperday.sterding.com/2015/06/median-filter-in-awk.html

Miljković, O. (2006). *IMAGE PRE-PROCESSING TOOL.* Megatrend University of Belgrade, College of Computer Science, Novi Beograd, Serbia.

Ministerio de Trabajo y Seguridad Social. (2017). *Lista de Salarios Mínimos por ocupación Año 2017.* http://www.mtss.go.cr/temas-laborales/salarios/lista_salarios_1_2017.PDF: Ministerio de Trabajo y Seguridad Social.

Montes, T. O., & Castillo, P. Á. (1996). La Morfología Matemática en el Tratamiento Digital de Imágenes. *Revista de la Facultad de Educación de Albacete*, págs. 241-256.

Morales, E., & González, J. (2013). *Aprendizaje Computacional.*

Morphological Image Processing. (s.f.). Obtenido de The University of Auckland: https://www.cs.auckland.ac.nz/courses/compsci773s1c/lectures/ImageProcessing-html/topic4.htm

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* The MIT Press.

Obando, M. G. (2017). Comunicación personal. (G. W. Feng, Entrevistador)

Patton, M. Q. (1996). *Utilization-Focused Evaluation* (3ra ed.). SAGE Publications.

Refaeilzadeh, P., Tang, L., & Lui, H. (2008). *Cross-Validation.* Arizona State University. Recuperado el 16 de Diciembre de 2017, de http://leitang.net/papers/ency-cross-validation.pdf

Rosti, G., Bevilacqua, G., Bidoli, P., Portalone, L., Santo, A., & Genestreti, G. (2006). Small cell lung cancer. *Annals of Oncology, 17: Suppl. 2*, i5–ii10.

Schiller, J. H., Harrington, D., Belani, C. P., Langer, C., Sandler, A., Krook, J., . . . Johnson, D. H. (2002). *Comparison of Four Chemotherapy Regimens for Advanced Non–Small-Cell Lung Cancer.*

Stufflebeam, D. L., & Shinkfield, A. J. (1987). *Evaluación sistemática: guía teórica y práctica.* San Sebastián de los Reyes, Madrid, España: Editorial Paidós.

Suárez, E. J. (2014). *Tutorial sobre Máquinas de Vectores de Soporte (SVM).* Universidad Nacional de Educación a Distancia, Departamento de Inteligencia Artificial, Madrid.

Toolox Image - A Toolbox for General Purpose Image Processing. (2008). Obtenido de MathWorks: Obtenida de http://in.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/16201/versions/3/previews/toolbox_image/html/content.html

Visualizador de imágenes médicas [Software]. (2017). Obtenido de MicroDicom: http://www.microdicom.com/

Wagner, T., & Lipinski, H.-G. (14 de Octubre de 2013). IJBlob: An ImageJ Library for Connected Component Analysis and Shape Analysis. *Journal of Open Research Software*(1), pág. 6. doi:http://doi.org/10.5334/jors.ae

Young, I. T., Gerbrands, J. J., & Vliet, L. J. (1995). *Fundamentals of Image Processing.*