UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

PRIORIZACIÓN DE PACIENTES DE CÁNCER DE PULMÓN, UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL SOBRE TC. DE TÓRAX

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MARIANO AGUSTÍN SUÁREZ SALAS

PROFESORA GUÍA:
Carolina Segovia Riquelme

MIEMBROS DE LA COMISIÓN:
Richard Weber Haas
Rodolfo Urrutia Uribe

SANTIAGO DE CHILE
2023

PRIORIZACIÓN DE PACIENTES DE CÁNCER DE PULMÓN, UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE LENGUAJE NATURAL SOBRE TC. DE TÓRAX

El cáncer de pulmón en Chile es una de las principales limitantes de la esperanza de vida de los habitantes del país, al ser la enfermedad con mayor nivel de mortalidad. Detectar tempranamente una sospecha de cáncer de pulmón puede incurrir en grandes mejoras para la esperanza y calidad de vida de las personas, así como el impacto de la enfermedad para el Estado y el Sistema de Salud Público.

Existe información clave para la detección temprana de cáncer de pulmón presente en los exámenes de TC. de tórax. Se propone generar un modelo predictivo de sospecha de cáncer de pulmón a través de técnicas de procesamiento de lenguaje natural, con la finalidad de priorizar casos sospechosos en el Hospital Doctor Sótero del Río, optando a un tratamiento oportuno y eficaz.

El diseño de modelos de predicción de sospecha de cáncer pulmonar se desarrolla en base a clasificadores clásicos del NLP como lo son *Random Forest* en conjunto con TF-IDF. Así como modelos de Redes Neuronales del estado del arte y *word embeddings*. Se evaluó el desempeño en base a la métrica *Recall* para la clase positiva, es decir, un paciente con sospecha de cáncer de pulmón. El mejor modelo predictivo de cáncer de pulmón posee una alta sensibilidad, y predice correctamente el 89 % de los casos positivos. Los datos utilizados corresponden a TC. de tórax etiquetados por Oncólogos y equipo médico de la Red de Salud. Al utilizar biopsias y otras fuentes de datos para la construcción de una etiqueta de cáncer general, el desempeño disminuye con un Recall del 63 %.

Se evidencia el potencial del NLP y el aprendizaje automático en la detección de cáncer de pulmón a partir de informes de TC. Se lograron resultados prometedores, pero es esencial considerar las implicaciones clínicas y abordar las limitaciones para una implementación exitosa en entornos médicos reales.

Dedicado a todos aquellos valientes que han batallado contra el cáncer,
y sus familias.

Agradecimientos

Quiero expresar mi profundo agradecimiento a todas las personas que han contribuido de manera significativa a la realización de esta tesis. En primer lugar, a Carolina, Richard y Rodolfo, por su orientación experta, paciencia y apoyo constante a lo largo de este proceso.

Agradezco sinceramente a Sebastián Santana, José Peña, y todo el equipo del proyecto CETRAP por sus valiosas contribuciones y sugerencias que enriquecieron este trabajo.

Mi reconocimiento especial a mis colegas y amigos que me acompañaron durante el proceso de formación universitaria, sin ellos no lo habría logrado.

Agradezco a mi familia por su constante apoyo emocional y comprensión durante este desafiante viaje académico.

Finalmente, dedico este logro a mi Yaya, que ya no esta con nosotros y a mi Tata que han enfrentado la batalla contra el cáncer. Su valentía y determinación son una inspiración constante. Que este trabajo contribuya, aunque sea modestamente, a la búsqueda de soluciones para aliviar el sufrimiento de aquellos que enfrentan esta enfermedad.

Gracias a todos los que formaron parte.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Antecedentes generales

El cáncer corresponde a un conjunto de enfermedades relacionadas a un proceso de división de células de forma descontrolada. En todo el mundo, el cáncer se ubica como una de las principales causas de muerte, y una barrera importante para aumentar la esperanza de vida. Las dos principales medidas de impacto del cáncer, corresponden a la mortalidad e incidencia, en dónde la primera representa al número de defunciones en cierto grupo de personas en determinado período, mientras que la incidencia se define como el número de casos nuevos de enfermedad que se diagnostican en un período. A nivel mundial, de todos los tipos de cáncer, es el **cáncer de pulmón** aquel con mayor mortalidad, encontrándose tercero en el nivel de incidencia, para el año 2020 (Ver figura 1.1).Ferlay et al. (2020)

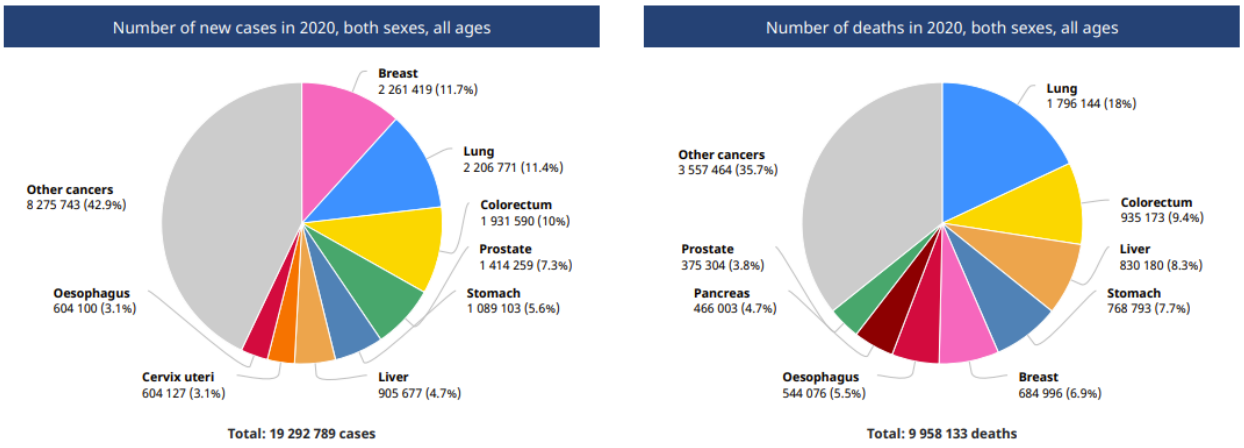

Figura 1.1: Mortalidad e incidencia mundial según el tipo de cáncer

A nivel nacional, el cáncer también representa un problema de salud pública, dada a su alta incidencia y mortalidad. Esta enfermedad corresponde a una importante problemática con alto impacto social y económico, debido a la repercusión y costos que afectan a los pacientes, sus familias, tanto como el país y el sistema de salud, según El PLAN NACIONAL DE CÁNCER 2018 – 2028 del Ministerio de Salud. Además, el **cáncer de pulmón** también

reporta la mayor tasa de mortalidad en Chile en el 2018, según lo reportado por Agencia Internacional para la Investigación en Cáncer (AIRC), Parra-Soto et al. (2020).

Los sistemas públicos de salud de Chile atienden a casi el 80 % de la población, correspondiente al porcentaje de la población inscrita en FONASA según el Instituto Nacional de Estadística (INE). Estas instituciones forman redes complejas que a menudo presentan problemas en la integración de datos y la gestión de casos oncológicos, lo que lleva a un diagnóstico tardío y bajas tasas de adherencia al tratamiento.

El proyecto de tesis fue desarrollado en el Hospital Sótero del Río (HSR), en conjunto con el Servicio de Salud Metropolitano Sur Oriente (SSMSO), siendo financiado por la fundación Chilesincáncer.

El proyecto en cuestión tiene por nombre: *"Desarrollo de una nueva estrategia para el diagnóstico precoz y el tratamiento del cáncer de pulmón sospechoso mediante inteligencia artificial en Chile"*. Y busca, mediante la utilización de procesamiento de lenguaje natural (NLP), predecir sospechas de cáncer pulmonar mediante la lectura automática de reportes asociados a exámenes de tomografías computarizadas de Tórax

Por su lado, Chilesincáncer es una fundación privada sin fines de lucro fundada el año 2016, que busca disminuir la desigualdad de oportunidades frente al cáncer. La fundación articula alianzas público-privadas para entregar mejores oportunidades de diagnóstico y tratamiento frente al cáncer para los adultos atendidos en el sistema público de salud. El propósito de la organización se declara: *"Que todos los chilenos tengan una atención oportuna y de calidad frente al cáncer"*, además, su misión corresponde a: *"Ser un referente para implementar modelos exitosos para la atención de pacientes con cáncer"*. Finalmente, la misión de Chilesincáncer corresponde a: *"Generar las mejores condiciones de infraestructura, capital humano y gestión para la atención de cáncer en Chile"*

1.1.1. Estrategias a nivel mundial

A nivel mundial, la Organización Mundial de la Salud (OMS), plantea seis líneas estratégicas para el control del cáncer:

1. **Prevención**: Controlar el consumo de tabaco, uso y abuso de alcohol, promover alimentación saludable, control de infecciones relacionadas o causantes de ciertos tipos de neoplasias (tumores).

2. **Detección temprana**: Tamizaje adaptado a la realidad local.

3. **Diagnóstico y tratamiento**: Con énfasis en proveer los recursos humanos y tecnologías necesarias para la correcta detección y manejo, teniendo en cuenta los patrones regionales de comportamiento de las enfermedades.

4. **Alivio del dolor y cuidados paliativos**: Se promueve que los principios del sistema de cuidados paliativos deben aplicarse tan temprano como sea posible, a todo paciente afectado por una enfermedad crónica potencialmente fatal.

5. **Investigación**; Considerando las distintas etapas que van desde los aspectos biomédicos básicos, hasta la evaluación poblacional de la aplicación de políticas de la salud

6. **Vigilancia epidemiolígica del cáncer**: Basada en la estructura de una agencia central que sistematice los datos nacionales

Esta enfermedad se encuentra en el foco mundial, por lo que grandes organizaciones y agencias como la Organización para la Cooperación y el Desarrollo Económico (OCDE) y la *European Partnership Action Against Cancer* (EPAAC) sugieren lineas de acción permanentes y guías para el control del cáncer, que MINSAL complementa con la OMS, y sistematiza de la siguiente manera en la figura 1.2. En esta misma se aprecian cuatro líneas transversales a las etapas del plan de manejo y control inherentes a la historia natural (evolución) de la enfermedad.

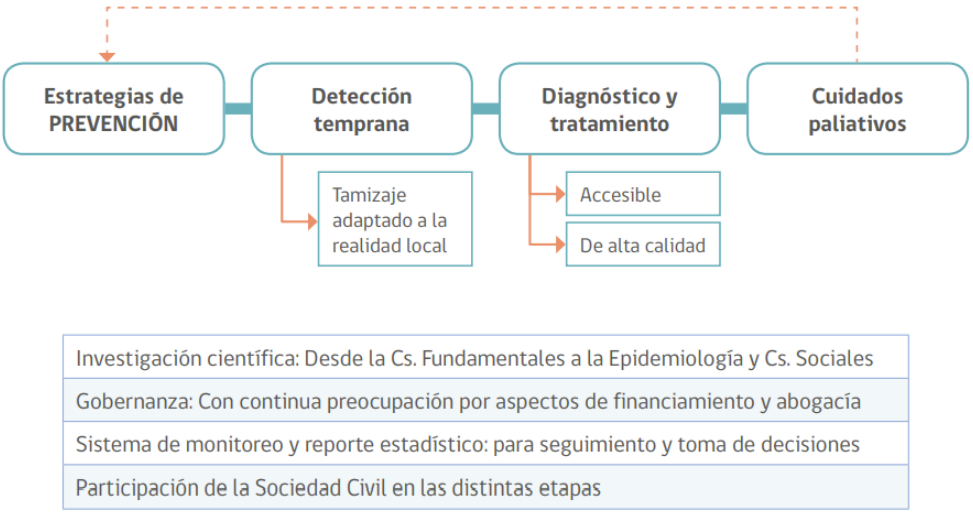

Figura 1.2: Sistematización de los énfasis de la OMS para el control del cáncer según sus distintas fuentes

En esta misma línea, el proyecto se enfoca en la etapa de **detección temprana**, fortaleciendo la pesquisa realizada para el **cáncer de tipo pulmonar**, siempre teniendo en cuenta las líneas transversales para un proyecto de control de cáncer.

1.1.2. Estrategias a nivel nacional

El cáncer de cualquier tipo, con la importancia e impacto que conlleva esta emergencia socio-sanitaria, es abordado por el Ministerio de Salud (MINSAL), a través del Plan Nacional de Cáncer 2018-2028, para abordar el problema de manera estratégica, contando con un plan de acción que permita el logro de los objetivos propuestos para la atención oportuna de esta patología en todo el país, con una mirada integral en la forma de enfrentar esta enfermedad, incluyendo la promoción de la salud y prevención, **detección precoz, diagnóstico oportuno de la enfermedad**, tratamiento adecuado, cuidados paliativos, seguimiento y

rehabilitación, garantizando el acceso a la atención que involucre al conjunto de la sociedad chilena.

1.1.3. Cáncer en HSR

El diagnóstico precoz se desarrolla en conjunto con el Hospital Dr. Sótero del Rio (HSR), establecimiento de salud de nivel terciario del Servicio Metropolitano Sur Oriente (SSMSO), que atiende una de las zonas más pobladas del país (Macrorregión centro). En este hospital, así como en el resto del país, el diagnóstico de cáncer de pulmón tiene grandes desafíos: **Un 50 % de los(as) pacientes oncológicos(as) reciben la primera atención relacionada con su cáncer en el servicio de urgencias, ya sea con síntomas no controlados o situaciones clínicas de riesgo vital**, lo que implica un diagnóstico tardío y aumenta la probabilidad de mortalidad relacionada con el cáncer. Esto se ve potenciado negativamente por varios factores, entre los que se cuentan una alta demanda asistencial y el aumento de las listas de espera asociado a la pandemia de la COVID-19

Según lo levantado a través de reuniones con el Hospital, lo anterior se ve potenciado por la alta demanda asistencial en los Sistemas de Salud, especialmente la red a la cual pertenece HSR (Red Metropolitana Sur Oriente), esto incurre en grandes listas de esperas para Oncología. Según el MINSAL, a nivel país las listas de esperas aumentaron aproximadamente en 81 % entre 2020 y 2022, desde un total de 4 mil personas a 7.262.

Por otro lado, y en base a lo levantado con el Servicio de Salud asociado al Hospital, otros factores que aportan negativamente al manejo de cáncer dentro de HSR corresponden a procesos de derivación de un paciente de una especialidad a otra ineficientes, la existencia de largos tiempos de espera y escasa capacidad de atención en relación a las necesidades de la población. Lo cual provoca que **pacientes con sospecha de cáncer diagnosticada no logren ingresar a tiempo a un tratamiento que le ayude mejorar de su enfermedad**.

1.1.4. Modelo de trazabilidad de Cáncer de pulmón

En 2020, se implementa en el Hospital un **modelo de trazabilidad de pacientes con cáncer de pulmón**, que busca identificar oportunamente personas con sospecha de cáncer pulmonar y realizar un seguimiento de su proceso diagnóstico-terapéutico, mediante la implementación de una herramienta tecnológica para monitorizar la gestión clínica de cada paciente (*CoTalker*), incorporando la optimización del proceso de diagnóstico, tratamiento y seguimiento, en base a gestión integral de casos por una profesional no-médico (enfermera). Hasta agosto de 2022, este modelo ya se ha utilizado exitosamente en 172 pacientes, mejorando la coordinación entre los diferentes agentes implicados en su gestión y posibilitando una atención más expedita y oportuna. Sin embargo, la derivación de pacientes a este modelo no es la óptima, debido en parte a la alta carga de trabajo que enfrentan los profesionales a cargo de remitir casos sospechosos (ej. servicios de urgencia, atención primaria, radiología, etc.). Por otro lado, este proceso es poco estandarizado, y las derivaciones se realizan por diferentes motivos, momentos, actores y unidades dentro del hospital.

1.2. Motivación del proyecto

Para abordar el grave problema sanitario que corresponde el cáncer de pulmón, se busca potenciar el modelo de trazabilidad implementado en el Hospital Sótero del Río. Para ello, se propone la utilización de técnicas de *Machine Learning* y Procesamiento de Lenguaje Natural (NLP, por su sigla en inglés) sobre el texto clínico no estructurado del reporte de Tomografías Computarizadas de Tórax ("TAC" o "TC" de Tórax), con la finalidad de generar un modelo analítico de sospecha de cáncer de pulmón, que sea capaz de procesar grandes volúmenes de informes de TC. Mejorando el *screening* realizado a los pacientes. Se debe destacar que un TC. de Tórax corresponde a una de las vías más frecuentes mediante la cual se investiga la posibilidad o sospecha de un cáncer de pulmón.

1.3. Justificación del Problema

La necesidad fundamental del sistema de detección de sospecha de cáncer de pulmón radica en el hecho de que los médicos pueden pasar por alto, o no darse cuenta a tiempo de las primeras señales de sospecha presentes en los informes de tomografías de tórax. Esto resulta en una demora en el inicio del tratamiento, lo cual es crucial. La *American Cancer Society* sostiene que "el tratamiento temprano del cáncer de pulmón puede aumentar las posibilidades de éxito en el tratamiento y mejorar las perspectivas a largo plazo" ("American Cancer Societ", n.d.).

Existen diversas causas que contribuyen a que los médicos no se percaten a tiempo de estas sospechas. Una de ellas es que los pacientes pueden no acudir a la consulta de oncología por diferentes motivos, como dificultades para acceder a la atención, falta de tiempo o listas de espera en los servicios de oncología. Además, hay casos en los que se realizan tomografías de tórax por otras enfermedades, como COVID-19, o por chequeos rutinarios, y estos exámenes pueden mostrar indicios tempranos de cáncer de pulmón, pero nadie los revisa en busca de esas sospechas.

En resumen, la justificación del sistema de detección de sospecha de cáncer de pulmón se basa en la necesidad de identificar oportunamente las primeras señales de sospecha presentes en los informes de tomografías de tórax, ya que los médicos pueden pasar por alto esta información debido a diversas razones, lo que resulta en una demora en el inicio del tratamiento y, potencialmente, en un peor pronóstico para los pacientes.

1.4. Objetivos

1.4.1. Objetivo General

El objetivo general del proyecto se plantea: *"Generar un modelo que detecte sospechas de cáncer de pulmón a través de técnicas de procesamiento de lenguaje natural, con la finalidad*

de priorizar casos sospechosos en el Hospital Doctor Sótero del Río, optando a un tratamiento oportuno y eficaz para el paciente oncológico."

1.4.2. Objetivos Específicos

Para llevar a cabo lo anterior, se detallan los siguientes objetivos específicos:

1. Realizar un levantamiento del proceso actual, así como una recopilación de datos para la predicción.
2. Construir los modelos de predicción de cáncer de pulmón, utilizando técnicas de procesamiento de lenguaje natural.
3. Evaluar el desempeño de los modelos en base a métricas clave.
4. Presentar los resultados y brindar recomendaciones a las partes involucradas.

1.5. Metodología

El trabajo realizado comienza con una profunda revisión bibliográfica sobre el Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés), además de su utilización en Oncología, particularmente para la predicción de cáncer. Posterirmente se emplea la metodología *Proceso Estándar Cross-Industry para Minería de Datos*, comúnmente abreviada CRISP-DM. Este proceso consta de 6 etapas que determinan el ciclo de vida de un proyecto de minería de datos, esta secuencia sin embargo no lineal, dado que siempre es necesario volver a iterar sobre etapas anteriores, en la figura 1.3 las flechas representan el flujo más común e importante entre las fases de la metodología. Además, dado que un proyecto de minería de datos no termina cuando se despliega el mejor modelo creado, si no la misma evaluación y despliegue pueden revelar nuevos *insights* del negocio o proceso.

A continuación, se detallan brevemente las 6 etapas de la metodología CRISP-DM:

1. **Entendimiento del negocio:** La fase inicial corresponde a entender los objetivos del proyecto y los requerimientos desde una perspectiva del negocio.
2. **Entendimiento de los datos:** Corresponde a una recopilación de datos y familiarización de estos a través de actividades que permiten determinar problemas en la calidad, y detectar *insights* sobre estos.
3. **Preparación de los datos:** Construcción del *dataset* que utiliza el modelo.
4. **Modelamiento:** En esta fase se seleccionan y aplican las distintas técnicas de modelamiento, así como los parámetros para ser calibrados a sus valores óptimos.
5. **Evaluación:** Corresponde a la evaluación de los modelos en base a si estos resuelven las necesidades del negocio

6. **Despliegue:** Dependiendo de los requerimientos del cliente, la etapa de despliegue puede ser tan simple como generar un reporte o tan compleja como la implementación de un procesos de minería de datos repetible a a través de la organización.

El proceso de CRISP-DM aplicado al contexto del problema presentado dentro del hospital se detalla a continuación:

- **Comprensión del proceso:** En primer lugar, la comprensión del comienza con entender el contexto de HSR dentro de la Red de Salud y el cáncer de pulmón en Chile. Posteriormente, mediante entrevistas y visitas al Hospital, se realiza el levantamiento de información para realizar el modelado del proceso de admisión de un paciente con cáncer de pulmón, con la finalidad de determinar las fuentes de datos para la clasificación, que consideran los informe de radiología (TACs) y su respectivas etiquetas para el entrenamiento y evaluación del modelo, esto quiere decir, una fuente que evidencie una sospecha de cáncer pulmonar dentro de la tomografía estudiada.

- **Comprensión de los datos:** Se realiza una recopilación de los datos, estudiando estos a través de un análisis exploratorio de los datos (EDA). En primer lugar se estudian las variables sin texto, como la distribución de edad de los pacientes, la serie de tiempo asociada a la demanda de TC de tórax y otros códigos asociados a los reportes de radiología. Luego se estudia el texto clínico, analizando el formato y estructura de este.

- **Preparación de datos:** Se preparan los datos a través de transformaciones adecuadas para los modelos a realizar. Esto considera un pre-procesamiento que elimine tildes y caracteres especiales dentro del texto, para luego generar los *embeddings* adecuados para utilizar como *input* para los modelos de Redes Neuronales. Para los modelos de árboles de decisión se realiza la vectorización del texto adecuada según el experimento.

- **Modelado:** Se seleccionan y construyen los modelos de predicción a utilizar, en base a la revisión bibliográfica realizada, es decir, *Random Forest*, Redes Convolucionales y Redes Neuronales Recurrentes de arquitectura LSTM.

- **Evaluación:** Evaluar el rendimiento de los modelos en la detección de sospecha de cáncer de pulmón, buscando aquel con el mejor desempeño en las métrica clave a dado el contexto del proyecto, descrita en la sección 2.1.7 Métricas de Evaluación, asegurando su generalización a través de la generación de conjuntos de entrenamiento, validación y testeo.

En la figura 1.3 se presenta el esquema de la metodología planteada, se debe tener en consideración que el proceso corresponde a un ciclo, debido a la naturaleza de los proyectos de *data mining*, además dentro del mismo se deben iterar ciertas etapas, en particular, una correcta comprensión del proceso es necesaria para de una correcta comprensión de los datos, y viceversa, sumado a lo mismo, la preparación de los datos y la creación de los modelos dependen uno del otro.

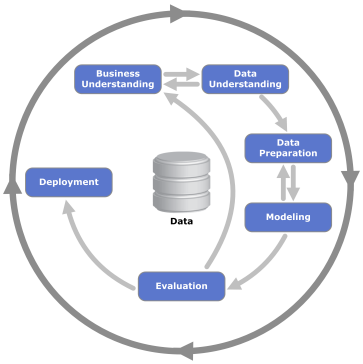

Figura 1.3: Diagrama metodología CRISP-DM

1.6. Alcances

En la realización del proyecto, se construirá un modelo de *machine learning* que identifique o detecte sospechas de cáncer de pulmón a través de reportes de radiología, imitando la "mirada del experto", pudiendo así **priorizar** la derivación de pacientes, brindando un tratamiento oportuno mediante una detección precoz.

En cuanto a los alcances de la memoria, esta considera únicamente la creación y evaluación de los modelos de sospecha de cáncer de pulmón, sin abordar el despliegue oficial de la herramienta para su uso diario. Además, se presentarán los resultados de esta al equipo médico del HSR y SSMSO.

1.7. Resultados esperados

A raíz del objetivo plantedo, se declara como resultados esperados, la creación de modelos analítico para detectar sospecha de cáncer de pulmón para HSR, el cual permita identificar correctamente a pacientes que necesiten una pronta derivación a la unidad oncológica.

Capítulo 2

Marco Conceptual

2.1. Procesamiento de Lenguaje Natural y Redes Neuronales

El Procesamiento de Lenguaje Natural (NLP) corresponde al campo que diseña métodos y algoritmos que toman como *input* o producen como *output* **datos de lenguaje natural** no estructurado (Golgberg, 2017), con la finalidad de resolver alguna tarea o *task* en particular, entre ellas se destacan el Reconocimiento de Entidades Nombradas (NER), y *Text Classification* o Clasificación de Texto. NER busca asignar categorías pre-definidas conocidas como entidades, en texto clínico puede ser utilizado para identificar en qué parte es del texto corresponden a enfermedades, personas, etc. Por otro lado *Text Classification* corresponde a analizar el texto en su totalidad, para luego asignar una etiqueta dentro de un set de categorías pre-definidas, como se puede apreciar, este tipo de tarea recibe una secuencia $x_1, .., x_n$ y asigna una etiqueta y, mientras que NER tiene como *output* una secuencia de etiquetas del mismo largo que el *input*, es decir $y_1, .., y_n$.

Hasta 2014 el estado del arte de NLP se basaba en la creación de atributos que representaran el texto (*feature engineering*) en conjunto con modelos de aprendizajes de máquinas como lo son Support Vector Machines (SVM) o Hidden Markov Models (HMM) (Golgberg, 2016). En la actualidad se utilizan modelos basados en aprendizaje profundo, que no requieren de la creación de atributos, pues realiza de manera automática representaciones del texto eficaces.

A continuación se abordarán los tópicos relevantes a la construcción de los modelos a utilizar, así como las métricas de evaluación.

2.1.1. *Text Vectorization*

Text Vectorization corresponde al proceso de convertir texto en una representación numérica con la finalidad de ser utilizado como *feature*. Una representación común es el *Bag of Words* (BOW), que captura la frecuencia con la que aparecen las palabras. Una desventaja que posee esta representación, corresponde a la pérdida del orden de las palabras, y por lo tanto, el significado lingüístico de la oración.

Para modelar que tan significativo es un término o palabra dentro de un documento, se utiliza la representación *Term Frecuency - Inverted Document Frecuency*. Siendo $tf_{i,j}$ la frecuencia del término t_i en el documento d_j. Definiendo N el número de documentos en el corpus, y n_i el numero de documentos que contienen el término t_i, se define la frecuencia invertida del documento:

$$idf_{t_i} = log_{10}(\frac{N}{n_i}) \tag{2.1}$$

Con lo cual un término que aparece en todos los documentos tiene *idf* igual a 0. Finalmente, el puntaje TF-IDF se calcula:

$$w(t_i, d_j) = tf_{ij} \times log_{10}(\frac{N}{n_i}) \tag{2.2}$$

Otro tipo de representación vectorial del texto corresponde a *word embeddings*, que serán abordados en la sección 2.1.3

2.1.2. Redes neuronales *Fully-Connected*

La arquitectura básica de Redes Neuronales corresponde al *Perceptrón*, el cual es una función lineal de sus *inputs*:

$$\begin{gathered} NN_{Perceptron}(x) = xW + b \\ x \in \mathbb{R}^{d_{in}}, W \in \mathbb{R}^{d_{in}} \times \mathbb{R}^{d_{out}}, b \in R^{d_{out}} \end{gathered} \tag{2.3}$$

Una Red Neuronal *Fully-Connected*, *Muti-Layer-Perceptron* o *Feedforward*, corresponde a un conjunto de estas funciones lineales separadas por funciones no-lineales $f(h)$ conocidas como funciones de activación, como se puede apreciar en la figura 2.1. Entre estas se mencionan la función de activación **sigmoide** $\sigma(x) = \frac{1}{1+e^{-x}}$ que transforma cada valor h al rango [0,1] y ReLU, que por otro lado se define como $ReLU(h) = max(0, h)$. El *output* también puede ser transformado por estas funciones, una transformación comun corresponde a *softmax*, donde para un vector $(x_1, ..., x_k)$:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{k} e^{x_j}} \tag{2.4}$$

Que entrega por resultado un vector no negativo de números reales que suman uno, pudiendo ser utilizado como una distribución de probabilidad discreta sobre k posibles resultados.

El entrenamiento de este tipo de modelos se realiza usando el descenso de gradiente (en general se usa el estocástico) para minimizar el error entre el output predicho y el output deseado, para ello se necesitan datos de entrenamiento con sus etiquetas. A continuación se detalla el método de gradiente estocástico para redes neuronales

- Los parámetros de la red $\Theta = [W, \vec{b}]$ son inicializados de manera aleatoria.
- Para cada ejemplo (x, y) se calcula su función de pérdida L que calcula cuanto se acerca

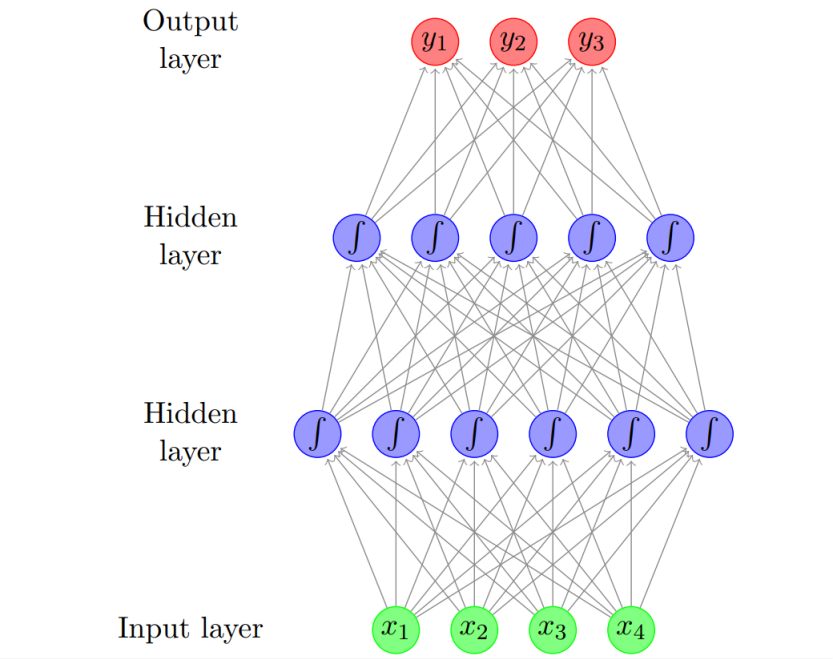

Figura 2.1: Red Neuronal *feed-fordward* con dos capas ocultas

la predicción $\hat{y}$ al valor real y, con los valores actuales de los parámetros de la red $\Theta = [W, \vec{b}]$.

- Se actualizan los parámetros hasta converger según la siguiente regla:

$$\Theta_i \leftarrow \Theta_i - \eta \frac{\partial L}{\partial \Theta}(\hat{y}, y) \tag{2.5}$$

donde η corresponde a un hiperparámetro denominado tasa de aprendizaje, que determina la velocidad de actualización, y $\frac{\partial L}{\partial \Theta}(\hat{y}, y)$ el gradiente función de pérdida L con respecto a Θ, el cual se calcula a traves del algoritmo de *backpropagation*

Debido a que la actualización de los parámetros para cada red para cada ejemplo resulta costoso y lento, se toma un conjunto de muestras o *batches* para los cuales se calcula el promedio del error, actualizando los parámetros para cada *batch*.

2.1.3. Word Embeddings

Para utilizar el texto como *input* para los modelos, se debe tratar con la *sparseness* del lenguaje, esto se refiere a que al haber una gran cantidad de palabras en el vocabulario del conjunto de textos (corpus), la representación de conteo de palabras como Bolsas de Palabras (BoW) o Matriz Palabra-Contexto generan atributos para los modelos con muy pocos valores distintos de 0, además de tener una alta dimensionalidad. Comúnmente, para Redes Neuronales se utiliza la representación de *word embeddings*, los cuales son vectores densos de baja dimensionalidad creados a partir de un mapeo de símbolos discretos (Jurafsky, D. and Martin, J. H., 2008). Existen dos maneras de obtener *word embeddings*:

- **Embedding layers**: Utilizar una capa de *embedding* dentro de una red neuronal para una tarea específica entrenada a partir de ejemplos etiquetados.
- **Embeddings pre-entrenados**: Solucionar una tarea predictiva a partir de corpus no etiquetados, como por ejemplo predecir la siguiente palabra en una oración, en donde los *words embeddings* se extraen desde la arquitectura entrenada con una gran cantidad de datos.

Estos dos enfoques se pueden utilizar en conjunto, utilizando *embeddings* pre-entrenados dentro de la red, y además utilizar una capa de *embeddings*. Finalmente, se destaca la utilización de *embeddings* contextualizados, los cuales se extraen a partir de arquitecturas complejas del estado del arte en Redes Neuronales, como lo son Bidirectional Encoder Representations from Transformer (BERT) o Embeddings from Language Model (ELMo), y permiten obtener *embeddings* sensibles al contexto, generando distintas representaciones para palabras que se escriben igual pero tienen distinto significado.

2.1.4. Redes Neuronales Convolucionales

Las redes neuronales convolucionales (CNNs) son una arquitectura de red para encontrar patrones dentro de la estructura del input, y por lo tanto, es la red predilecta a utilizar para resolver tareas de reconocimiento de imágenes, permitiendo identificar patrones sin importar en que posición en la que estos ocurren. Una CNN utilizada en NLP captura los n-gramas (puede ser de palabras o caracteres) que aportan mayor información para la tarea predictiva.

Al utilizar este tipo de arquitecturas en el contexto de NLP, se tienen 2 operadores o capas dentro del contexto de red neuronal, estos son la capa Convolucional y la capa de *Pooling*. Recordando que una oración o documento es modelada como una secuencia de *word embeddings*, entonces:

- **Capa Convolucional:** La capa convolucional aplica funciones no-lineales o filtros mapeando una ventana de k palabras (*k-gramas*) a valores escalares, estos filtros capturan propiedades relevantes de las palabras en la ventana. Sobre texto, este tipo de operación convolucional se realiza en una única dimensión, aplicando la misma función paramatrizada sobre todos los *k-gramas* de la secuencia. Esto crea una representación de m vectores, sensibles al orden de las palabras dentro de la ventana y que aportan información para la tarea predictiva. Una limitante de este método, corresponde a que se puede extraer la misma representación para un *k-grama* independiente de su pocición en la secuencia.
- **Capa Pooling:** La capa de *pooling*, como dice su nombre es una combinación de los m vectores resultantes de la capa convolucional, para construir un único vector $\vec{c}$ que captura la información más importante dentro de la secuencia para la tarea. Para ello se puede utilizar la operación *max pooling* donde se toma el máximo de cada dimensión, o *average pooling* que toma el promedio de los valores para cada índice

Posteriormente, el vector $\vec{c}$ se utiliza como *input* en otra capa, por ejemplo una capa *Fully-Connected*, la cual puede ser utilizada para la predicción necesaria. El proceso de entrenamiento para una CNN utiliza la función de pérdida L para optimizar los parámetros del filtro de la convolución, así como los pesos de la ecuación 2.3, a través de todas las capas mencionadas.

2.1.5. Redes Neuronales Recurrentes

Las redes neuronales recurrentes (RNNs) tienen una arquitectura que el permite capturar las propiedades asociadas a la estructura de secuencia de los *inputs*, en un único vector, en particular las arquitecturas con compuertas tales como LSTM y GRU, capturar las regularidades estadísticas en la secuencia de entrada (u oración) (Goldberg,2016). Lo anterior se debe a la recurrencia dentro de la arquitectura, donde para calcular cada vector de estado $\vec{s}_i$, se utiliza el vector de estado anterior $\vec{s}_{i-1}$ y el *input* $\vec{x}_i$, de la manera:

$$\begin{aligned} RNN^*(\vec{x}_{1:n}; \vec{s}_0) &= \vec{y}_{1:n} \\ \vec{y}_i &= O(\vec{s}_i) \\ \vec{s}_i &= R(\vec{s}_{i-1}, \vec{x}_i) \\ \vec{x}_i \in \mathbb{R}^{d_{in}}, \quad \vec{y}_i \in \mathbb{R}^{d_{out}}, \quad \vec{s}_i &\in \mathbb{R}^{f(d_{out})} \end{aligned} \tag{2.6}$$

Considerando R la función recurrente y O la función para obtener el input del estado i, como se puede apreciar en la figura:

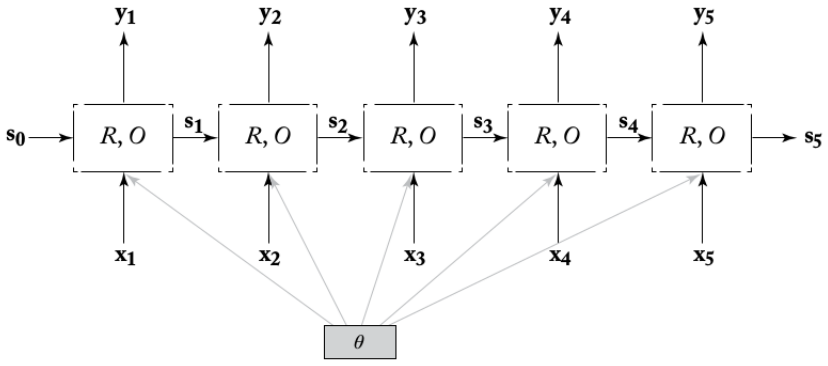

Figura 2.2: Arquitectura RNN para una secuencia finita

La formulación mas simple de una RNN se conoce como Red de Elman:

$$\begin{aligned} \vec{s}_i &= R_{SRNN}(\vec{x}_i, \vec{s}_{i-1}) = g(\vec{s}_{i-1}W^s + \vec{x}_i W^x + \vec{b}) \\ \vec{y}_i &= O_{SRNN}(\vec{s}_i) = \vec{s}_i \\ \vec{s}_i, \vec{y}_i \in \mathcal{R}^{d_s}, \quad \vec{x}_i \in \mathcal{R}^{d_x}, \quad W^x \in \mathcal{R}^{d_x \times d_s}, \quad W^s &\in \mathcal{R}^{d_s \times d_s}, \vec{b} \in \mathcal{R}^{d_s} \end{aligned} \tag{2.7}$$

Donde se transforma linealmente cada estado $\vec{s}_i$ e *inputs* $\vec{x}_i$, para luego ser sumados junto a un sesgo $\vec{b}$, calculando el estado siguiente a través de una función de activación g, comúnmente *tanh* o *ReLU*. Esta arquitectura provee un buen desempeño para las tareas de *sequence tagging* como lo es NER.

Sin embargo, estas arquitecturas sufren del problema de desvanecimiento del gradiente debido a la multiplicación repetida de la matriz W, para ello se utiliza el mecanismo de compuerta para permitir a cada celda de la red *olvidar* información anterior. Se utilizará en los modelos la arquitectura Long Short-Term Memory (LSTM) (Hochreiter, S. and Schmidhuber, J. 1997), que se define:

$$
\begin{aligned}
\boldsymbol{s_j} = R_{\text{LSTM}}(\boldsymbol{s_{j-1}}, \boldsymbol{x_j}) &= [\boldsymbol{c_j}; \boldsymbol{h_j}] \\
\boldsymbol{c_j} &= \boldsymbol{f} \odot \boldsymbol{c_{j-1}} + \boldsymbol{i} \odot \boldsymbol{z} \\
\boldsymbol{h_j} &= \boldsymbol{o} \odot \tanh(\boldsymbol{c_j}) \\
\boldsymbol{i} &= \sigma(\boldsymbol{x_j}\boldsymbol{W^{xi}} + \boldsymbol{h_{j-1}}\boldsymbol{W^{hi}}) \\
\boldsymbol{f} &= \sigma(\boldsymbol{x_j}\boldsymbol{W^{xf}} + \boldsymbol{h_{j-1}}\boldsymbol{W^{hf}}) \\
\boldsymbol{o} &= \sigma(\boldsymbol{x_j}\boldsymbol{W^{xo}} + \boldsymbol{h_{j-1}}\boldsymbol{W^{ho}}) \\
\boldsymbol{z} &= \tanh(\boldsymbol{x_j}\boldsymbol{W^{xz}} + \boldsymbol{h_{j-1}}\boldsymbol{W^{hz}})
\end{aligned}
$$

$$
\boldsymbol{y_j} = O_{\text{LSTM}}(\boldsymbol{s_j}) = \boldsymbol{h_j}
$$

$$
\boldsymbol{s_j} \in \mathbb{R}^{2 \cdot d_h}, \ \boldsymbol{x_i} \in \mathbb{R}^{d_x}, \ \boldsymbol{c_j}, \boldsymbol{h_j}, \boldsymbol{i}, \boldsymbol{f}, \boldsymbol{o}, \boldsymbol{z} \in \mathbb{R}^{d_h}, \ \boldsymbol{W^{xo}} \in \mathbb{R}^{d_x \times d_h}, \ \boldsymbol{W^{ho}} \in \mathbb{R}^{d_h \times d_h}.
$$

Figura 2.3: Formulación matemática arquitectura LSTM

Donde en el estado j, el vector $\vec{c}_j$ representa el componente de memoria y $\vec{h}_j$ el estado oculto. Las compuertas $\vec{i}$, $\vec{f}$ y $\vec{o}$ controlan el ***i**nput*, ***f**orget* u olvidar, y ***o**utput*.

En la figura 2.4[1] se puede observar un esquema del funcionamiento de una celda LSTM

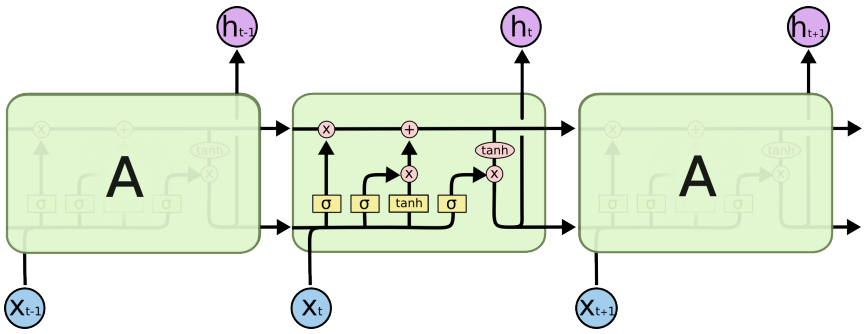

Figura 2.4: Celda LSTM

[1] fuente: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Para realizar una predicción considerando una tarea de clasificación, se requiere utilizar una red *feed-fordward* en conjunto con una función de activación *softmax* que reciba como *input* el estado final de RNN, cuyo *output* en la última capa sea de igual a la cantidad categorías a predecir. Suponiendo una red *fully-connected* de una capa (MLP), la etiqueta k a predecir dado un documento $\vec{w}_{1:n}$ se calcula:

$$\begin{aligned} p(label = k|\vec{w}_{1:n}) &= \hat{\vec{y}}_{[k]} \\ \hat{\vec{y}} &= softmax(MLP(RNN(\vec{x}_{1:n}))) \\ \vec{x}_{1:n} &= E_{[w_1]}, \ldots, E_{[w_n]} \end{aligned} \tag{2.8}$$

Donde E corresponde a la matriz de *embeddings*.

2.1.5.1. Encoder-Decoder

El poder predictivo de las RNNs, se ve incrementado de gran manera al ser utilizado en una estructura *encoder-decoder*. Esto se realiza utilizando dos RNNs, en donde la RNN que corresponde al *encoder* codifica el input a un vector $\vec{c}$, mientras que el *decoder* decodifica el output del *encoder*, generando la predicción, lo cual se aprecia en la figura 2.5:

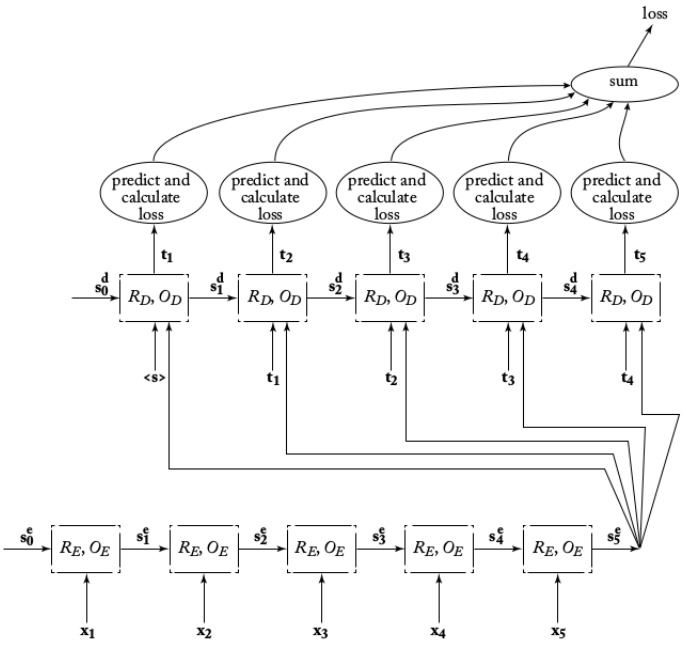

Figura 2.5: Entrenamiento arquitectura *encoder-decoder*

Puesto que en la arquitectura de *encoder-decoder* el vector $\vec{c}$ contiene toda la información para la generación de nuevas secuencias, esto genera problema para extraer información en oraciones largas, lo cual se resuelve utilizando el mecanismo de *atención* propuesto por Bengio, Y et. al. (2015), este mecanismo permite al *decoder* "mirar hacia atrás" accediendo a los estados ocultos del *encoder* basado en su estado actual. Considerando una secuencia $\vec{x}_{1:n}$, en el estado j, el *decoder* utiliza un promedio ponderado de los vectores $\vec{c}_{1:n}$, en donde los

pesos de *atención* $\vec{\alpha}^j$ (donde la suma de estos pesos es igual a la unidad) se eligen mediante el mecanismo de *atención*, es decir:

$$\vec{c}^j = \sum_{i=1}^{n} \vec{\alpha}^j_{[i]} \cdot \vec{c}_i \tag{2.9}$$

2.1.6. Transformers

La arquitectura del *Transformer* utiliza el mecanismo de *atención* para descartar totalmente la componente de *encoder-decoder* inherente a las RNNs que la componen (Vaswani et al., 2017). Si bien una RNN procesa el input secuencialmente, el *Transformer* permite al *encoder* y *decoder* "ver"la secuencia en su totalidad utilizando la *atención*.

El mecanismo de *atención* utilizado en este modelo se denomina *Multi-Head Attention*, y puede ser interpretado como una forma de cómputo de la relevancia de **valores** (V), basado en **llaves** (K) y ***queries*** (Q). Para aprender diversas representaciones *Multi-Head Attention* aplica distintas transformaciones lineales a estos componentes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.10}$$

Donde Q corresponde a la matriz de *queries*, K y V las matrices de llaves y valores, a su vez d_k es la dimensión K y V.

Dicho esto, la arquitectura de *Transformer* propuesta por (Vaswani et al., 2017) presente en la figura 2.6 utiliza también el diseño de *encoder-decoder*, pero utilizando el mecanismo de atención propuesto en vez de RNNs.

2.1.6.1. *BERT transformer*

El mecanismo de *transformer* permite la construcción del modelo *Bidirectional Encoder Representations From Transformers* o BERT. Esta arquitectura fue desarrollada por Google Devlin et.a al 2018 y cuenta con 12 *encoders* y 12 capas de auto-atención bidireccional. La red se utiliza realizando un *fine tuning* de la capa de *embedding* y la capa de de output para la clasificación de sospecha de cáncer (clasificación binaria). La arquitectura de BERT se muestra en la figura 2.7

2.1.7. Métricas de evaluación

Para evaluar el desempeño de los modelos en la identificación de casos positivos y negativos, es necesario establecer métricas de evaluación. La elección de métricas debe considerar el *trade-off* existente entre una correcta detección temprana y la minimización de diagnósticos erróneos. La sensibilidad y la especificidad son métricas críticas para minimizar los falsos negativos y falsos positivos, respectivamente. Además, el AUC-ROC proporciona una evaluación más completa del rendimiento del modelo en la detección de casos de cáncer de pulmón.

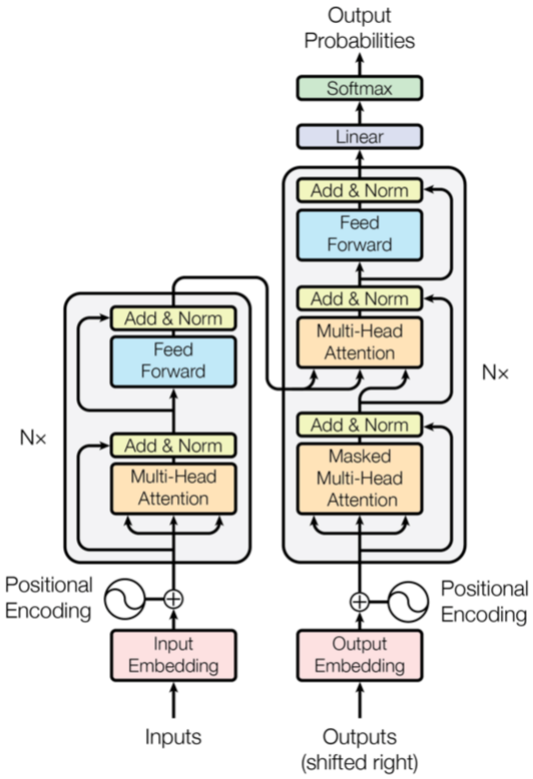

Figura 2.6: Arquitectura de *Transformer*

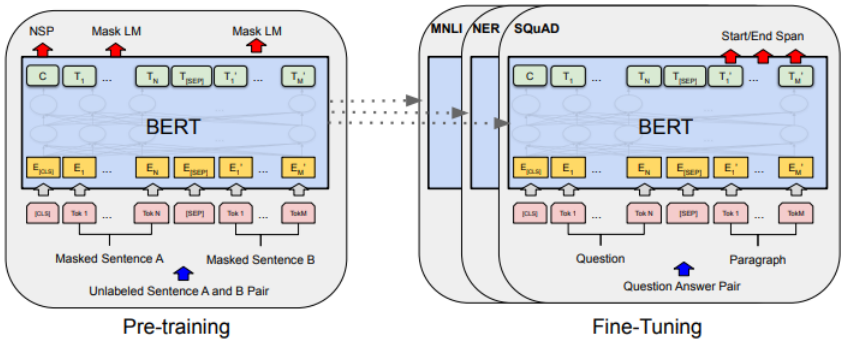

Figura 2.7: Arquitectura BERT

Dada una matriz de confusión (tabla 2.1), que corresponde a una tabla de contingencia de los valores reales, y los valores predichos, las métricas de desempeño se detallan a continuación:

2.1.7.1. Sensibilidad

La sensibilidad (o *Recall* en inglés) mide la capacidad del modelo para detectar positivos reales (pacientes con sospecha de cáncer de pulmón). En un contexto médico, la sensibilidad es crítica, puesto que se requiere **minimizar los falsos negativos**. En otras palabras, se busca evitar que pacientes con cáncer de pulmón sean clasificados erróneamente como negativos, ya

Tabla 2.1: Matriz de Confusión

		Valor predicho		
		p	**n**	**total**
Valor Real	**p′**	Verdadero Positivo	Falso Negativo	P′
	n′	Falso Positivo	Verdadero Negativo	N′
	total	P	N	

que esto podría retrasar el tratamiento y tener consecuencias graves para su salud. Se calcula como:

$$Sensibilidad = \frac{VerdaderosPositivos}{VerdaderosPositivos + FalsosNegativos}$$

2.1.7.2. Precisión

La precisión mide la proporción de predicciones positivas que son verdaderamente positivas. En el contexto médico, una alta precisión es importante para garantizar que los pacientes identificados como sospechosos de cáncer realmente tengan un alto riesgo de la enfermedad. Esto ayuda a evitar procedimientos invasivos innecesarios, como por ejemplo una biopsia. Esta métrica se calcula como:

$$Precisión = \frac{VerdaderosPositivos}{VerdaderosPositivos + FalsosPositivos}$$

2.1.7.3. Área bajo la Curva ROC

La curva ROC y el AUC-ROC evalúan la capacidad del modelo para distinguir entre las clases. Un AUC-ROC cercano a 1 indica un buen rendimiento en la discriminación entre pacientes con y sin cáncer. Es especialmente útil cuando se ajustan umbrales de probabilidad para equilibrar la sensibilidad y la especificidad. Un AUC-ROC de 0.5 indica un rendimiento aleatorio, mientras que un valor cercano a 1 indica un rendimiento excelente en la discriminación.

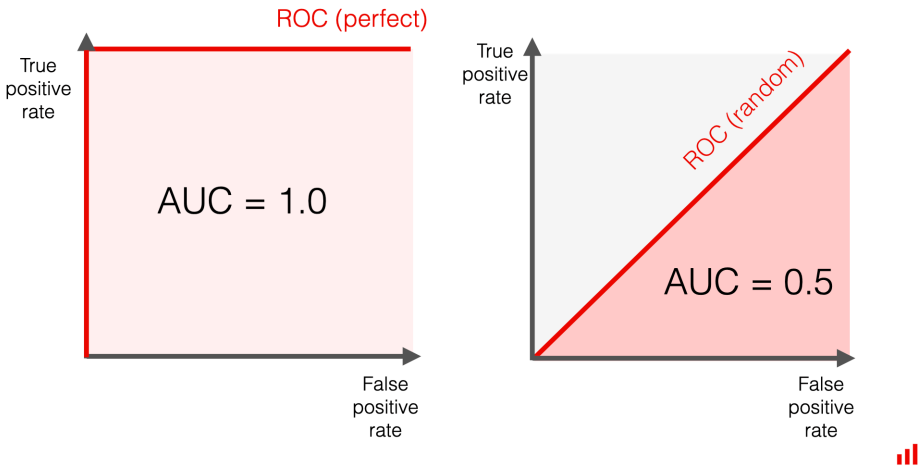

Figura 2.8: Ejemplo AUC-ROC

2.2. Revisión de literatura

La detección y predicción del cáncer de pulmón han sido áreas de interés clave en la investigación médica. Se han aplicado diversas técnicas de aprendizaje automático para mejorar la precisión en esta tarea. Estudios recientes han destacado la efectividad de las redes neuronales convolucionales (CNN) y el aprendizaje profundo en la detección de tumores malignos a partir de imágenes de tomografías computarizadas (Gao et al., 2018). También se ha explorado el análisis radiómico, que implica la extracción de características cuantitativas de imágenes médicas junto con algoritmos de aprendizaje automático, demostrando una precisión significativa en la predicción temprana del cáncer pulmonar (Li et al., 2020).

Además, se ha investigado el uso del procesamiento de lenguaje natural para analizar informes radiológicos basados en texto y predecir la incidencia y recurrencia del cáncer de pulmón (Choi et al., 2018). Estos enfoques han demostrado ser efectivos en la extracción de información relevante y la predicción de resultados clínicos. Otras investigaciones han aplicado técnicas de aprendizaje automático como Support Vector Machines (SVM) y Random Forests para clasificar informes y lograr una precisión significativa en la predicción de la recurrencia del cáncer de pulmón (Nishio et al., 2020).

También se ha utilizado el procesamiento de lenguaje natural para predecir el riesgo de cáncer de pulmón a partir de datos de historias clínicas electrónicas, destacando la importancia de las técnicas de NLP en la identificación de pacientes con mayor riesgo (Tseng et al., 2019). Estos estudios subrayan el potencial de las técnicas de aprendizaje automático y NLP en la mejora de la detección y predicción del cáncer de pulmón, lo que puede tener un impacto significativo en la práctica médica y la atención a pacientes.

Existe una diversidad de enfoques utilizados en la predicción de cáncer pulmonar a través del uso de tomografías computarizadas. Ya sea mediante procesamiento de lenguaje natural o visión computacional, los modelos desarrollados en estos estudios han demostrado la capa-

cidad de detectar y predecir el cáncer de pulmón con resultados prometedores, lo que resalta el potencial de estas técnicas en el campo de la medicina y la oncología.

Se deben mencionar también las limitantes al utilizar técnicas de NLP específicamente para la oncología, entre estas se mencionan la falta de datos etiquetados públicos, la nomenclatura no estándar utilizada en los reportes y problemas de interoperabilidad entre las plataformas tecnológicas utilizadas.

Capítulo 3

Desarrollo metodológico

En este capítulo se detalla el trabajo realizado, abordando las etapas clave de la metodología CRISP-DM especificadas (ver figura 1.3). A continuación, se presenta un esquema conceptual que relaciona las distintas secciones del capítulo:

- Entendimiento del contexto: En esta sección, se profundiza en el contexto del cáncer de pulmón y su impacto a nivel mundial y nacional. Se justifica el problema abordado y se describe la situación actual del cáncer en el hospital donde se lleva a cabo el estudio. Además, se explora la ruta de un paciente con sospecha de cáncer de pulmón, desde su ingreso al hospital hasta el proceso de etapificación del cáncer y determinación del tratamiento.

- Fuentes de datos identificadas: Aquí se presentan las fuentes de datos utilizadas en el estudio, incluyendo los informes de Tomografía Computarizada de Tórax, los ingresos a la plataforma Cotalker, las interconsultas y las biopsias. Se discute la importancia de utilizar estas fuentes de datos para identificar automáticamente las sospechas de cáncer de pulmón y se describe la construcción de la etiqueta de sospecha.

- Comprensión de los datos: En esta sección, se analizan en detalle las fuentes de datos utilizadas. Se describe la estructura y el contenido de los informes de Tomografía Computarizada de Tórax, se estudia la demanda de estos exámenes y se analiza la distribución de la edad de los pacientes. Además, se examinan los n-gramas más comunes en los reportes de Tomografía para comprender mejor el texto clínico.

- Preparación de los datos: Aquí se aborda el proceso de preparación de los datos, que incluye el cruce entre las diferentes fuentes de datos, el preprocesamiento del texto clínico, la partición del dataset en conjuntos de entrenamiento y validación, y el balanceo de clases utilizando técnicas de undersampling y oversampling.

- Modelamiento: En esta sección se describen los modelos utilizados para la clasificación de sospecha de cáncer de pulmón. Se utilizan modelos de redes neuronales convolucionales, redes neuronales recurrentes y transformadores BERT. Además, se implementan

modelos de clasificación basados en árboles de decisión y se exploran diferentes atributos y técnicas de clasificación.

3.1. Entendimiento del contexto

Para entender el contexto en el cual se desarrolla y diseñan los modelos de predicción de sospecha de cáncer de pulmón a implementar, es fundamental primero conocer la **ruta de un paciente con sospecha de cáncer pulmonar**, a través de entrevistas con el equipo médico partícipe.

3.1.1. Ruta de un paciente con sospecha de cáncer de pulmón

Todo comienza con el **ingreso** del paciente al Complejo asistencial HSR, este al ser un centro de salud terciario, está destinado a la ejecución de intervenciones quirúrgicas y atenciones de mayor complejidad y especialización, debido a esto, cuenta con un servicio de urgencia. En consecuencia el Hospital recibe derivaciones de pacientes con sospecha de cáncer de pulmón, en forma de Interconsultas desde otros centros de la Red de Salud o ingresos a través del Servicio de Urgencia Adulto (SUA). A nivel Metropolitano Sur Oriente, se destacan los centros de Atención Primaria de Salud (APS), como lo son los Centros de Salud Familiar (CESFAM), Centro Comunitario de Salud Familiar (CECOSF) o Postas de Salud Rural (PSR).

Una vez el paciente ingresa al Hospital por sospecha de cáncer, ya sea por SUA, interconsultas u otros, se continúa con el proceso de **etapificación** del cáncer. Este se define por el Instituto Nacional del Cáncer de Estados Unidos, como determinar la etapa o estadio del cáncer, refiriéndose a la extensión, tamaño y ramificación de este. Para esto se le solicitan al paciente distintos exámenes, en este caso, la detección de un cáncer broncopulmonar es realizada mediante una Tomografía Computarizada de Tórax, se pueden solicitar además otros tipos de exámenes o el historial médico del paciente para apoyar el diagnóstico del paciente.

Posteriormente, todo paciente con sospecha de cáncer pasa a través del **comité oncológico**, donde se evalúa la etapa del cáncer y se determina el tratamiento a seguir, a continuación se agrupan los tratamientos a realizar para un paciente de cáncer broncopulmonar:

1. **Seguimiento**: la etapa del cáncer y/o tamaño no requieren una atención inmediata por lo que el paciente entra a seguimiento.

2. **Cirugía**: realización de la operación quirúrgica, que tambien considera la radiocirugía, y puede culminar en el seguimiento del paciente tratado, o la exclusión del cáncer de cáncer de células pequeñas de pulmón (CCPP).

3. **Oncología**: atención de especialidad del cáncer, el paciente puede acudir a presentar exámenes o después de la cirugía, para ser derivado a los tratamientos anteriormente mencionados.

Puede ocurrir excepciones que finalizan el proceso, tales como cáncer descartado o un eventual fallecimiento del paciente.

3.1.2. Modelo de Trazabilidad y *Cotalker*

El principal problema mencionado en la sección 1.1.4 del Modelo de trazabilidad de cáncer de pulmón, corresponde a la derivación e ingreso de pacientes a este modelo. Hasta la fecha, no se efectúa un protocolo definido para derivación de pacientes con sospecha de cáncer de pulmón al modelo de trazabilidad, considerando la importancia de priorizar y tratar lo antes posible esta enfermedad.

Dicho esto, el flujo del modelo de trazabilidad, comienza cuando un paciente con **sospecha de cáncer de pulmón** es ingresado en la plataforma *Cotalker*, como se puede apreciar en la figura 3.2):

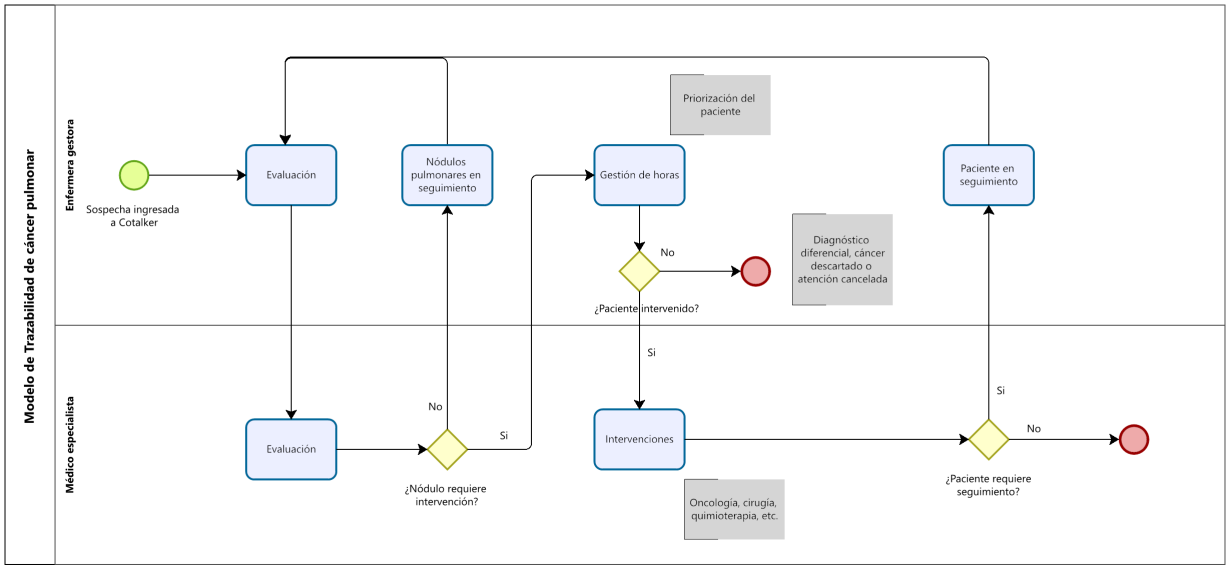

Figura 3.1: Flujo Paciente Cáncer Broncopulmonar

Posterior al ingreso a la plataforma, los gestores de casos analizan los exámenes y antecedentes del paciente, si se confirma como sospechoso, estos son luego revisados por el Médico Especialista, en este caso el Cirujano de Pulmón del Hospital. En caso de requerir de una pronta conducta por especialista como cirugía u oncología, se realiza la gestión de enfermería para priorizar la toma de horas del paciente y mejorar la comunicación con el mismo.

El BPMN se resume en el siguiente diagrama (figura 3.2):

Como se puede apreciar, cada etapa puede llegar a requerir una ventana no menor de tiempo al considerar la listas de espera en la especialidad, la disponibilidad para la realización de los exámenes solicitados, entre otros.

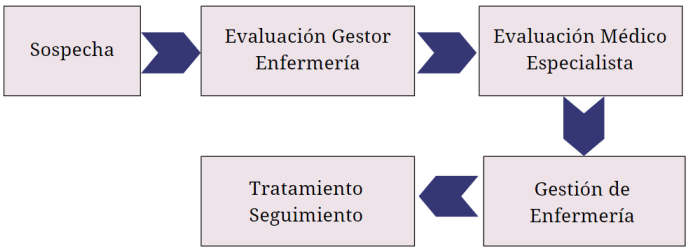

Figura 3.2: Diagrama simplificado modelo de trazabilidad

3.1.3. Fuentes de datos identificadas

A través del entendimiento del contexto realizado, se evidencia la necesidad de identificar automáticamente las sospechas de cáncer de pulmón utilizando NLP, a modo de mejorar la pesquisa de esta enfermedad y en consecuencia, que los pacientes puedan optar un tratamiento oportuno. Para resolver dicha tarea, se necesita como insumos *datos etiquetados*, es decir, datos sobre los cuáles el algoritmo pueda predecir si existe o no una sospecha de cáncer de pulmón, y comparar con una etiqueta real.

Naturalmente, los datos necesarios para estudiar la sospecha de cáncer de pulmón en un paciente es el examen de TC. de tórax. La etiqueta o **label** que indica si el paciente realmente presentaba una sospecha de cáncer de pulmón puede ser construida de diversas formas.

De mayor a menor nivel de confianza para la construcción de la etiqueta, el mejor método corresponde a un etiquetado manual de los TACs de tórax, ya sea por el médico especialista o el gestor de enfermería. Otros métodos, corresponden a tomar como fuente de sospecha, que el paciente pase a través de distintas etapas de la ruta del paciente con sospecha de cáncer de pulmón (3.1.1). Indicaría una sospecha entonces, que el paciente se realice una tomografía computarizada de pulmón, y que luego en una ventana de tiempo este: acuda a Oncología, sea presentado en Comité Oncológico, se realice una Biopsia, acuda a cirugías, entre otros.

Se debe recalcar que la fuente más confiable, corresponde a los pacientes que ingresan al flujo de trazabilidad mediante *Cotalker*, debido a que es el destino final de los casos que identificará la solución en un futuro. La limitación que tiene utilizar este método como única fuente de datos recae en la cantidad de ejemplos positivos (con sospecha de cáncer de pulmón), por lo cual se decide utilizar todas las fuentes de datos posibles para etiquetar sospechas.

Con las fuentes de datos identificadas, se crea una etiqueta binaria, es decir, que toma el valor numérico 1 si el paciente tiene una sospecha de cáncer de pulmón y 0 si no.

3.2. Comprensión de los datos

Los datos utilizados fueron proporcionados por la Unidad de Salud Digital del SSMSO[2]. En esta sección, se abordarán las fuentes de datos utilizadas, la construcción de la etiqueta

[2] página web: https://saluddigital.ssmso.cl

y un análisis de la base de datos final.

3.2.1. Fuentes de datos

3.2.1.1. Tomografías Computarizadas de Tórax

Los reportes de Tomografía Axial Computarizada de Tórax (TAC de Tórax) se utilizan como fuente de datos para capturar la información (*features*) para realizar la predicción de una sospecha de cáncer de pulmón. Estos exámenes se encuentran en formato de texto escrito no estructurado, y están compuestos de tres partes:

- **Antecedentes**: Información anterior del paciente, por ejemplo si es fumador o tiene ya cáncer detectado. Este parte resulta de suma importancia pues permiten tener una visión más completa y contextualizada de la situación clínica del paciente, lo que ayuda a mejorar la precisión y la relevancia del informe.
- **Hallazgos**: Descripción de la imagen, tanto de irregularidades como imágenes sin novedad (evaluación de los pulmones, estado del corazón y aorta entre otros).
- **Impresión**: Diagnóstico general realizado por el radiólogo que resume los hallazgos. Corresponde a una evaluación médica fundamentada que puede tener un impacto significativo en el diagnóstico y tratamiento del paciente.

A continuación, se muestran dos ejemplos de reportes de TAC de tórax anonimizados. El primer reporte corresponde a una sospecha negativa de cáncer de pulmón, donde el examen es realizado a partir de un resultado positivo de Covid-19.

El segundo reporte presentado a continuación, se realiza debido a la expulsión de sangre al toser (hemoptisis) y en el cuál se encuentra una sospecha de cáncer pulmonar (nódulo pulmonar derecho):

Como se puede apreciar, los reportes de TAC al ser texto libre, difieren entre sí en lo que respecta a la descripción de enfermedades o hallazgos presentes, un mismo radiólogo por ejemplo, puede referirse a un cáncer de pulmón como nódulo pulmonar o masa pulmonar. Existen casos en los cuáles los no se encuentran las tres partes del relato descritas anteriormente.

Entre las otras variables presentes en el *dataset*, se destacan el ID del paciente, que se utiliza como identificador para mantener la anonimidad de los pacientes, la edad y la fecha en la que se realiza el TAC, entre otras.

3.2.1.2. Ingreso a *Cotalker*

La fuente de datos más fidedigna para levantar una sospecha de cáncer de pulmón corresponde a pacientes ingresados a *Cotalker*. Debido a que en un futuro, los casos considerados como positivos por el algoritmo ingresarían al modelo de trazabilidad, luego de su revisión.

Esta base de datos cuenta con el ID del paciente, la fecha de ingreso, el origen de la sospecha, como urgencias, programa de tabaquismo, etc. Finalmente, el *dataset* cuenta con el

TOMOGRAFÍA COMPUTADA DE TORAX

Antecedentes clínicos: Covid - 19 (+).

Hallazgos:
Pulmones de volumen normal.
No se observan masas ni nódulos en el parénquima pulmonar.
Se observan opacidades parenquimatosas con atenuacion en vidrio esmerilado bilaterales, de distribucion central y periferica, asociado a engrosamiento septal intra e interlobulillar.
No hay derrame pleural ni neumotórax.
Corazón de tamaño y configuración normal.
No hay derrame pericárdico.
Moderada a acentuada ateromatosis aortocoronaria.
Traquea y grandes bronquios de calibre normal, permeables.
No hay adenopatías mediastínicas, hiliares ni axilares.
Aorta, tronco de arteria pulmonar y ramas principales, permeables y calibre normal.
En los cortes complementarios del abdomen no se identificaron alteraciones significativas.

Impresión:
1. Signos de una alveolitis bilateral consistente con neumonia viral (Covid - 19) conocida.

Figura 3.3: Ejemplo TAC sin sospecha

Tomografía computada tórax del 06-12-2021:

Antecedente clínico: Hemoptisis.

Pulmones son de volumen y arquitectura general normal sin identificar foco de condensación ni masas.
Nódulo subpleural subsólido en LM imagen 127 serie alta resolución de 9 mm.
No observo otros nódulos pulmonares sospechosos.
No hay derrame pleural ni neumotórax.
Tráquea y bronquios principales permeables.
Corazón de tamaño normal sin derrame pericárdico.
Aorta torácica calibre normal con ateromatosis.
Tronco arteria pulmonar y vasos supraaórticos de calibre conservado.
No hay adenopatías mediastínicas, hiliares ni axilares según criterios tomográficos de tamaño.
Linfonodo subcarinal de 9 mm.
En los cortes complementarios del abdomen no se identifican alteraciones significado patológico.

Figura 3.4: Ejemplo TAC con sospecha

estado del paciente, siguiendo la notación de cada una de las etapas del flujo de trazabilidad (figura 3.2).

3.2.1.3. Interconsultas

Como fue mencionado anteriormente, el Hospital Sótero del Río, en su calidad de establecimiento de salud terciario, recibe derivaciones de otros establecimientos. Se utiliza como

fuente de datos, las interconsultas que además del ID y la fecha de la derivación, incluyen:

- **Especialidad de origen**: Especialidad de la que fue derivado, por ejemplo, medicina familiar corresponde a interconsulta proveniente de un CESFAM. Las principales especialidades de origen corresponden a oncología médica, cirugía general, medicina interna y broncopulmonar.
- **Especialidad de destino**: Las especialidades de destino presentes en la base de datos proporcionada corresponden a tres, oncología médica, broncopulmonar y cirugía de tórax.
- **Sospecha de diagnóstico**: Este atributo corresponde a texto libre escrito por el especialista de origen, el cual sintetiza el por qué de la derivación, este campo tiene información valiosa para identificar sospechas de cáncer de pulmón

Estos tres campos descritos se utilizan en la creación de una etiqueta de sospecha de cáncer de pulmón, cuyo método se describe en la sección 3.2.2.

3.2.1.4. Biopsias

Una biopsia corresponde se define como "*extracción de células o tejidos para ser examinados por un patólogo*"(Instituto Nacional del Cáncer de Estados Unidos, s.f.). Este examen sirve para estudiar el tejido extraído para confirmar o descartar algún tipo de cáncer, o derechamente extraer la masa en su totalidad.

El equipo médico del hospital, asegura que se puede considerar como caso sospechoso de cáncer de pulmón, un paciente que se realizó un TAC de tórax, y en una ventana de tiempo se realizó una **biopsia de cualquier tipo**, no necesariamente de tejido pulmonar, pues esto significaría que se puede estar estudiando la diseminación del cáncer o metástasis (localizaciones secundarias).

3.2.2. Construcción de la etiqueta

La construcción de la etiqueta u *output* (y), se realizó con la validación del equipo médico perteneciente al proyecto. Esta utiliza las fuentes de datos de la sección 3.1.3, se define como:

$$y = \begin{cases} 1, & \text{si existe una } \textbf{sospecha de cáncer} \\ 0, & \text{en caso contrario} \end{cases} \tag{3.1}$$

Se utiliza la notación de **caso positivo** en los casos $y = 1$ y **caso negativo** para $y = 0$.

3.2.2.1. Etiqueta para el estudio de cáncer pulmonar

Para la realización de los experimentos, se utilizaron dos tipos de etiquetas (y). En el caso de predicción de cáncer pulmonar, se utilizará la etiqueta descrita en el cuál el equipo médico compuesto por oncólogos, médicos broncopulmonar y enfermeras gestoras, etiquetaron

manualmente. Es decir, bajo el criterio unificado de los integrantes a través de la revisión de los reportes, se etiqueta como **caso positivo** si este **posee una sospecha de cáncer pulmonar**.

Puede ocurrir el caso en el que, el cáncer pulmonar corresponde a una localización secundaria producto de otro tipo de cáncer, este registro se etiqueta como caso negativo. Lo mismo sucede con los TAC de tórax que presentan cáncer de mama, pues estarían interfiriendo con la predicción del modelo de sospecha de cáncer específico para pulmón.

3.2.2.2. Etiqueta para el estudio de cualquier tipo de cáncer

A modo de utilizar todas las fuentes de datos a disposición, se crea una etiqueta en base a una serie de reglas descritas a continuación. Se etiqueta como **caso positivo** cualquier TAC que en un período de **0 a 90 días** luego de su realización, incurrió en uno o más de los siguientes tratamientos:

1. **Ingreso a la plataforma *Cotalker***: como ya fue mencionado, este tipo de paciente corresponde al mejor ejemplo posible, debido a estos casos fueron ya revisados manualmente por la enfermera gestora o el médico cirujano de pulmón, que efectivamente representan una sospecha de cáncer pulmonar.

2. **Derivaciones por interconsulta**: para considar como sospechosa una derivación, se utilizan dos criterios. Primero, utilizando el origen-destino de las interconsultas, se considera como positivo si fue derivado desde la especialidad de Oncología a Cirugía de Tórax o Broncopulmonar. Segundo, para cualquier tipo de especialidad origen-destino, se realiza una búsqueda en la sospecha de diagnóstico, con palabras claves identificadas por el equipo médico (presentes en el Anexo B), considerándose como caso positivo la presencia de alguna de estas *keywords* en el texto.

3. **Realización de una Biopsia**: la realización de una biopsia, ya sea de chequeo de tejido o células, tanto como para extirpar cualquier tipo de masa, da indicios de sospecha positiva en el TAC de tórax según el equipo médico.

La ventana de 90 días fue definida por el equipo médico, a modo de asegurar que los tratamientos sean relevantes al TAC analizado en cuestión.

3.2.3. Análisis de fuentes de datos

El análisis de datos fue realizado utilizando los los *dataset* de TACs, ingresos *Cotalker*, interconsultas y biopsias cuya cantidad de registros se reporta en la tabla 3.1. Se debe destacar que estos registros se encuentran anonimizados bajo un ID único, con el fin de realizar un cruce para la contrucción de la etiqueta de cancer en general, sin embargo algunos pacientes presentan más de un examen del mismo tipo y en cuyo caso, se decidió mantener el primer examen realizado en orden temporal, eliminando los registros con ID duplicado.

Tabla 3.1: Cantidad de registros por fuente y duplicidad

Fuente de datos	**Cantidad total**	**Total sin duplicados**
TACs	12.260	9181
Biopsias	8520	7339
Interconsultas	1755	1366
Cotalker	197	195

En la tabla 3.2, se puede apreciar el cruce realizado, recordando que el filtro del período corresponde a los 90 días **después** de la realización del TAC.:

Tabla 3.2: Cantidad de registros cruzados por fuente de datos, según filtro

Fuente de datos	**Sin filtro de período**	**Con filtro de período**
Biopsias	3.280	940
Interconsultas	835	303
Cotalker	179	72

La tabla 3.2 da pié para el análisis de la diferencia entre la fecha de realización del TC. de Tórax, y la fecha en la que se identifica la fuente sospechosa. La figura 3.5 muestra dicha distribución, tomando la mayor diferencia entre la fecha del evento y el TAC. La figura da cuenta de la gran cantidad de observaciones con diferencia de días negativas, esto puede justificarse a través de TC. de evolución o seguimiento del cáncer. Además, se observa que existe un pequeño volumen de observaciones con fecha positiva, pero que superan el período de 90 días para considerar el TAC como sospechoso de cáncer pulmonar.

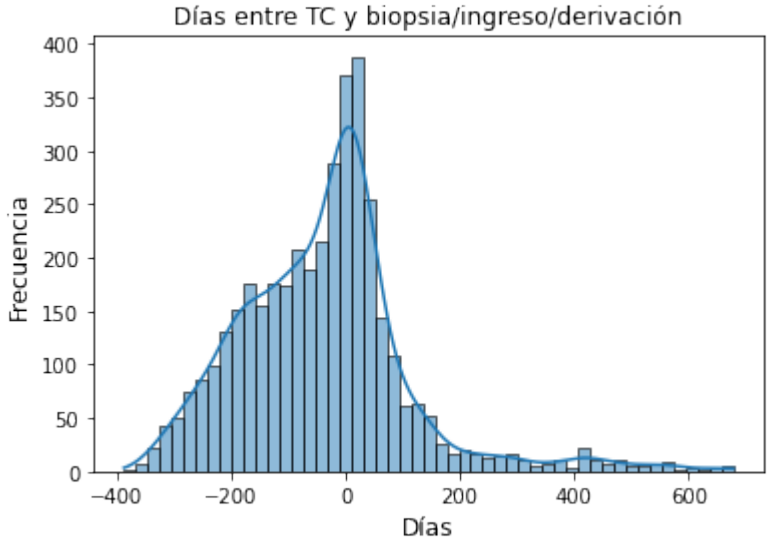

Figura 3.5: Histograma diferencia de días entre TC. y Sospecha

3.2.3.1. Demanda de exámenes TAC. de Tórax

En la figura 3.6, se aprecia la demanda diaria de TC. de Tórax, con un promedio 37 exámenes por día. Como se puede apreciar, es de gran volumen la cantidad de exámenes que tendría que revisar el personal médico para poder evaluar una sospecha de cáncer pulmonar, con lo cual se reitera la importancia del proyecto.

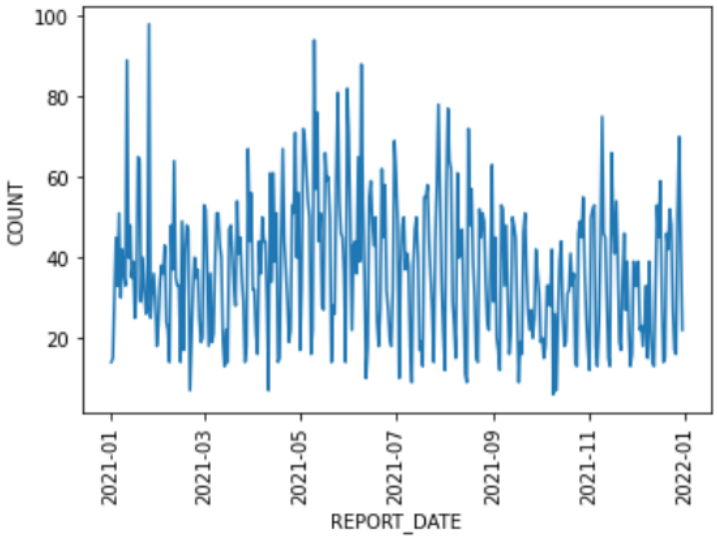

Figura 3.6: Serie de tiempo Tomografías realizadas en el Hospital

3.2.3.2. Edad de los pacientes

Resulta de interés, estudiar la edad de los pacientes frente a una sospecha de cáncer pulmonar. Según la Sociedad Americana del Cáncer, la mayoría de los casos de cáncer pulmonar son para pacientes con 65 años o más; un pequeño número de personas diagnosticadas son menores a 45. La media de la variable edad de los casos con cáncer de pulmón en Estados Unidos para el 2023 es de 70 años aproximadamente.

Analizando la figura 3.7, no se aprecian grandes diferencias en la distribución de la edad para ambas categorías, donde ambas tienen un promedio de 60 años, con una desviación estándar de 16 años para la clase negativa y 15 para la positiva.

3.2.3.3. Reporte de TACs de Tórax

La variable de texto libre de reportes de TC. de tórax, es el principal insumo para generar los vectores de atributos para entrenar el algoritmo de clasificación de sospecha de cáncer pulmonar. Para realizar un análisis de la variable de texto, se separa este en sus partes de antecedentes, hallazgos e impresiones, y se estudian los *n-gramas* (conjuntos de palabras consecutivas en un documento) más comunes:

La separación de las partes del relato se justifica debido a que las palabras tendrán distinta importancia según si aparecen es antecedentes, o impresiones, a modo de ejemplo, es distinto

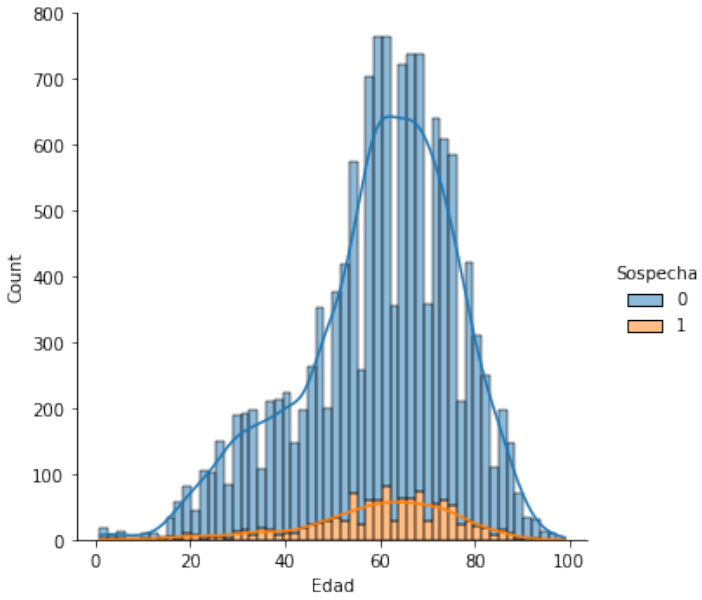

Figura 3.7: Histograma edades según tipo de Sospecha

encontrar la palabra nódulos en la primera parte del relato a encontrarlo en impresiones, pues si se menciona en antecedentes puede ser debido a un seguimiento, mientras que si aparece en impresiones se puede referir a que fue encontrado en el examen. Por su lado, en hallazgo se suelen repetir muchas palabras que describen el tamaño de los órganos y otras observaciones de la imagen.

Como se puede apreciar en la figura 3.8, en antecedentes, las frases más frecuentes muestran el estudio de nódulos pulmonares, escritos de diversas formas. Mientras que en hallazgos, se repiten las medidas que tiene los órganos de la imagen, independiente a si se encuentra una anomalía o no. Finalmente, las impresiones son las más variadas, detallando una serie de enfermedades que los radiólogos identifican en las imágenes.

3.2.3.4. Distribución de la etiqueta

Se estudia por separado la distribución de la etiqueta de sospecha de cáncer de pulmón y sospecha de cáncer:

1. **Etiqueta de cáncer de pulmón:** La distribución de la etiqueta real, correspondiente a la revisión de TACs y su clasificación mediante la decisión del equipo médico se presenta en la figura 3.9. De un total de 2418 casos revisados, 235 (9.7 %) corresponden a una sospecha positiva de cáncer de pulmón

2. **Etiqueta de cáncer:** La distribución de la etiqueta de cáncer construida en base a los eventos sospechosos, realizados en la ventana posterior a la realización del TAC se presenta en la figura 3.10. Del total de 9210 TACs, 1320 (14 %) corresponden a una sospecha positiva de cáncer de pulmón

Figura 3.8: Nube de palabras, *tri-gramas* según parte del relato

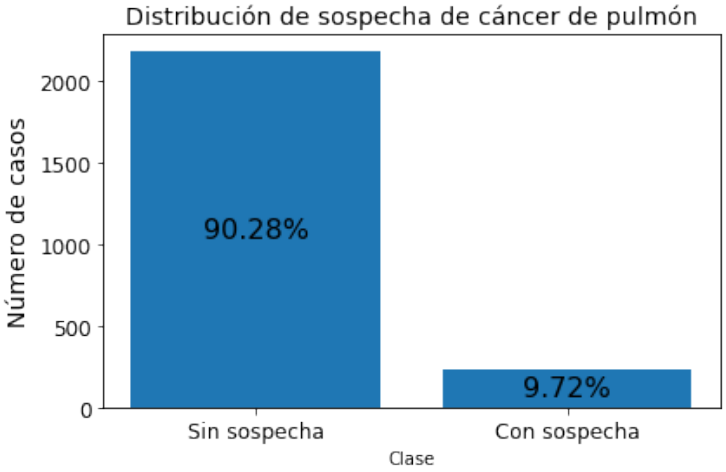

Figura 3.9: Distribución etiqueta de cáncer de pulmón

Si bien, para cualquier tipo de cáncer existe un mayor volumen de datos etiquetados, en ambos casos se puede apreciar un problema de desbalance de clases inherente al contexto.

Al estudiar la creación de la etiqueta, la figura 3.11 muestra el problema ya evidenciado en la sección 3.2.3. Existen una cantidad no menor de registros que se consideran como no sospechosos o casos negativos, que se debe al periodo de 90 días (positivos) definido. Por otro lado, se reitera la proporción en la que se identifican los casos sopechosos de la tabla 3.2, siendo las biopsias el grueso de las fuentes sospechosas, seguidas por las interconsultas y finalmente los ingresos por *Cotalker*. Finalmente, se debe mencionar que la suma de las

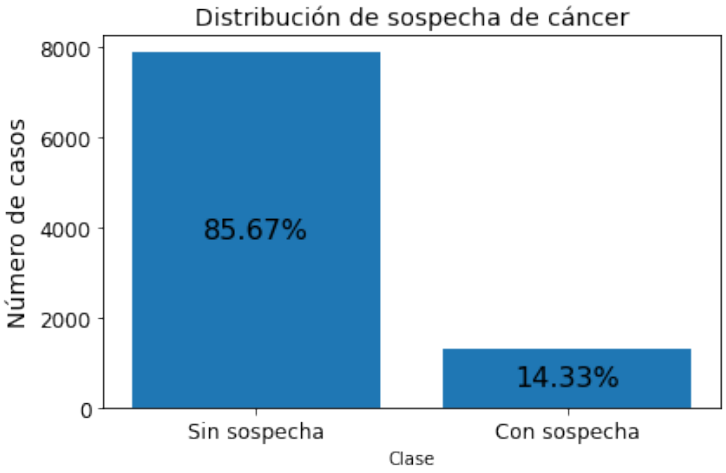

Figura 3.10: Distribución etiqueta de cáncer

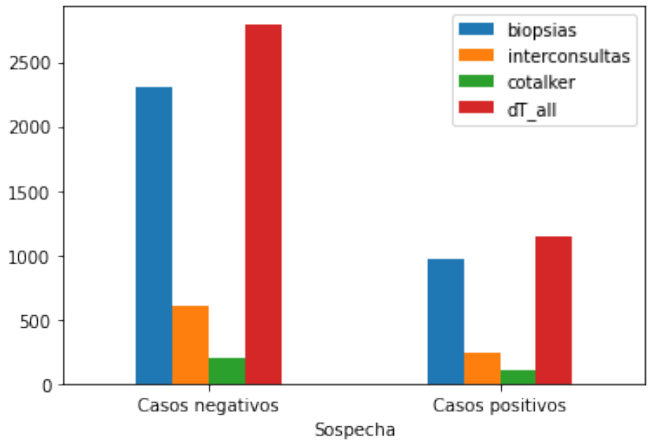

Figura 3.11: Etiqueta según fuente de origen y sospecha

barras del número de biopsias, interconsultas e ingresos por *Cotalker* de la figura 3.11, no coincide con el tamaño de la barra roja, debido a existen casos en los que el paciente posee más de una fuente sospechosa.

3.3. Preparación de los datos

La fase inicial del procesamiento de datos implica la eliminación de IDs duplicados de pacientes, conservando el último registro temporal. Luego, se cruza la información entre fuentes según los identificadores de los pacientes. En cuanto al texto clínico, se realiza un preprocesamiento que implica la eliminación de números y caracteres especiales, estandarización del texto a minúsculas y eliminación de tildes. Esta elección se basa en criterios como normalización del texto para reducir variabilidad, simplificación del análisis, reducción de ruido, uniformidad de tokens y eficiencia computacional. Para la construcción de modelos de clasificación, resulta beneficioso estructurar los informes en antecedentes, hallazgos e impresiones.

Esto posibilita un contexto clínico completo, la selección de características relevantes, la interpretación efectiva de hallazgos y la obtención de un resumen para la toma de decisiones clínicas. En el ámbito del aprendizaje automático, se realiza una partición del dataset en conjuntos de entrenamiento (80 %) y validación (20 %). Esto facilita la evaluación imparcial del modelo, previene el sobreajuste, mide el rendimiento real, permite la selección y ajuste de hiperparámetros, y facilita la comparación justa de modelos. La utilización de conjuntos de prueba separados es esencial para la evaluación objetiva y efectiva del rendimiento del modelo en situaciones del mundo real.

Con la finalidad de mejorar el rendimiento de los modelos, de acuerdo con la literatura, se realiza un balanceo de clases primero realizando un *undersampling* aleatorio sobre el conjunto de entrenamiento y la clase mayoritaria, posteriormente, realiza un *oversampling* utilizando la técnica *Synthetic Minority Oversampling Technique* (SMOTE), para generar más observaciones de la clase minoritaria. Se debe recalcar que el conjunto de testeo no es alterado para la generalización del modelo y la obtención de métricas realistar de su rendimiento.

Para modelos de Redes Neuronales, la preparación de los datos se resume en la utilización de *word embeddings*, para transformar el texto en vectores densos que captan el contexto de las palabras dentro de los documentos.

Para modelos clásicos de clasificación como lo es *Random Forest*, se utiliza *Text Vectorization*, en su versión *Term Frequency-Inverse Document Frequency* (TF-IDF), que cuantifica la relevancia de las palabras del documento.

Estas representaciones del texto, en conjunto con sus etiquetas correspondientes son necesarias para el modelamiento de los distintos experimentos.

3.4. Modelamiento

El modelamiento realizado considera las técnicas abordadas en el Capitulo 2, el *pipeline* de los modelos comienza una vez se tienen los datos con su etiqueta correspondiente, y el texto clínico ya preprocesado (*Dataset*, en la figura 3.12). Posteriormente, se separan los conjuntos de entrenamiento y testeo en razón 8:2.

Como el contexto del problema de clasificación presenta clases desbalanceadas, se aplican las estrategias de **reequilibrio de clases** para mejorar el rendimiento de los modelos. Esto se realiza únicamente en el conjunto de entrenamiento y se utilizan las siguientes técnicas:

1. ***Random Undersampling*:** Esta técnica reduce el tamaño de la clase mayoritaria (casos sin sospecha de cáncer de pulmón), seleccionando aleatoriamente ejemplos de esa clase hasta que su número sea similar al de la clase minoritaria. Su finalidad es disminuir la influencia de la clase mayoritaria en el modelo y permitir que el modelo se concentre más en aprender la clase minoritaria (casos con sospecha de cáncer de pulmón)

2. ***Random Oversampling*:** Esta técnica aumenta el número de ejemplos de la clase minoritaria seleccionando aleatoriamente muestras de esta clase y agregándolas al conjunto

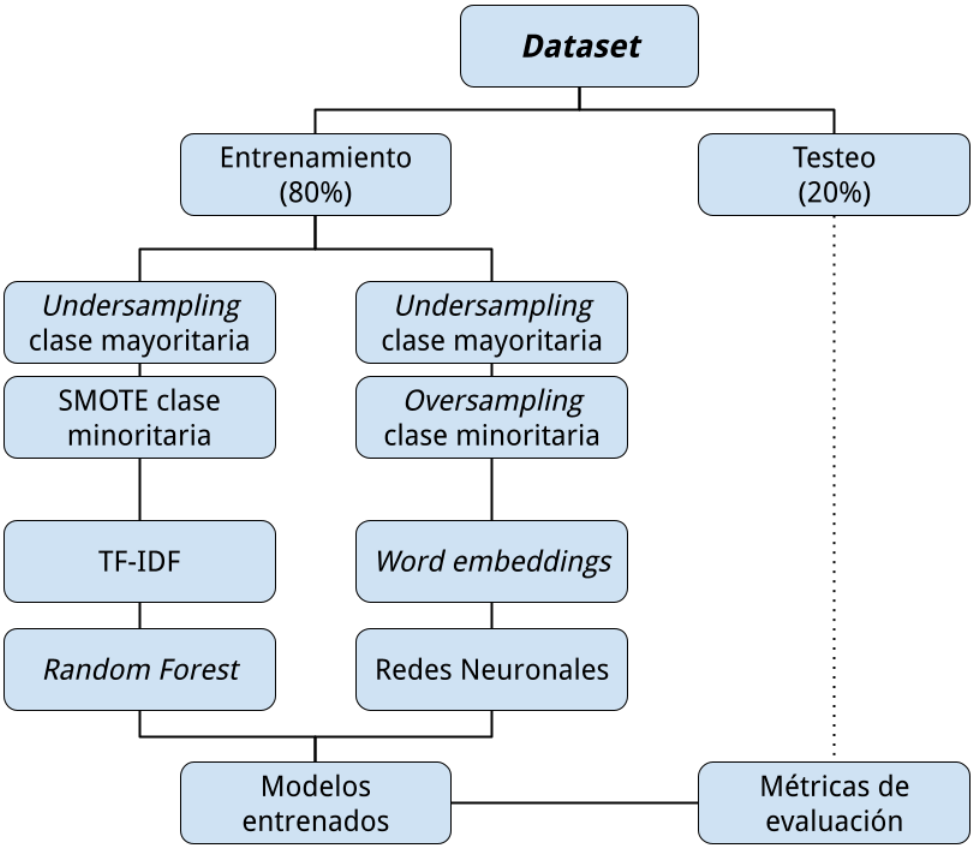

Figura 3.12: Diagrama de los modelos realizados

de datos original. Esto equilibra la proporción entre las clases, lo que puede mejorar la capacidad del modelo para aprender de manera equitativa de ambas clases. Sin embargo, el *Random Oversampling* también puede llevar a un aumento en el riesgo de sobreajuste, ya que introduce datos duplicados en la clase minoritaria.

3. **SMOTE**: SMOTE, o *Synthetic Minority Over-sampling Technique* (Técnica de Sobremuestreo Sintético de la Clase Minoritaria, en español), crea muestras sintéticas de la clase minoritaria mediante la interpolación de ejemplos cercanos en el espacio de características. Funciona seleccionando un ejemplo de la clase minoritaria y generando ejemplos adicionales al crear combinaciones lineales entre ese ejemplo y sus vecinos más cercanos. Estas muestras sintéticas ayudan a equilibrar la proporción de clases en el conjunto de datos, lo que puede mejorar la capacidad del modelo para aprender de manera equitativa de ambas clases. SMOTE es una técnica valiosa cuando se trata de abordar desbalances de clases. Como SMOTE requiere de una representación de vectorial del texto para la interpolación, sólo se utiliza esta técnica previo a los modelos de *Random Forest* debido a la vectorización TF-IDF.

En el conjunto de entrenamiento se calibran los distintos modelos y arquitecturas tanto

de árboles de decisión como redes neuronales:

1. **Árboles de decisión (*Random Forest*)**: Para el primer caso, el modelamiento considera el balanceo de clases ya descrito, así como el cálculo de la matriz *tf-idf*, para posteriormente encontrar los mejores parámetros para *Random Forest*, los parámetros calibrados corresponden a:

   a) *n_stimators*: Número de árboles de decisión que se construirán en el bosque. Un valor más alto generalmente mejora el rendimiento del modelo, pero también aumenta el tiempo de entrenamiento y la complejidad.

   b) *max_features:* Determina el número máximo de características que se utilizarán para dividir un nodo en cada árbol

   c) *max_depth:* Define la profundidad máxima de cada árbol en el bosque. Limita la cantidad de divisiones que puede hacer cada árbol, se debe considerar que un valor alto puede llevar al sobreajuste del modelo.

   d) *bootstrap:* Controla si se utiliza el muestreo con reemplazo (*bootstrapping*) al construir árboles. *Bootstrapping* implica muestrear aleatoriamente con reemplazo del conjunto de entrenamiento para construir cada árbol.

   Para obtener la mejor combinación de parametros, se utiliza la técnica de búsqueda de hiperparámetros (*RandomizedSearchCV*) de validación cruzada, eligiendo el mejor modelo en base a la métrica especificada. Finalmente, dicho modelo entrenado se utiliza para la clasificación del conjunto de testeo, para calcular su desempeño en base a las métricas de evaluación.

2. **Redes Neuronales**: se utilizan modelos del estado del arte en NLP revisados en la literatura, en particular, a través de la librería *Autokeras* que permite realizar una *Búsqueda de Arquitectura Neuronal* (NAS, por sus siglas en inglés), mediante optimización bayesiana, Jin et al. (2019). Dentro de los modelos probados, se destacan Redes Neuronales Convolucionales, Redes Neuronales Recurrentes y *BERT transformer*. Los parámetros del modelo se describen a continuación:

   a) *max_trials*: Este parámetro controla el número máximo de modelos (arquitecturas de NN) diferentes que se probarán durante la búsqueda de hiperparámetros. Cuanto mayor sea el número, más modelos se evaluarán.

   b) *metrics*: Define la función objetivo que se utilizará para optimizar el modelo. En este caso, se fija *Recall* como métrica.

   Al igual que para los modelos de árboles de decisión, una vez se obtienen el mejor modelo en el conjunto de entrenamiento, se prueba su rendimiento en el conjunto de testeo, midiendo su desempeño en base a las métricas de evaluación.

Capítulo 4

Resultados

4.1. Modelos de cáncer de pulmón

Para el *dataset* de casos revisados por el equipo médico, que entrena un modelo analítico para la sospecha de cáncer de pulmón, se listan los modelos utilizados (tabla 4.1) así como el tamaño de sus conjuntos de entrenamiento y testeo. Se debe mencionar que el tamaño del entrenamiento varía entre experimentos debido al balanceo de clases realizado.

Tabla 4.1: Tamaño conjuntos entrenamiento y testeo

Dataset	**Modelo**	**Entrenamiento**	**Testeo**
C. pulmón	TF-IDF *Random Forest*	1934	484
C. pulmón *balanced*	TF-IDF y *Random Forest*	728	484
C. pulmón *balanced*	TF-IDF, *TF* y Separación Relato	1456	484
C. pulmón	Redes Neuronales	1934	484
C. pulmón *balanced*	Redes Neuronales	1743	484

Los resultados de la tabla 4.2 evidencian que el mejor modelo corresponde a *Random Forest* junto a la representación *tf-idf*, balanceo de clases y separación del relato (destacado en negrita). Este modelo obtuvo el mejor *Recall* para los casos con sospecha de cáncer de pulmón. Los modelos 2 y 4 en la tabla si bien presentan un alto valor para el *Recall* de casos positivos, son descartados debido a que realizan una predicción unidireccional. Otro factor importante a la hora de elegir el mejor modelo corresponde al tiempo ejecución, existiendo una diferencia sustancial entre los modelos clásicos utilizando *tf-idf* y *Random Forest,* en comparación con modelos más complejos como lo es *BERT*. El primero demora del orden de minutos, mientras que el entrenamiento de un modelo *BERT* puede tardar horas, sin significar una mejora significativa. Se debe tener en cuenta también la escalabilidad de modelo elegido, pues ante un eventual despliegue un modelo más complejo necesitará de un mejor *hardware* y recursos.

Tabla 4.2: Resultados modelos cáncer de pulmón

Modelo	Caso	Precision	Recall	F1-score	Soporte
Random Forest	Sin sospecha	0.99	0.99	0.96	431
	Con sospecha	0.82	0.53	0.64	53
Random Forest	Sin sospecha	0.99	0.55	0.71	431
y balanceo	Con sospecha	0.21	0.96	0.34	53
***Random Forest*,**	Sin sospecha	0.98	0.77	0.86	431
balanceo y sep.[a]	**Con sospecha**	0.32	**0.89**	0.47	53
Red Convolucional	Sin sospecha	0.50	0.01	0.02	440
y balanceo	Con sospecha	0.08	0.91	0.15	44
BERT	Sin sospecha	0.97	0.55	0.70	440
y balanceo	Con sospecha	0.15	0.82	0.26	44

[a] separación del relato

4.1.1. Análisis de Sensibilidad

El valor de AUC (Área bajo la Curva ROC) de 0.91 presentado por el mejor modelo (figura 4.1), sugiere que el modelo tiene una sólida capacidad discriminativa entre casos positivos y negativos. Además, un valor de Recall de 0.89 para la clase con sospecha es un indicativo de que el modelo puede identificar correctamente el 89 % de los casos positivos (sospecha de cáncer de pulmón). Este es un valor bastante alto y es fundamental en un contexto médico, ya que minimiza la probabilidad de falsos negativos, es decir, de no identificar a pacientes con sospecha de cáncer de pulmón.

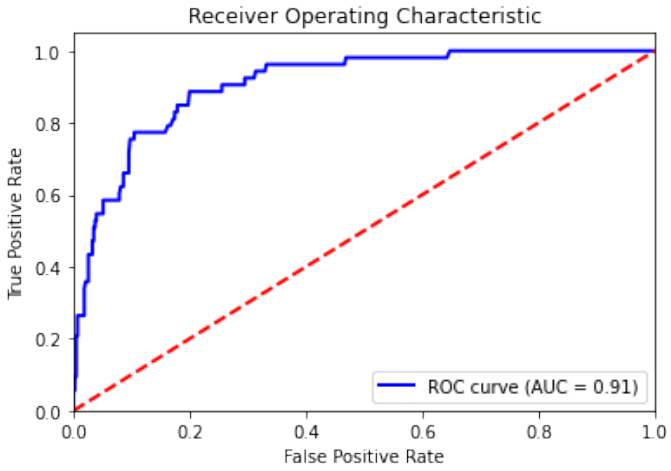

Figura 4.1: Curva ROC

La Curva de Precisión-Recall ayuda a seleccionar un punto de operación óptimo según las necesidades clínicas y la tolerancia a los errores de predicción. Dado un umbral de probabilidad (valor que define cuan segura debe ser una predicción para considerarse como positiva),

un umbral más bajo puede ser preferible si se desea maximizar el *Recall* para no perder casos de cáncer, a pesar de que haya más falsos positivos. Por otro lado, un umbral más alto puede ser preferible si se busca una alta precisión y estás dispuesto a tolerar menos falsos positivos, aunque eso signifique perder algunos casos de cáncer.

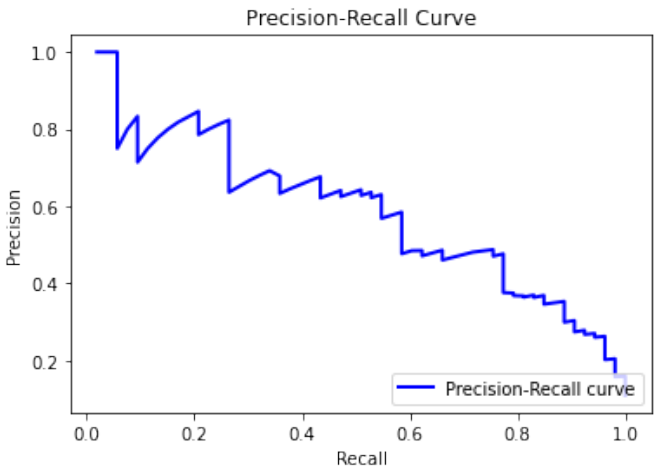

Figura 4.2: Curva Precisión-Recall

El umbral de probabilidades para el mejor modelo fue fijado en 0.3, al estudiar la sensibilidad de la curva Presición-Recall, obteniendo una Precisión del 0.32 y un Recall del 0.89. Por otro lado, el valor que maximiza el *Recall* para este modelo, corresponde a un umbral de 0.11 obteniendo una Precisión del 0.16 y un Recall del 100 %.

4.2. Modelos de cáncer en general

Para el *dataset* construido utilizando las múltiples fuentes de datos a disposición, que tiene como etiqueta positiva a un paciente sospechoso de cualquier tipo de cáncer. Se presentan los resultados en la tabla 4.3, se puede apreciar que el rendimiento de los modelos disminuyen considerablemente, obteniendo los mejores resultados para el modelo de *Random Forest* que incorpora el *resampling* de clases, la edad de los pacientes y la separación del relato.

El mejor modelo obtiene un *Recall* del 0.63, para la clase positiva. Es decir, de cada 3 pacientes con una posible sospecha de cáncer, el modelo identifica 2 correctamente. Por otro lado presenta una Precisión de 0.25.

Como se puede apreciar, el modelo que presenta el mayor *recall* para la clase con sospecha corresponde a la arquitectura *BERT* con un 40 %. El desempeño de este modelo no presenta un pronóstico acertado ni confiable para los casos de sospecha de cáncer pulmonar, entre las posibles explicaciones, se menciona el evidente desbalance de clases, la cantidad de datos para el entrenamiento de los modelos de redes neuronales y la construcción de la etiqueta como un factor importante para el bajo desempeño.

Tabla 4.3: Resultados clasificación *Random Forest*

Modelo	**Caso**	**Precision**	**Recall**	**F1-score**	**Soporte**
Random Forest	Sin sospecha	0.94	0.99	0.96	2490
	Con sospecha	0.74	0.38	0.50	237
Random Forest	Sin sospecha	0.96	0.83	0.89	2490
y balanceo	Con sospecha	0.26	0.61	0.36	237
Random Forest,	Sin sospecha	0.96	0.83	0.89	2490
balanceo y edad	Con sospecha	0.26	0.62	0.36	237
Random Forest,	Sin sospecha	0.96	0.82	0.88	2490
balanceo, edad y sep.[a]	**Con sospecha**	0.25	**0.63**	0.36	237
Convolucional	Sin sospecha	0.94	0.96	0.95	2495
	Con sospecha	0.48	0.36	0.41	232
BERT transformer	Sin sospecha	0.94	0.95	0.95	2495
	Con sospecha	0.44	0.40	0.42	232

[a] separación del relato

Capítulo 5

Discusión

Hasta la fecha, no esta existe una propuesta de sistema de detección temprana de cáncer de pulmón por el Plan Nacional de Cáncer 2018-2028. A diferencia de las pesquisas realizadas por el Ministerio de Salud para el cáncer de mama y cervicouterino. Por lo tanto los pacientes ingresan a los sistemas de salud del país después de la presencia de síntomas

La solución propuesta permite procesar grandes volúmenes de texto, automatizando la tarea de evaluación de TACs de Tórax para confirmar o descartar una sospecha de cáncer de pulmón. Se debe mencionar que la herramienta no remplaza el trabajo realizado por el equipo médico, sino debe ser interpretado como una ayuda para este, que levanta una alerta de casos sospechosos para su posterior evaluación y eventual ingreso en el modelo de trazabilidad de cáncer de pulmón que posee HSR.

El proyecto logró un alto rendimiento en la tarea de clasificación binaria, identificando pacientes con sospecha de cáncer de pulmón. Sin embargo, este éxito conlleva importantes consideraciones y desafíos. Se demostró el potencial del NLP y el aprendizaje automático en la detección de cáncer de pulmón a partir de informes de TC. Se lograron resultados prometedores, pero es esencial considerar las implicaciones clínicas y abordar las limitaciones para una implementación exitosa en entornos médicos reales. El trabajo de titulo podría aportar al desarrollo de políticas públicas asociadas al cáncer de pulmón, complementando el trabajo ya realizado en áreas de prevensión de esta enfermedad, como por ejemplo los programas de tabaquismo.

Esta enfermedad tiene un alto impacto social y económico, tanto para las personas, establecimientos de salud y el estado. Para las personas, se espera una mejoría en la calidad de vida, mediante la anticipación a un cáncer pulmonar, previo a la presencia de síntomas. Para los establecimientos de salud, la priorización de pacientes y un tamizaje eficiente permiten enfocar los recursos, bajando la demanda de atenciones y costos operativos. Finalmente para el estado, el detectar la enfermedad en una etapa temprana, disminuye los costos de medicamentos para el tratamiento de la enfermedad en una etapa crónica.

En cuanto al modelo de cáncer en general, los resultados obtenidos son un buen acercamiento a modelos de predicción de otros tipos de cáncer, a pesar del bajo rendimiento en

un contexto clínico. Extender esta herramienta de detección de cáncer de pulmón a otros tipos de cáncer implica varios desafíos, pero es factible con ciertos ajustes y consideraciones utilizando la metodología del proyecto. Se debe primero recolectar datos clínicos, radiológicos y de otros tipos para el cáncer específico a detectar, estos datos pueden incluir informes médicos, imágenes médicas, historias clínicas electrónicas y otros registros médicos.

Cada tipo de cáncer puede tener características específicas en los datos que son relevantes para la detección, por lo que será necesario el desarrollo de nuevas *features*. Además, se debe adaptar los modelos de aprendizaje automático y algoritmos para la detección de nuevos tipos de cáncer, lo que significa ajustar hiperparámetros, seleccionar características relevantes y entrenar modelos específicos. La colaboración con expertos en oncología y otros profesionales médicos es esencial para el éxito de este tipo de proyectos. En la literatura, el Machine Learning destaca por ser altamente efectivo en la predicción de varios tipos de cáncer, incluyendo el cáncer de mama, cerebral, de pulmón, hígado y prostata.

Capítulo 6

Conclusiones

El presente trabajo de titulo se centra en el desarrollo de una herramienta de detección de cáncer de pulmón utilizando reportes de tomografías computarizadas de tórax (TC), tomadas en el Hospital Sótero del Río de la región Metropolitana de Chile. El uso de modelos de aprendizaje automático, como *Random Forest* y Redes Neuronales, condujo a la construcción de clasificadores precisos.

El mejor modelo predictivo de cáncer de pulmón posee una alta sensibilidad, y predice correctamente el 89 % de los casos positivos. Los datos utilizados corresponden a TC. de tórax etiquetados por Oncólogos y equipo médico de la Red de Salud. Al utilizar biopsias y otras fuentes de datos, el desempeño disminuye con un Recall del 63 %.

Se espera que los modelos presentados puedan ser utilizados en un futuro para el apoyo en la detección temprana de cáncer de pulmón, en conjunto con las estrategias de seguimiento y trazabilidad existentes dentro del hospital, mejorando la esperanza de vida de los pacientes del país.

6.1. Trabajo futuro

El trabajo futuro con mayor relevancia corresponde a la implementación de la herramienta en el Hospital Sótero del Río, que escapa de los alcances del trabajo de titulo. Para ello se debe trabajar de forma integrada con tanto con HSR, como la red de salud SSMSO.

La continua revisión y entrenamiento de los modelos resultará clave, trabajando constantemente por la mejora de los modelos de predicción utilizados, probando con distintas arquitecturas, clasificadores, técnicas de muestreo para texto. No se descarta el uso de otros insumos, cómo lo serían imágenes o fichas clínicas.

Bibliografía

[1] Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2020). *Global Cancer Observatory: Cancer Today*. Lyon, France: International Agency for Research on Cancer. [Disponible en línea: https://gco.iarc.fr/today]

[2] *PLAN NACIONAL DE CÁNCER 2018 – 2028* (2019). Ministerio de Salud. [Disponible en línea: PLAN NACIONAL DE CÁNCER 2018 - 2028]

[3] Parra-Soto, Solange, Petermann-Rocha, Fanny, Martínez-Sanguinetti, María Adela, Leiva-Ordeñez, Ana María, Troncoso-Pantoja, Claudia, Ulloa, Natalia, Diaz-Martínez, Ximena, & Celis-Morales, Carlos. (2020). Cancer in Chile and worldwide: *an overview of the current and future epidemiological context. Revista médica de Chile, 148*(10), 1489-1495. [DOI: 10.4067/S0034-98872020001001489]

[4] *Chilesincáncer*. (s.f). Quienes Somos [Disponible en línea: https://chilesincancer.cl]

[5] Yim W, Yetisgen M, Harris WP, Kwan SW. (2016) Natural Language Processing in Oncology: *A Review. JAMA Oncol.* ;2(6):797–804. [DOI: 10.1001/jamaoncol.2016.0213]

[6] Rogers, Anna; Kovaleva, Olga; Rumshisky, Anna (2020)."A Primer in BERTology: *What We Know About How BERT Works". Transactions of the Association for Computational Linguistics*. 8: 842–866. [arXiv:2002.12327]

[7] Solarte-Pabón, O., Montenegro, O., Blazquez-Herranz, A., Saputro, H., Rodriguez-González, A., & Menasalvas, E. (2021). Information extraction from Spanish radiology reports using multilingual BERT. *CLEF eHealth.* [Disponible en: http://ceur-ws.org/Vol-2936/paper-69.pdf]

[8] Montazeri, M., Afraz, A., Mahboob Farimani, R., & Ghasemian, F. (2021). Natural Language Processing Systems for Diagnosing and Determining Level of Lung Cancer: *A Systematic Review. Frontiers in Health Informatics, 10*(1), 68. [DOI: 10.30699/fhi.v10i1.264]

[9] Bitterman D., Miller T., Mak R., Savova G. (2021). Clinical Natural Language Processing for Radiation Oncology: *A Review and Practical Primer*. [DOI: 10.1016/j.ijrobp.2021.01.044]

[10] Goldberg, Y. (2016). *Neural Network Methods for Natural Languaje Processing.*

[11] Jurafsky. D. and Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*

[12] Hochreiter, S. and Schmidhuber, J (1997). *Long short-term memory.*

[13] *Cancer staging.* (s.f). National Cancer Institute.

[14] American Cancer Society. (2023). *Key Statistics for Lung Cancer*

[15] Jin, H., Song, Q., & Hu, X. (2019). Auto-keras: *An efficient neural architecture search system.* [Disponible en línea: https://arxiv.org/abs/1806.10282]
American Cancer Society. (s.f.). *Treating Early-Stage Non-Small Cell Lung Cancer.* [Disponible en línea: https://www.cancer.org/cancer/lung-cancer/treating-early.html]

Anexos

Anexo A. Tomografía Computarizada de Tórax

TOMOGRAFIA COMPUTADA DE TÓRAX del 01-01-2021:

Antecedentes clínicos: Disnea. Insuficiencia renal aguda.

Hallazgos:

Examen realizado sin uso contraste endovenoso.
Leve a moderado derrame pleural derecho y leve derrame pleural izquierdo.
Atelectasias pasivas en ambos lóbulos inferiores.
Pulmones de arquitectura normal.
Engrosamiento intersticial fino bilateral mayor hacia ambas bases.
Tenues opacidades con densidad en "vidrio esmerilado" en porciones dependientes de lóbulos superiores y en menor medida en lóbulos inferiores.
Moderada cardiomegalia con crecimiento predominante de cavidades izquierdas. No hay derrame pericárdico.
Ateromatosis cálcica aórtica y coronaria.
Pequeñas adenopatías mediastínicas de aspecto inespecífico, la mayor de ellas de 1,1 x 1,2 cm en situación paratraqueal derecha.
Hígado de tamaño y morfología normal.
Vesícula biliar, vía biliar, páncreas, bazo y glándulas suprarrenales de caracteres tomográficos normales en este estudio no contrastado.
Ambos riñones de tamaño normal. El riñón derecho mide 10,7 cm y el izquierdo 12,3 cm. No hay hidroureteronefrosis.
Vejiga en repleción, de paredes finas. Fosas isquiorrectales libres.
No hay líquido libre intraabdominal. Próstata y vesículas seminales sin alteraciones evidentes. Asas de intestino delgado y grueso de calibre normal.
No hay adenopatías mesentéricas ni retroperitoneales.

Impresión:

Moderada cardiomegalia con crecimiento predominante de cavidades izquierdas.
Leve a moderado derrame pleural derecho y leve derrame pleural izquierdo.

Figura A.1: Ejemplo TAC. de Tórax

Anexo B. Lista de palabras sospechosas

NÓDULO PULMONAR EN ESTUDIO
Tumor de comportamiento incierto o desconocido de la traquea, de los bronquios y del pulmon
Tumor maligno de los bronquios o del pulmón
CÁNCER DE PULMÓN
HALLAZGOS ANORMALES EN DIAGNOSTICO POR IMAGEN DEL PULMON
nodulo pulmonar
neoplasia broncogénica
nódulo cavitado
nódulo pulmonar solitario
nódulo subsólido
metastasis en pulmon
metástasis pulmonares
implantes pleurales
Masa pulmonar
adenopatías hiliares
masa pulmonar
CA PULMON
Tumor maligno de los bronquios y del pulmón, (de los órganos respiratorios e intratoracicos)
TU PULMONAR EN ESTUDIO
Tumor maligno de los bronquios o del pulmón, parte no especificada
METÁSTASIS PULMONAR
Neoplasia maligna
MULTIPLES NODULOS PULMONARES Y MASA ESTRENAL METASTASICO
OBS CA PULMON
CANCER DE PULMON DE CÉLULAS GRANDES
paciente en estudio de lesión pumonar
NÓD PULMONAR
CANCER ALVEOLAR DEL PULMON CANCER DE PULMON

Motivos de consultas: Hemoptisis, Neumonía a repetición, Baja de peso en estudio, Cáncer pulmonar en estudio, Nódulo pulmonar en estudio, Nódulo crecimiento lepídico.

Anexo C. Hiperparámetros del mejor modelo *Random Forest*

Hiperparámetros obtenidos a partir del *Grid Search* realizado:

```
{'n_estimators': 104,
 'max_features': 'sqrt',
 'max_depth': None,
 'bootstrap': True}
```