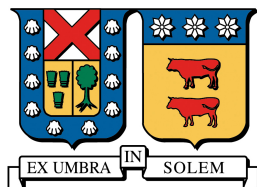


UNIVERSIDAD TÉCNICA FEDERICO SANTA
MARÍA

DEPARTAMENTO DE ELECTRÓNICA Y TELEMÁTICA
VALPARAÍSO – CHILE



PRIORIZACIÓN DE ATENCIÓN EN PACIENTES CON
CÁNCER DE MAMA:
EXTRACCIÓN DE INFORMACIÓN E INTEGRACIÓN DE
MÓDULOS DE PROCESAMIENTO

KATTYA ANETTE HERRERA SALAZAR

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL TELEMÁTICA

PROFESOR GUÍA
WERNER CREIXELL FUENTES

PROFESOR CORREFERENTE
NICOLÁS GÁLVEZ RAMÍREZ

OCTUBRE 2023

A mi Mamá.

Agradecimientos

Agradezco a mi familia, especialmente a mi mamá por ser el pilar fundamental en mi vida, gracias a sus enseñanzas y fortaleza pude completar esta etapa. A mi papá, quien no alcanzó a verme completar esta etapa, pero que sé que me está acompañando. Toda mi familia, quienes siempre me apoyaron y confiaron en mí, soy lo que soy gracias a ustedes y les estoy eternamente agradecida. También a mi Tía quién siempre ha estado presente y es prácticamente mi segunda madre, a mis hermanos y a mis primos.

Agradezco a mis compañeras y compañeros de la universidad, a Luis y Bastián, el *Equipo Asobi* fue inseparable hasta el final y siempre estuvieron para mí. A los participantes de la iniciativa *IRE*, me es imposible nombrarlos a todos pero sepan que estoy agradecida por siempre entregarme buenos momentos y acompañarme en los más difíciles. A todos mis amigos gracias.

Finalmente, agradezco a Nasus por su cariño incondicional, por todos los momentos compartidos y por toda la felicidad que trae a mi vida.

Resumen

El cáncer de mama es el cáncer más frecuente en mujeres y una de las principales causas de mortalidad [1]. Se trata de un problema a nivel mundial y nacional, y se prevé que el número de nuevos casos continuará aumentando en los próximos años. Por lo tanto, es imperativo contar con medidas para detener este crecimiento.

La mortalidad asociada a esta enfermedad puede prevenirse mediante una detección temprana, ya que en estos casos se logra una tasa de curación sobre el 90 % [2]. El examen fundamental para esta detección temprana es la mamografía, considerado el más efectivo en relación a su costo. Sin embargo, tras la obtención de los resultados de este examen, surge un problema: no existe un sistema de priorización en la atención. Por lo general, los pacientes son atendidos siguiendo un sistema de 'orden de llegada'. En este contexto, donde el tiempo juega un papel crucial para los pacientes, esta situación puede perjudicarlos.

Con esta idea en mente, se propone el diseño e implementación de un sistema que tomará como entrada el texto de los informes de mamografías. Esta información se someterá a un proceso de limpieza y estandarización para extraer el BI-RADS, el cual será utilizado para la asignación de prioridad. Sin embargo, surge un desafío cuando no se encuentra información de BI-RADS en todos los informes. En estos casos, será necesario realizar una predicción para determinar su clasificación.

Una vez completadas ambas etapas, los resultados se mostrarán en una plataforma que permitirá a los médicos visualizar el estado de sus pacientes y su historial de exámenes. Esta plataforma funcionará como una valiosa herramienta de apoyo en el campo de la salud, proporcionando una visión general y detallada que facilitará la toma de decisiones.

Este escrito se centrará en el procesamiento de datos y en la integración de los diferentes módulos para permitir que este procesamiento se realice de manera directa desde la plataforma.

Por último, es importante mencionar que este proyecto de memoria se desarrolla en el marco del Programa de Memorias Multidisciplinarias, en colaboración con la Fundación Arturo López Pérez, con el objetivo de abordar el problema previamente planteado.

Abstract

Breast cancer is the most common cancer in women and one of the leading causes of mortality. It is a global and national issue, and it is anticipated that the number of new cases will continue to increase in the coming years. Therefore, it is imperative to have measures in place to halt this growth.

Mortality associated with this disease can be prevented through early detection, as in these cases, a cure rate of over 90 % is achieved. The key examination for this early detection is mammography, considered the most cost-effective. However, a problem arises after obtaining the results of this examination: there is no prioritization system in place for patient care. Typically, patients are attended to on a 'first-come, first-served' basis. In this context, where time plays a crucial role for patients, this situation can be detrimental to them.

With this idea in mind, the design and implementation of a system are proposed, which will take the text from mammography reports as input. This information will undergo a cleaning and standardization process to extract the BI-RADS, which will be used for priority assignment. However, a challenge arises when BI-RADS information is not found in all reports. In such cases, it will be necessary to make a prediction to determine its location.

Once both stages are completed, the results will be displayed on a platform that allows doctors to visualize the status of their patients and their examination history. This platform will serve as a valuable support tool in the field of healthcare, providing an overview and detailed insight to facilitate decision-making.

This document will focus on data processing and the integration of different modules to enable this processing to be carried out directly from the platform.

Finally, it is important to mention that this thesis project is being developed as part of the Multidisciplinary Thesis Program, in collaboration with the Arturo López Pérez Foundation, with the aim of addressing the previously stated problem.

Glosario

- **BI-RADS:** Sus siglas provienen de *Breast Imaging Reporting and Data System*. Este sistema de informes se utiliza para describir de manera estandarizada los resultados de las mamografías, ecografías mamarias o imágenes por resonancia magnética de las mamas, clasifica los resultados de las pruebas según 7 categorías, que van desde normal o benigno (no canceroso) hasta altamente sospechoso o maligno (cáncer).
- **JavaScript** Lenguaje de programación interpretado. Se define como orientado a objetos, basado en prototipos, imperativo, débilmente tipado y dinámico.
- **KNN:** Clasificador de aprendizaje supervisado, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual
- **MNB:** Multinomial Naive Bayes es un algoritmo de aprendizaje probabilístico que se utiliza sobretodo en procesamiento del lenguaje natural.
- **NLP:** Natural Language Processing o Procesamiento del Lenguaje Natural es una disciplina que se enfoca principalmente en la comprensión, el manejo y la generación del lenguaje natural por parte de las máquinas.
- **Pandas:** Biblioteca de software escrita para el lenguaje de programación Python utilizada para la manipulación y el análisis de datos.
- **Python** Lenguaje de programación interpretado, orientado a objetos, de alto nivel y semántica dinámica.
- **SVM:** Support Vector Machines (o Máquinas de Vectores de Soporte en español) es un algoritmo de aprendizaje automático que es utilizado para realizar análisis de clasificación, regresión y detección de valores atípicos.
- **TNM:** Sistema de estadificación del cáncer que permite comprender su gravedad. Donde:
 - T indica el tamaño y la extensión del tumor.
 - N indica el número de ganglios linfáticos cercanos que son cancerosos.
 - M indica la presencia de metástasis.

Índice general

1. Introducción	8
1.1. Contexto y Motivación	8
1.2. Problemática por resolver	9
1.3. Acercamiento a la Solución	10
1.4. Objetivos	11
1.4.1. Objetivo General	11
1.4.2. Objetivos Específicos	11
1.5. Contribuciones	12
1.6. Estructura de la memoria	13
2. Marco Teórico	14
2.1. Funcionamiento Actual	14
2.2. Estado del Arte y la Técnica	15
2.3. Discusión	20
3. Ingeniería de Requisitos	21
3.1. Requisitos Funcionales	21
3.2. Requisitos No Funcionales	22
3.3. Requisitos de Ambiente	23
3.3.1. Hardware de Desarrollo	23
3.3.2. Software de Desarrollo	23
3.4. Perfiles de Usuario	24
4. Diseño de la Solución	25
4.1. Diagrama de Contexto	25
4.2. Diagrama de Arquitectura	27

4.3. Enumeración de Módulos	28
4.4. Matriz de Requisitos Funcionales y Módulos	28
4.5. Módulos	29
4.5.1. Limpieza de Datos	29
4.5.2. Identificación y Extracción de Clasificación BI-RADS	29
4.5.3. Predicción de Prioridad	30
4.5.4. Asignación de Prioridad de los Pacientes	30
4.5.5. Interfaz Usuaría	31
4.6. Gestión de Riesgos	32
4.6.1. Supuestos	32
4.6.2. Dependencias	32
4.6.3. Restricciones	32
4.6.4. Riesgos	33
5. Desarrollo de Solución	34
5.1. Interpretación de los datos	36
5.2. Limpieza de los datos	37
5.3. Extracción de Información	39
5.4. Asignación de Prioridad	40
5.5. Integración en plataforma web	42
6. Conclusiones y Trabajo Futuro	44
6.1. Conclusión	44
6.2. Trabajo Futuro	45
Bibliografía	46

Capítulo 1

Introducción

1.1. Contexto y Motivación

En los últimos años, el cáncer se ha convertido en una de las principales causas de mortalidad en todo el mundo [3], y lamentablemente, el pronóstico para los próximos años no es alentador. Esto se debe a factores como las tasas de crecimiento demográfico, los estilos de vida y el envejecimiento de la población. Se estima que para el año 2040, el riesgo de padecer esta enfermedad aumentará en un 56.4 % en comparación con el año 2020 [4]. Nuestro país no está exento de esta tendencia preocupante. Según los datos de Globocan 2020 [5], en ese año se registraron más de 54,000 nuevos casos de cáncer en el país. El cáncer de mama fue el tipo más común entre las mujeres, con alrededor de 5,300 nuevos casos anuales.

Para el cáncer de mama, la mamografía destaca como el examen más efectivo, en términos de costo para la detección en etapas tempranas. Esto adquiere una importancia crucial, dado que, a pesar de ser una de las principales causas de muerte en mujeres, presenta una tasa de curación que supera el 90 % [2] cuando se detecta de manera temprana. Sin embargo, aún no existe un procedimiento estandarizado para orientar el tratamiento y priorizar los casos de mayor urgencia.

1.2. Problemática por resolver

En la mayoría de los casos de cáncer de mama, el primer paso para el diagnóstico es la mamografía. Este examen implica el uso de rayos X para obtener una imagen de la mama, que posteriormente es evaluada por un radiólogo. El radiólogo categoriza los resultados mediante un sistema de evaluación llamado BI-RADS y elabora un informe que generalmente describe el mamograma utilizando términos médicos estandarizados.

En esta primera etapa, los resultados del BI-RADS se dividen en siete categorías. Estas categorías van desde hallazgo negativo hasta altamente sospechoso o maligno. Además, se incluyen casos no concluyentes que requieren estudios adicionales para su evaluación.

El problema que surge a partir de estos resultados es la falta de un sistema de priorización en la atención. Frecuentemente, los pacientes son tratados en orden de llegada, sin tener en cuenta la categorización de sus resultados. Esto puede ocasionar retrasos en el proceso y perjudicar directamente a los pacientes. En esta enfermedad, el factor tiempo es fundamental. Si el tratamiento no se inicia de manera temprana, el paciente podría avanzar a un estadio más avanzado y de difícil tratamiento.

1.3. Acercamiento a la Solución

Con el objetivo de resolver el problema anteriormente planteado, se propone diseñar e implementar un sistema que reciba como entrada el texto de los informes de mamografías y determine la priorización del paciente según su categoría BI-RADS.

Para poder realizar esto, se cuenta con dos etapas. En primer lugar, un algoritmo recibe los informes de donde se extrae el BI-RADS, para asignar el nivel de prioridad a los pacientes. En segundo lugar, se cuenta con una plataforma de visualización donde se puede observar el estado de los pacientes y el historial de exámenes relacionados de un paciente en particular.

Es necesario mencionar que no todos los informes cuentan con la clasificación BI-RADS, por lo que también se realizará una predicción de la prioridad que deben tener aquellos pacientes.

1.4. Objetivos

1.4.1. Objetivo General

- Generar un sistema de información integrado que permita la lectura de informes de mamografías y que en base a la información obtenida permita la priorización de los pacientes.

1.4.2. Objetivos Específicos

- Diseñar e implementar un módulo para la limpieza y estandarización del texto de los informes de mamografías.
- Diseñar e implementar un módulo para la extracción de información de los informes de mamografías.
- Integrar módulos de limpieza y extracción con el modelo de predicción.
- Integrar funcionalidades de los módulos en una plataforma web.

1.5. Contribuciones

Esta memoria fue desarrollada en conjunto entre Bastián Cornejo, Alejandro Viertel y la autora de este documento.

El desarrollo del sistema se realizó de manera modular, donde Alejandro Viertel va a trabajar con el *Módulo de Predicción*, Bastián Cornejo con *Módulo de interfaz Usuaria* y Katty Herrera con *Módulo de Procesamiento de datos e integración de módulos*.

El equipo completo se encargó de mantener reuniones periódicas con la Fundación Arturo López Perez (FALP) quienes nos plantearon este desafío, para obtener retroalimentación durante todo el proceso de desarrollo.

Si se desea conocer en detalle los aspectos técnicos relevantes al trabajo de Bastián y Alejandro, se pueden consultar sus escritos de memoria, que se encontrará disponible en el Repositorio USM al finalizar sus respectivos procesos de titulación.

1.6. Estructura de la memoria

En el capítulo 1 de la memoria se encuentra el contexto y la motivación que llevan a la problemática que nuestro equipo buscó resolver durante este proceso. Además, se da un primer acercamiento a lo que se quiere lograr con la solución y cómo se llevará a cabo.

En el capítulo 2 se presenta la investigación realizada para conocer el estado actual en nuestra área de interés, qué trabajos o investigaciones existentes se asemejan a lo que estamos buscando y cómo podemos iniciar nuestro trabajo a partir de lo aprendido.

En el capítulo 3 se pueden apreciar los requisitos recopilados a partir de lo solicitado por la contraparte de este desafío y quiénes serán los usuarios finales de este sistema.

En el capítulo 4 se muestra el diseño generado para la solución propuesta, cómo se estructura y cuáles serán sus funcionalidades y alcance.

El capítulo 5 presenta el trabajo desarrollado, explicando el proceso desde un inicio y cómo se llevó a cabo finalmente la parte en la que está enfocado este trabajo.

Finalmente, en el capítulo 6 se encuentran las conclusiones obtenidas a partir de este trabajo y se mencionan las posibles áreas de trabajo futuro que se pueden realizar en este proyecto.

Capítulo 2

Marco Teórico

2.1. Funcionamiento Actual

En el contexto del cáncer de mama, la mamografía desempeña un papel fundamental en la detección temprana de la enfermedad. Sin embargo, una vez completado este examen, es esencial que el médico oriente al paciente sobre los pasos a seguir según el resultado obtenido. Esto puede implicar programar controles periódicos basados en el historial del paciente o la realización de exámenes adicionales para complementar el diagnóstico. El problema radica en la ausencia de un protocolo para establecer la prioridad en la atención de los pacientes.

En nuestro país, la atención a pacientes con sospecha de cáncer de mama o casos confirmados suele basarse en el principio de “orden de llegada”. Este enfoque resulta ineficiente y puede perjudicar a aquellos pacientes que requieren procedimientos de manera más urgente. En resumen, no existe un sistema establecido para asignar prioridad en la atención médica de pacientes con cáncer de mama, lo que subraya la necesidad de abordar esta deficiencia en la atención sanitaria.

2.2. Estado del Arte y la Técnica

En el ámbito de la salud, abunda la información en forma de textos no estructurados. Con el tiempo, se han desarrollado técnicas para extraer datos de estos textos con el fin de realizar análisis y predicciones.

Entre las técnicas de procesamiento de lenguaje natural (*NLP* por sus siglas en inglés) utilizadas en la actualidad, destaca el empleo de *word embeddings* [6]. Esta técnica convierte palabras en representaciones numéricas asignando valores similares a términos con significados semejantes. El propósito es capturar el significado semántico de las palabras en un texto. Un ejemplo de esta técnica es el sistema "Apache cTAKES" [7], una plataforma para extraer información de documentos médicos electrónicos. Este sistema se utiliza, por ejemplo, en la detección de la recurrencia del cáncer de mama en documentos clínicos [8].

Además, se emplean otras técnicas como *Word2vec*, que se utiliza para adquirir propiedades semánticas de las palabras, y *TFIDF*, que se emplea para determinar la relevancia de los elementos mediante el análisis de su frecuencia de aparición. La aplicación de estas técnicas puede potenciar el desempeño de clasificadores como *Support Vector Machine (SVM)*, *Naive Bayesian (NB)*, *Extreme Gradient Boosting (XGBoost)*, y *Multilevel Fuzzy Min-Max Neural Network (MLF)*. Este mejoramiento se evidencia en un análisis llevado a cabo por la *Universidad Islámica Azad* [9], en el cual se comparan las técnicas combinadas con los informes del sistema de información hospitalaria mediante diversos clasificadores.

Otra técnica utilizada implica el entrenamiento de redes neuronales mediante *word embeddings*. Uno de los modelos más sobresalientes en este ámbito es la red XLNET [10], que representa una mejora del modelo BERT (*Bidirectional Encoder Representations from Transformers*) [11], ambos diseñados por Google. El modelo BERT hace uso de *transformers*, un mecanismo de atención que aprende las relaciones contextuales entre las palabras de un texto. A diferencia de los modelos direccionales que leen la entrada de texto secuencialmente (de izquierda a derecha o de derecha a izquierda), BERT analiza toda la secuencia de palabras a la vez, considerando el contexto tanto a la izquierda como a la derecha de cada palabra [12]. La mejora que ofrece XLNET radica en que es un modelo auto regresivo que emplea la técnica *Transformers - XL* para trabajar con una secuencia de texto y obtener información de secuencias previas para realizar predicciones [13].

Las aplicaciones y el rendimiento de las técnicas XLNET y BERT en el ámbito de la salud para analizar textos no estructurados han sido exhaustivamente investigados por universidades e instituciones de investigación. Ejemplos de estos estudios incluyen la investigación realizada por la Universidad de

Minnesota, que emplea BERT para la extracción de fenotipos relacionados con el cáncer [14], y otra investigación llevada a cabo por la Universidad de Harvard, donde se utiliza XLNET para el modelado de notas clínicas y la predicción de la ventilación mecánica prolongada [15].

En el campo de la extracción de información de informes médicos, se encuentra el sistema MedLEE (Medical Language Extraction and Encoding) [16], una solución privada desarrollada por la Universidad de Columbia. MedLEE se encarga de extraer y codificar información clínica de informes médicos, originalmente diseñado para la detección de tuberculosis a partir de informes radiológicos, pero en la actualidad ha ampliado su aplicación a una variedad más amplia de tipos de informes, incluyendo el ámbito del cáncer de mama y su uso en mamografías. Como alternativas de código abierto a MedLEE, destacan cTAKES y HITEEx (Health Information Text Extraction) [17]. HITEEx es un software desarrollado específicamente para resolver desafíos en el dominio médico, extrayendo información como diagnósticos y medicamentos. Este software se basa en GATE (General Architecture for Text Engineering) [18], un sistema de código abierto que proporciona una infraestructura para el desarrollo e implementación de componentes de software para el procesamiento del lenguaje natural.

Debido a que los informes clínicos son textos no estructurados que varían según la institución y el médico que los redacta, el uso de redes neuronales puede no ser la elección más adecuada para desarrollar una solución genérica. La diversidad en el formato de estos documentos requiere una gran cantidad de datos para adaptar una red neuronal a este problema de manera efectiva. Por esta razón, varias investigaciones han propuesto enfoques alternativos, todos centrados en la misma idea fundamental: crear un sistema que extraiga términos médicos de los informes clínicos y, a través de algoritmos personalizados que analicen la presencia de estos términos, asigne una puntuación al informe para determinar el nivel de gravedad o la afección del paciente.

En agosto de 2021, un equipo de investigadores del *McGill University Health Centre* en Montreal, Canadá, desarrolló un algoritmo de procesamiento del lenguaje natural (NLP) que se puede aplicar de manera generalizada para extraer información relacionada con el dolor de documentos clínicos de pacientes con cáncer [19]. Para comprender la terminología utilizada en los informes médicos, los autores recurrieron al repositorio de vocabularios biomédicos conocido como el *Unified Medical Language System (UMLS)* [20], el cual fue desarrollado por la Biblioteca Nacional de Medicina de los Estados Unidos.

En el proceso de extracción de términos médicos de los documentos clínicos, los investigadores emplearon diversas herramientas. En primer lugar, utilizaron MetaMap [21], un programa diseñado específicamente para acceder a los conceptos médicos del conjunto de datos UMLS. Además, desarrollaron un algoritmo en Python haciendo uso de la biblioteca spaCy [22]. La principal función de spaCy fue realizar tareas como la eliminación de espacios en blanco, la detección de caracteres especiales, la

identificación de abreviaturas médicas y la determinación de si los conceptos médicos extraídos indicaban negación o no.

Los resultados obtenidos se integraron en un pipeline de procesamiento del lenguaje natural (NLP) que previamente había sido entrenado y validado utilizando conjuntos de datos públicos, en este caso, MIMIC-III e i2b2 [23, 24]. Este pipeline incluía un clasificador basado en reglas y una serie de tablas que contenían conceptos médicos clasificados por enfermedades. Se llevó a cabo una verificación para determinar si los conceptos extraídos en el paso anterior coincidían con los conceptos presentes en estas tablas o si se trataba de menciones de afecciones no relevantes. Esto condujo a la asignación de un valor denominado *score pain*, que reflejaba la relevancia del dolor en el informe clínico. Finalmente, utilizando esta puntuación, se calculó la intensidad promedio del dolor en el informe clínico, lo que permitió determinar si se identificaba alguna enfermedad relevante en el paciente a través del informe o no.

Desde un enfoque probabilístico, es importante destacar las redes bayesianas o de creencias, que son modelos gráficos utilizados para representar variables estocásticas mediante nodos y conexiones que indican las influencias probabilísticas directas entre estas variables. Estas redes resultan muy útiles para el razonamiento en situaciones de incertidumbre, particularmente en sistemas de apoyo para la toma de decisiones médicas.

El concepto fundamental detrás de las redes bayesianas es asignar valores de probabilidad a las variables para expresar el grado de creencia en diferentes estados posibles. Por ejemplo, consideremos un nodo que representa la variable "metástasis", que puede tener dos estados: "sí" o "no". Cada uno de estos estados se asocia con un valor de probabilidad, y la suma de estos valores siempre equivale a 1. Estas redes permiten identificar relaciones entre diversos diagnósticos, hallazgos de imágenes médicas o resultados de análisis de laboratorio. Además, se emplean para planificar exámenes de diagnóstico, tomar decisiones sobre intervenciones terapéuticas y ayudar en la interpretación de informes de diagnóstico y hallazgos médicos.

Un ejemplo destacado es el consorcio MIDAS, que ha desarrollado MammoNet [25], un modelo de cáncer de mama que se vale de redes bayesianas para asistir en el diagnóstico a partir de mamografías. Es relevante señalar que este modelo se construyó con una cantidad limitada de datos de pacientes y continúa en desarrollo por parte del consorcio. Su objetivo es obtener un conjunto más amplio de datos probabilísticos a partir de una extensa serie de casos. Aunque el modelo inicial se basa en datos limitados y no puede considerarse completamente concluyente, demuestra el potencial de las redes bayesianas como herramienta de apoyo en el diagnóstico y la planificación de procesos de diagnóstico adicionales.

Recientemente, se ha propuesto un sistema predictivo basado en modelos bayesianos para la detección temprana del cáncer de mama [26]. Este sistema se centra principalmente en variables como la edad, el peso, el índice de masa corporal, el consumo de alcohol, la historia familiar de cáncer, la menopausia, los embarazos, entre otros. Para su funcionamiento, emplea dos bases de datos separadas, una para pacientes con cáncer y otra para aquellos sin la enfermedad, utilizando una de ellas para entrenar y realizar pruebas. El sistema se vale de un clasificador bayesiano que utiliza 24 características para clasificar a los pacientes en la base de datos. Luego, se lleva a cabo una validación para determinar la precisión del sistema, obteniendo un alto porcentaje de aciertos incluso con un mínimo de 5 características, que por sí solas no serían suficientes para identificar el cáncer.

En el ámbito de las herramientas para el trabajo relacionado con NLP, una de las más utilizadas en la actualidad es la biblioteca de Python llamada spaCy [22]. Se trata de una biblioteca de código abierto ampliamente utilizada en el procesamiento avanzado del lenguaje natural (NLP) y el aprendizaje automático. spaCy se emplea para desarrollar sistemas de extracción de información, comprensión del lenguaje natural y preprocesamiento de texto para modelos de aprendizaje automático.

En el entorno de Python, el Natural Language Toolkit (NLTK) [27] ha sido una plataforma principal para el procesamiento general de texto. Consiste en una serie de bibliotecas NLP altamente funcionales diseñadas para el análisis de texto basado en reglas y aprendizaje automático. A pesar de su éxito en numerosos proyectos e investigaciones, NLTK tiene ciertas limitaciones, como su enfoque en el procesamiento de texto como listas de cadenas y la falta de una arquitectura unificada, lo que puede dificultar su capacidad para escalar y manejar grandes conjuntos de datos.

Por otro lado, spaCy ofrece una arquitectura robusta que permite construir y compartir módulos de procesamiento de lenguaje natural (NLP) personalizados y de alto rendimiento. Se basa en un enfoque orientado a objetos para el procesamiento de texto, lo que facilita la integración de métodos estadísticos y de aprendizaje automático con el procesamiento basado en reglas. Además, spaCy utiliza Cython, un compilador estático optimizador para Python, que genera código C o C++ altamente eficiente, lo que resulta en un rendimiento excepcional. En resumen, spaCy proporciona un marco de trabajo para la construcción modular de pipelines de procesamiento de lenguaje natural personalizados.

Reconociendo la necesidad de adaptar el procesamiento de lenguaje natural a diferentes dominios, han surgido varios modelos y conjuntos de herramientas especializados en textos técnicos. En nuestro contexto, destaca la biblioteca medspaCy [28, 29], que se basa en spaCy y proporciona capacidades avanzadas para el análisis de informes clínicos. Los componentes de medspaCy tienen una configuración predeterminada y ofrecen numerosos parámetros de personalización opcionales. Esto permite una configuración rápida y un aprendizaje ágil, al tiempo que se adapta a necesidades específicas. En

resumen, medspaCy facilita la detección de términos médicos y el análisis exhaustivo de informes clínicos utilizando las herramientas de procesamiento de lenguaje natural que spaCy ofrece.

En el contexto de las redes neuronales, otra alternativa a considerar es el uso de modelos pre-entrenados desarrollados por John Snow Labs, específicamente Spark NLP [30, 31]. Spark NLP es una biblioteca de Python que hace uso de componentes de Spark ML, como los transformadores y estimadores previamente mencionados. Su enfoque de “entrenamiento y validación” asegura una partición adecuada de los datos para el entrenamiento de modelos. Es importante destacar que una limitación de los modelos pre-entrenados es que la mayoría de ellos están disponibles en inglés, lo que implica que las entidades deben ser traducidas para adaptarse a otros idiomas, lo que puede generar variaciones en los resultados. Sin embargo, existe un modelo de interés basado en esta biblioteca denominado *Detectar morfología de tumores en textos en español* [32], diseñado específicamente para identificar entidades relacionadas con la morfología de los tumores en español. Este modelo fue entrenado utilizando el conjunto de datos CANTEMIST [33], que se centra en el análisis oncológico en español.

2.3. Discusión

El objetivo de este trabajo es asignar prioridades a los pacientes utilizando la información contenida en sus informes de mamografía, lo que implica la extracción de datos clave como el BI-RADS y la predicción de este valor o una prioridad basada en dicho parámetro.

En este contexto, se han mencionado herramientas como *medspaCy* que pueden ser útiles para extraer información relevante de los informes médicos. Sin embargo, es importante tener en cuenta que esta herramienta no admite el idioma español de forma nativa, lo que requeriría adaptaciones y modificaciones considerables para su aplicación efectiva en dicho idioma. Además, personalizar todas las funcionalidades puede ser un proceso que demande tiempo, especialmente al tratar grandes volúmenes de datos.

En cuanto a la predicción, aunque los ejemplos previamente mencionados no se ajusten exactamente a su caso, proporcionan un punto de partida para explorar enfoques de Machine Learning y Redes Neuronales aplicables a su contexto. Sin embargo, es importante destacar que los datos disponibles presentan desafíos específicos, como el desbalanceo, que implica una gran diferencia en la cantidad de datos entre las diferentes clasificaciones, especialmente las más altas que son críticas. Ante esta situación, es apropiado considerar el uso de clasificadores como SVM (Support Vector Machine), KNN (K-Nearest Neighbors) y MNB (Multinomial Naive Bayes) para evaluar los resultados y abordar este desbalance de datos.

Capítulo 3

Ingeniería de Requisitos

3.1. Requisitos Funcionales

RF1 El sistema debe poder predecir la prioridad en aquellos informes donde no se indique explícitamente el BI-RADS.

RF2 El sistema debe solicitar los datos correspondientes del paciente para iniciar el análisis.

RF3 El sistema debe tener la capacidad de identificar los conceptos y diagnósticos médicos.

RF4 El sistema debe entregar información del historial de un paciente.

RF5 El sistema permitirá al médico visualizar a todos sus pacientes donde podrá ver toda la información de este incluyendo su nivel de prioridad.

3.2. Requisitos No Funcionales

RNF1 El sistema no debe otorgar falsos negativos.

RNF2 El sistema debe poder funcionar montado en el servidor interno de la FALP.

3.3. Requisitos de Ambiente

3.3.1. Hardware de Desarrollo

Debido a que no se utilizan redes neuronales o procesamiento de alto calibre hasta el momento, no se requiere de alguna GPU o especificaciones en particular.

El hardware de desarrollo necesario es una computadora que soporte el software de desarrollo detallado a continuación.

3.3.2. Software de Desarrollo

Respecto al software de desarrollo requerido básico se cuenta con lo siguiente:

- Python 3
 - Jupyter
 - Scikit-learn
 - Pandas
 - Numpy
 - Pickle
 - Flask
- JavaScript
 - React
 - Node.Js

3.4. Perfiles de Usuario

Usuario	Perfil				
	Socio-económico y Cultural	Ocupacional	Etario	Características físicas, fisiológicas, psicológicas	Otros
Médico Oncólogo	Trabajador en la FALP	Médico Oncólogo	25 - 65	Busca realizar su trabajo de forma más eficiente para así diagnosticar a un mayor número de pacientes en un menor tiempo	Profesional integral con conocimiento en el área oncológica, especialmente en el diagnóstico y tratamiento del cáncer de mama
Radiólogo	Trabajador en la FALP	Radiólogo Médico	25 - 65	Busca realizar su trabajo de forma más eficiente para así diagnosticar a un mayor número de pacientes en un menor tiempo	Profesional integral con conocimiento en el área oncológica, especialmente en el diagnóstico y tratamiento del cáncer de mama

Tabla 3.1: Perfiles de Usuario

Capítulo 4

Diseño de la Solución

4.1. Diagrama de Contexto

El diagrama de contexto nos permite visualizar las interacciones entre los diferentes componentes y actores con nuestro sistema a desarrollar. En la Figura 4.1 se puede observar este diagrama enfocado en las interacciones con nuestra Aplicación Web.

Se presentan dos actores: el trabajador de FALP, responsable de subir la información a la plataforma para que sea procesada utilizando la aplicación web, y el médico a cargo, quien puede revisar la prioridad otorgada a sus pacientes y acceder al historial de alguno de ellos a partir de la información obtenida de los informes de mamografías.

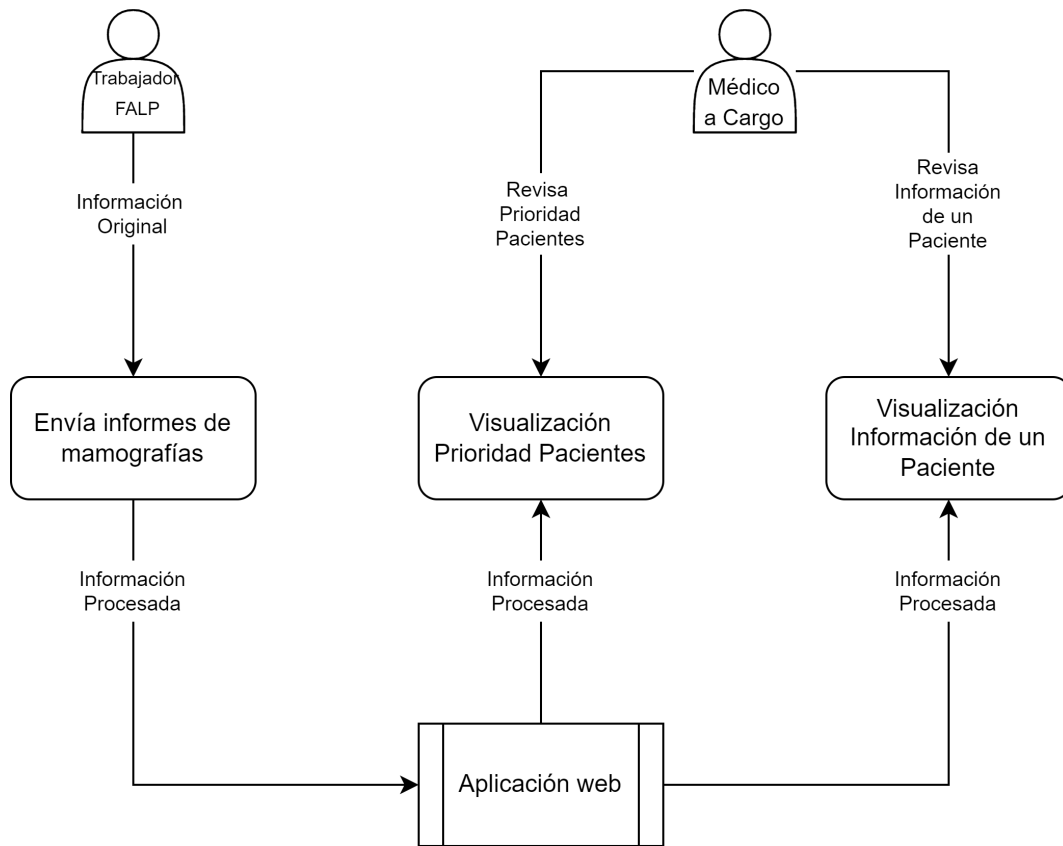


Figura 4.1: Diagrama de contexto del sistema.

4.2. Diagrama de Arquitectura

En el siguiente diagrama se pueden observar los distintos módulos y componentes del sistema. Permitiendo mostrar como interactúan entre sí, iniciando con los informes de mamografía, su limpieza y extracción de información para luego ser dividido en dos módulos paralelos donde por una parte se asigna prioridad a partir de la información extraída y por otro lado se utiliza el modulo de predicción de prioridad para luego unir ambos resultados y proceder a la visualización en la interfaz usuaria.

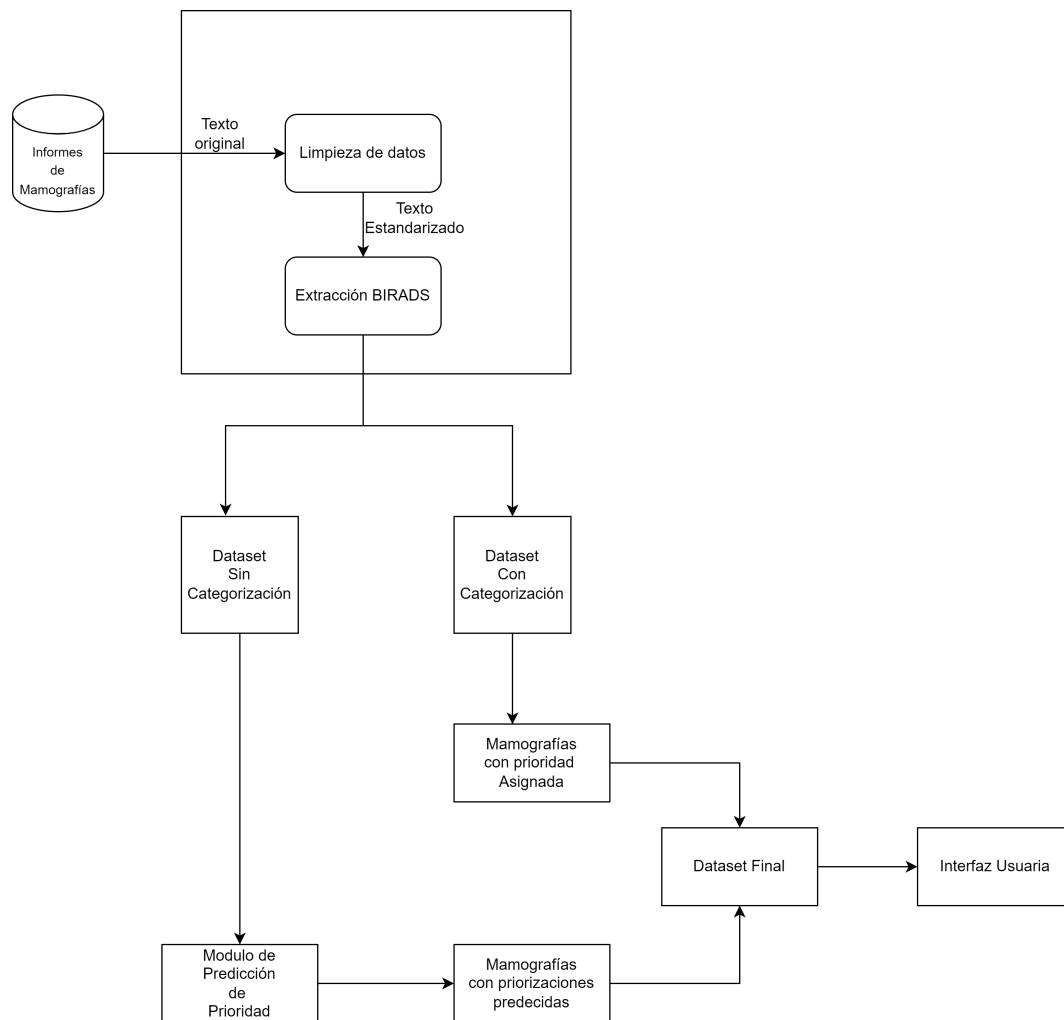


Figura 4.2: Diagrama de arquitectura del sistema.

4.3. Enumeración de Módulos

Módulo	Propósito	Sección
Limpieza de Datos	Este módulo permite corregir datos no válidos, incompletos o incorrectos, eliminación de caracteres o datos innecesarios	5.1
Identificación y Extracción de Clasificación BI-RADS	Este módulo permite obtener la clasificación BI-RADS de los informes e identificar cuáles no cuentan con dicha clasificación	5.2
Predicción de Prioridad	Predice la categoría en la que se encuentra un paciente cuando no se menciona su BI-RADS	5.3
Asignación de Prioridad de los Pacientes	Prioriza los pacientes de cáncer de mama en base a su BI-RADS	5.4
Interfaz Usuaría	Permite que el médico visualice la priorización y los datos de los pacientes que están a su cargo	5.5

Tabla 4.1: Enumeración de Módulos

4.4. Matriz de Requisitos Funcionales y Módulos

	Limpieza de Datos	Identificación y Extracción de BI-RADS	Predicción de Prioridad	Asignación de Prioridad de los Pacientes	Interfaz Usuaría
RF1			X		
RF2	X				
RF3		X		X	
RF4					X
RF5					X

Tabla 4.2: Matriz de Requisitos Funcionales y Módulos

4.5. Módulos

4.5.1. Limpieza de Datos

- **Propósito:** Este módulo permite corregir datos no válidos, incompletos o incorrectos, así como eliminación de caracteres o datos innecesarios.
- **Alcance:** Es el punto de inicio del proceso, donde se preparan los datos para ser utilizados en el análisis posterior.
- **Dependencias:** No depende de otros módulos.
- **Supuestos:** No se conoce de antemano la ubicación de la información relevante del documento médico. Por ejemplo: En el caso de una mamografía, el texto del informe médico puede estar después de las instrucciones del documento, mientras que en otro caso puede presentarse antes de las instrucciones.
- **Restricciones:** Los informes deben contar con una extensión mínima.
- **Estructura General:** Este módulo toma por entrada el dataset que contiene los informes de exámenes, corrige datos erróneos, elimina caracteres innecesarios y estandariza los datos.

4.5.2. Identificación y Extracción de Clasificación BI-RADS

- **Propósito:** Este módulo tiene como propósito obtener la clasificación BI-RADS de los informes e identificar aquellos que no cuentan con dicha clasificación.
- **Alcance:** Esta etapa es el inicio del análisis, extrayendo los BI-RADS de los informes para poder ordenarlos por su prioridad e identificar aquellos que no cuentan con la clasificación BI-RADS para continuar hacia la predicción de prioridad.
- **Supuestos:** La limpieza realizada en el módulo 4.5.1 funciona correctamente y estandariza el texto, lo cual facilita su análisis.
- **Restricciones:** Los valores de BI-RADS identificados deben estar dentro del rango 0-6; de lo contrario, serán considerados como no clasificados.
- **Estructura General:** Este módulo cuenta con el dataset limpio y estandarizado como entrada, para así realizar la búsqueda de la mención de BI-RADS en los informes y entregar como salida

dos datasets: uno con los informes donde se encontró la clasificación y otro con los informes sin clasificación.

4.5.3. Predicción de Prioridad

- **Propósito:** Este módulo tiene como objetivo predecir la prioridad de un paciente cuando no se encuentra mención de su clasificación BI-RADS en el informe de mamografía.
- **Alcance:** Permite realizar una clasificación de prioridad a pacientes cuyos informes no mencionan su nivel de BI-RADS.
- **Dependencias:** Depende del módulo 4.5.2.
- **Supuestos:** Los resultados entregados por el módulo 4.5.2 son correctos.
- **Restricciones:** Se espera que los resultados falsos negativos, es decir, cuando se predice una prioridad menor a la real, sean mínimos.
- **Estructura General:** Toma como entrada los resultados del módulo 4.5.2. A partir de estos datos, se realizará la predicción de la categoría en la que se encuentra el paciente utilizando clasificadores.

4.5.4. Asignación de Prioridad de los Pacientes

- **Propósito:** Este módulo tiene como objetivo priorizar a los pacientes con cáncer de mama en función de su clasificación BI-RADS asociada.
- **Alcance:** Una vez realizada la extracción de BI-RADS, se encarga de asignar la prioridad a aquellos informes donde se identificó su clasificación.
- **Dependencias:** Depende del módulo 4.5.2.
- **Restricciones:** Para diferenciar a los pacientes que cuenten con la misma prioridad, se debe mostrar información adicional en la visualización.
- **Estructura General:** A partir de la información obtenida por el módulo 4.5.2, este módulo realiza la priorización de los pacientes.

4.5.5. Interfaz Usuaría

- **Propósito:** Permite que el médico visualice la priorización y los datos de los pacientes que están a su cargo.
- **Alcance:** En este módulo se visualizarán los datos de los pacientes ordenados por la prioridad que les fue otorgada anteriormente.
- **Dependencias:** Depende de los resultados de priorización otorgados por el módulo 4.5.3 y 4.5.4.
- **Supuestos:** Se tiene como entrada los resultados de los dos módulos anteriores.
- **Restricciones:** Todos los datos entregados a este módulo cuentan con una prioridad asignada.
- **Estructura General:** El médico puede elegir visualizar los datos de todos los pacientes, se mostrarán en una tabla con algunos datos personales y la prioridad que tienen. Por otro lado, podrá seleccionar un paciente en particular y se mostrará el historial de informes con lo que se podrá visualizar la evolución con el tiempo.

4.6. Gestión de Riesgos

4.6.1. Supuestos

1. Se contará con el software definido en *Requisitos de Ambiente* para desarrollar el proyecto.
2. Se contará con el hardware definido en *Requisitos de Ambiente* para desarrollar el proyecto.
3. Habrá comunicación efectiva con el cliente y se comprenderán sus requerimientos.
4. Habrá comunicación efectiva entre los integrantes del equipo para cumplir con los objetivos.
5. Es necesario contar con la carga de documentos para ser analizados.
6. Los datos otorgados contarán con datos adecuados y de calidad para el proyecto.
7. Los usuarios sabrán utilizar la aplicación.
8. Los usuarios comprenderán la salida que les muestra la plataforma.

4.6.2. Dependencias

1. Todos los participantes del equipo contarán con conexión a Internet.
2. Todos los participantes contarán con el hardware adecuado para trabajar en el proyecto.
3. La contraparte facilitará una base de datos con los datos a utilizar en el proyecto.

4.6.3. Restricciones

1. Se utilizará el lenguaje de programación Python para el procesamiento de los datos, ya que, es un lenguaje en el que todos los integrantes del equipo tienen experiencia y cuenta con muchas herramientas útiles en este contexto.
2. Se utilizará el lenguaje de programación JavaScript para el desarrollo de la plataforma web, puesto que es soportado por la mayoría de los navegadores y cuenta con múltiples herramientas para facilitar el desarrollo.

4.6.4. Riesgos

Riesgo	Medida de Mitigación
Los integrantes del equipo realizan diferentes interpretaciones respecto al trabajo a realizar	Se realiza una reunión semanal de equipo y se definen canales de comunicación rápidos manteniendo así una comunicación activa.
No hay una buena comprensión de los requisitos del cliente	Se realizan reuniones semanales en las que se plantean distintos tópicos e inquietudes. También se cuenta con un canal de comunicación rápido que sirve para mitigar las dudas que surjan.
Alguno de los integrantes del equipo no cuenta con conexión a internet	Dependiendo de la naturaleza de la tarea a realizar dicho integrante puede trabajar de forma <i>offline</i> , utilizar datos móviles o utilizar la conexión a internet de la universidad.

Tabla 4.3: Riesgos

Capítulo 5

Desarrollo de Solución

En este capítulo, se detallará el proceso de desarrollo de la solución planteada. Sin embargo, antes es necesario mencionar que este planteamiento no fue el inicial con el que comenzamos a trabajar, por lo que también se comentará brevemente el proceso realizado.

Inicialmente, este proyecto tenía un objetivo distinto donde se buscaba cubrir todo el proceso desde el diagnóstico del cáncer hasta la asignación de tratamiento de los pacientes considerando su prioridad, lo cual era mucho más extenso y los datos con los que contábamos (Informes de mamografías) no permitían realizar algo así.

Decidimos acotar el problema a otorgar prioridad a los pacientes a partir del informe de mamografía. Aquí es donde surge la intención de utilizar el BI-RADS para esto, ya que la otra opción sería utilizar el estadio otorgado por la clasificación TNM, pero esta no se puede obtener solamente a partir de un informe de mamografía, por lo que se necesitarían otros tipos de exámenes.

En cuanto a la extracción de información, contemplamos utilizar *medspaCy*, realizando algunas modificaciones para utilizarlo en español. Logramos obtener algunos resultados básicos con la identificación de algunos términos médicos, pero no se lograba aprovechar completamente su utilidad debido a los problemas con el cambio de idioma. Además, no estábamos obteniendo la información que necesitamos (BI-RADS) y el tiempo de procesamiento era muy alto en comparación a realizar este análisis con herramientas de *Python*.

Es importante recordar que el trabajo realizado en esta memoria corresponde al procesamiento de los datos y la integración de estos con los módulos de predicción e interfaz gráfica. A continuación, se detallará el funcionamiento según lo mencionado anteriormente. Si desea comprender en profundidad el desarrollo del modelo de predicción o el funcionamiento de la interfaz gráfica, se recomienda revisar el trabajo realizado por Alejandro Viertel y Bastián Cornejo, respectivamente.

5.1. Interpretación de los datos

El primer paso es comprender el tipo de dato con el que estamos trabajando. A continuación se describirán los tipos de datos con los que contamos:

Nombre	Descripción
Fecha Firma	Fecha en la que el informe fue ingresado al sistema.
Study ID	Identificador del informe.
Type	Tipo de examen de mamografía.
Informe	Texto correspondiente al informe de mamografía.
Fecha Diag	Fecha cuando se realizó diagnóstico en caso de ocurrir.
Diag	Indica si existe un diagnóstico de tumor maligno.
Fecha de Nacimiento	Fecha de nacimiento del paciente.
ID	Identificador del paciente.

Tabla 5.1: Dataset a utilizar.

En general, para el análisis, los datos más importantes a considerar son los informes, ya que contienen una descripción de los resultados observados en el examen, de donde se puede extraer información relevante, como la clasificación BI-RADS. Además, se utiliza la edad del paciente al momento de ordenar según prioridad, datos que son anexados luego de la limpieza de datos, extracción de información y asignación de prioridad.

Para los pasos posteriores, como la asignación de prioridad, es necesario tener conocimiento sobre qué significan las categorías de BI-RADS existentes en los informes. El sistema de clasificación BI-RADS se divide en 7 categorías que van desde la categoría 0 hasta la 6. En la tabla a continuación, se puede observar la definición de cada una de ellas:

Categoría	Definición
0	Es necesario realizar una evaluación de imagen adicional y/o comparar con mamografías anteriores.
1	Hallazgo negativo, es decir, no hay ninguna anomalía importante que reportar.
2	Hallazgo benigno (no canceroso).
3	Hallazgo probablemente benigno. Se cuenta con una probabilidad mayor al 98 % de ser benigno.
4	Hallazgo sospechoso, se sugiere realizar una biopsia. Esta categoría se divide en tres sub-categorías: -4A: Baja sospecha de malignidad. -4B: Moderada sospecha de malignidad. -4C: Alta sospecha de malignidad.
5	Anomalía altamente sugestiva de malignidad.
6	Diagnóstico maligno comprobado por biopsia.

Tabla 5.2: Categorías BI-RADS.

5.2. Limpieza de los datos

En esta sección se realiza la estandarización de los informes para posteriormente llevar a cabo la extracción de información relevante. Los procesos que se llevan a cabo son los siguientes:

- Eliminación de puntuación: Se eliminan todos los signos de puntuación del texto.
- Texto en minúsculas: Se convierte todo el texto del informe a minúsculas para eliminar las diferencias existentes debido a la manera en que el médico escribió el informe.
- Eliminación de instrucciones: Los informes contienen instrucciones para el proceso de registro del informe por parte del médico, consta de 7 pasos a seguir y quedan registrados textualmente en el informe, por lo que son eliminadas y se extrae un dato que será detallado en la siguiente sección.
- Arreglos de formatos: Se realiza el ajuste necesario al formato de la fecha para que sea identificado como un formato de tiempo y así tener la posibilidad de ordenar los datos a partir de este parámetro. Además, se convierte la fecha de nacimiento en la edad del paciente, ya que es más claro para los médicos utilizarlo de esta manera.

Todos estos pasos detallados anteriormente son aplicados a todos los informes en el dataset, independientemente de su contenido o si contienen instrucciones al inicio, en la mitad o al final del texto. Más allá de lo especificado, no se elimina más información para no perder datos que pueden ser importantes.

A continuación, en la Figura 5.1, se muestra un ejemplo de un informe en su estado inicial, sin ningún tipo de formato y con las instrucciones presentes entre el texto del informe. Además, se puede observar que en este caso de ejemplo se indica el BI-RADS en la primera línea.

Informe de mamografía original

Mamas densas, lo cual puede disminuir la sensibilidad de la mamografía. Sin signos de malignidad. BIRADS 2 ACR C Se complementó con ecografía. Controlar en forma anual. MAMOGRAFIA BILATERAL ECOTOMOGRAFIA MAMARIA PASO 1: Si identifica un estudio con resultado crítico, seleccione la alternativa "SI" desde la siguiente lista. NO PASO 2: Para continuar con el flujo resultado crítico, dirijase a la "pestaña con el nombre del estudio". Una vez que identifique la pestaña del estudio debe: PASO 3: Seleccione una o más patologías críticas de la lista. PASO 4: Seleccione una acción de contacto de la lista: "personal interno" o "con el paciente". PASO 5: Agregue un comentario en el campo de texto libre (Ej.: Nombre o cargo de la persona que contactó). PASO 6: Para guardar el comentario, debe seleccionar: "aceptar y público". PASO 7: Para finalizar, presione aprobar. Paciente referida por control. Antecedente de cirugía mamaria derecha por nódulo benigno. Las mamas son heterogéneamente densas, lo que puede ocultar algunos nódulos pequeños. No se observan nódulos dominantes evidentes. No se observan microcalcificaciones agrupadas que sean sospechosas. Calcificación aislada bilateral. Regiones axilares bajas normales. Ecografía mamaria. Patrón glandular ecogénico heterogéneo. No se observan nódulos sólidos. No se observan quistes en la actualidad. Regiones retroareolares libres. Regiones axilares bajas normales.

Figura 5.1: Ejemplo de informe original.

Luego de pasar por este proceso de limpieza, se puede notar en la Figura 5.2 que el informe queda completamente en minúsculas, sin ningún tipo de símbolo y sin las instrucciones de los pasos a seguir que se encontraban a mitad del reporte.

Informe de mamografía obtenido

mamas densas lo cual puede disminuir la sensibilidad de la mamografía sin signos de malignidad birads 2 acr c se complemento con ecografía controlar en forma anual mamografía bilateral ecotomografía mamaria paciente referida por control antecedente de cirugía mamaria derecha por nódulo benigno las mamas son heterogeneamente densas lo que puede ocultar algunos nodulos pequeños no se observan nodulos dominantes evidentes no se observan microcalcificaciones agrupadas que sean sospechosas calcificacion aislada bilateral regiones axilares bajas normales ecografía mamaria patron glandular ecogenico heterogeneo no se observan nodulos solidos no se observan quistes en la actualidad regiones retroareolares libres regiones axilares bajas normales

Figura 5.2: Ejemplo de informe obtenido.

5.3. Extracción de Información

Una vez contamos con el texto limpio y estandarizado, procedemos a la extracción de información. Como se mencionó en la sección anterior, el texto original incluye unas instrucciones a seguir, estas instrucciones cuentan con una opción la cual solicita indicar si es un caso crítico o no. Esta respuesta se extrae y se añade un nuevo parámetro “Critico”, que almacena los valores “Si” , “No” o “Unknown” en caso de que no exista este dato en el informe. En este caso, no se encontró una relación directa de este parámetro con la clasificación BI-RADS del paciente, por lo que no fue utilizado en este análisis, pero de todas formas se mantuvo su extracción para futuros análisis.

La información más importante extraída del texto corresponde a la clasificación BI-RADS indicado en el informe. Para poder proceder a la extracción de este dato, se consideran múltiples opciones con distintas escrituras o errores de la palabra *birad* para la búsqueda. Una vez encontrado este término, se guarda el número que lo acompaña el cual correspondería a la clasificación. En caso de encontrarse el término más de una vez en el texto se registra el más alto.

La extracción el BI-RADS consta de un manejo de errores, ya que este parámetro solo puede ser clasificado del 0-6. Si se obtiene un número distinto a estos, se reemplaza por -1, que será utilizado como valor por defecto. Este valor BI-RADS debería siempre ser anotado por los médicos que revisan los exámenes, pero la realidad es que no siempre es así, por lo que en algunos informes no existe mención a esta clasificación, y estos casos también son considerados como -1.

Luego de pasar por el proceso de limpieza y extracción, el dataset con el que estamos trabajando obtiene el formato observado en la siguiente tabla:

Nombre	Descripción
Fecha Firma	Fecha en la que el informe fue ingresado al sistema.
Study ID	Identificador del informe.
Type	Tipo de examen de mamografía.
Informe	Texto correspondiente al informe de mamografía.
Fecha Diag	Fecha cuando se realizó diagnóstico en caso de ocurrir.
Diag	Indica si existe un diagnóstico de tumor maligno.
ID	Identificador del paciente.
Edad	Corresponde a la edad del paciente calculada según la fecha de nacimiento y el día actual.
Critico	Valor crítico extraído de las instrucciones.
BIRADS	Valor de BIRADS extraído a partir de los informes.

Tabla 5.3: Dataset luego de limpieza y extracción.

5.4. Asignación de Prioridad

Una vez extraída la clasificación BI-RADS, se le asignará a cada informe un nivel de prioridad para el paciente. Este nivel de prioridad esta determinado según lo siguiente:

Valor BI-RADS	Valor Priorización
4, 5, 6	3
3	2
1,2	1
0	0

Tabla 5.4: Asignación de Prioridad.

Ahora, considerando lo mencionado anteriormente, en algunos casos no se cuenta con el valor BI-RADS en el informe, por lo que se ha implementado un modelo de predicción de prioridad a partir de los informes de textos utilizando SVM:Lineal (para más información respecto a la elección de este modelo para la clasificación consultar el trabajo de Alejandro Viertel)

De esta forma, todos los informes contarán con un parámetro que indicará la prioridad, siendo 3 la prioridad más alta, 1 prioridad más baja y 0 en caso de que se necesitan más exámenes para tomar una decisión. Esto está planteado de esta manera debido a la naturaleza del BI-RADS, en la cual a más alta sea la clasificación, es más probable es la existencia de un tumor maligno. Por ejemplo, el BI-RADS 6 indica la confirmación de esto con una biopsia complementaria. Por lo tanto, es necesario otorgar la prioridad a aquellos pacientes que tengan un mayor riesgo para así ser tratados rápidamente.

A continuación, se puede observar el ejemplo mencionado al inicio en la limpieza. En este caso, se extrajo el BI-RADS 2 y el parámetro crítico “no” a partir del informe. Luego, como se determinó BIRADS 2 la prioridad asignada es 1.

Informe	Critico	BIRADS	Prioridad
mamas densas lo cual puede disminuir la sensibilidad de la mamografia sin signos de malignidad birads 2 acr c se complemento con ecografia controlar en forma anual mamografia bilateral ecotomografia mamaria paciente referida por control antecedente de cirugia mamaria derecha por nodule benigno las mamas son heterogeneamente densas lo que puede ocultar algunos nodulos pequeños no se observan nodulos dominantes evidentes no se observan microcalcificaciones agrupadas que sean sospechosas calcificacion aislada bilateral regiones axilares bajas normales ecografia mamaria patron glandular ecogenico heterogeneo no se observan nodulos solidos no se observan quistes en la actualidad regiones retroareolares libres regiones axilares bajas normales	no	2	1

Tabla 5.5: Ejemplo de informe luego de extraer información y asignar prioridad.

5.5. Integración en plataforma web

Es necesario poder visualizar la información, por lo que se desarrolló una plataforma web donde el médico puede visualizar a sus pacientes en orden de prioridad, además de observar información relevante para la toma de decisiones. La información que se muestra es la siguiente:

- Identificación del paciente, en este caso se puede observar una ID codificada para proteger la identidad de la persona.
- Fecha en la que fue ingresado el informe
- Edad del paciente
- BI-RADS identificado en el informe
- Prioridad asignada al paciente

Por otra parte, se muestra el informe completo del paciente si se extiende la celda para su visualización, permitiendo al médico leer detalladamente el informe si es necesario. Además, se pueden mostrar todos los informes correspondiente al mismo paciente si se selecciona su ID, facilitando observar su historial.

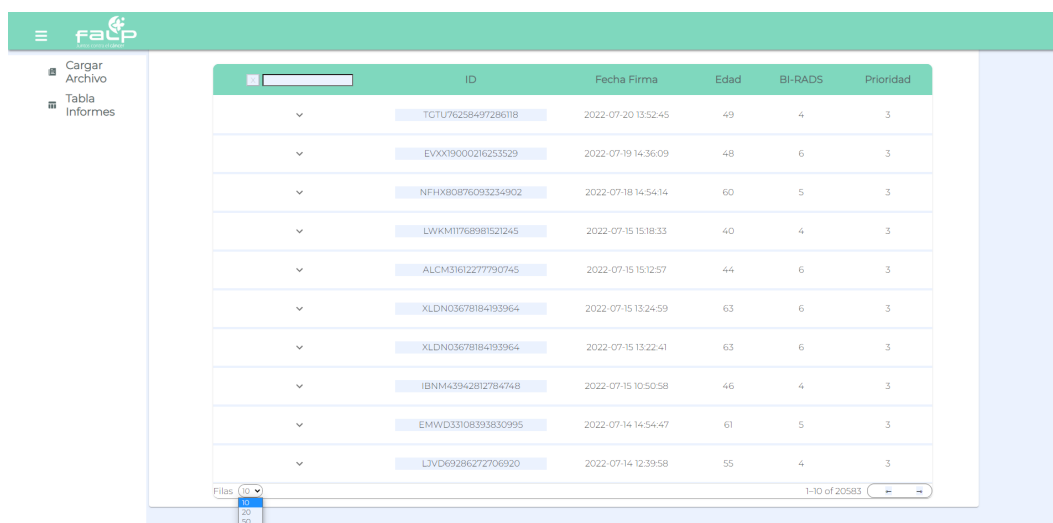
Inicialmente, la plataforma web estaba enfocada únicamente en la visualización de los datos anteriormente procesados. Luego este enfoque cambió, necesitando que la plataforma recibiera los datos y realizara el procesamiento internamente.

Como se puede observar a continuación en la Figura 5.3, es necesario cargar un archivo con el dataset a analizar.



Figura 5.3: Carga de archivos.

Una vez se sube el archivo se realiza automáticamente el proceso detallado anteriormente para el procesamiento de los datos y obtener los resultados deseados, ahora al dirigirse a la pestaña de Tabla Informes se pueden observar los datos de los pacientes ya priorizados como se puede ver en la Figura 5.4.



ID	Fecha Firma	Edad	BI-RADS	Prioridad
TGTU76258497286118	2022-07-20 13:52:45	49	4	3
EVXX9000216253529	2022-07-19 14:36:09	48	6	3
NFH080876093234902	2022-07-18 14:54:14	60	5	3
LWKM1768981521245	2022-07-15 15:18:33	40	4	3
ALCM3161227790745	2022-07-15 15:12:57	44	6	3
XLDN03678184193964	2022-07-15 13:24:59	63	6	3
XLDN03678184193964	2022-07-15 13:22:41	63	6	3
IBNM43942812784748	2022-07-15 10:50:58	46	4	3
EMWD33108393830995	2022-07-14 14:54:47	61	5	3
LVD69286272706920	2022-07-14 12:39:58	55	4	3

Figura 5.4: Carga de archivos.

Para más información respecto al funcionamiento y desarrollo de la plataforma web se recomienda consultar el trabajo realizado por Bastián Cornejo.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusión

Actualmente, en nuestro país no existe ningún sistema estandarizado para mejorar la atención de pacientes con cáncer de mama. Todo el proceso se basa en la asignación de turnos por orden de llegada, lo cual resulta ineficiente y puede perjudicar a las personas con sospecha o diagnóstico confirmado de esta enfermedad. En este contexto, un diagnóstico oportuno puede marcar la diferencia.

La solución desarrollada en este trabajo de memoria, en colaboración con Bastián Cornejo y Alejandro Viertel, proporciona una priorización efectiva para mejorar la continuidad del proceso de atención en los pacientes que más lo necesitan. Utilizamos datos de informes de mamografía, lo que facilita la estandarización de esta solución, dado que este examen es fundamental en todo el proceso y se realiza en prácticamente todas las instituciones de salud relacionadas.

Este trabajo también brinda apoyo a los médicos, permitiéndoles obtener una visión general del estado de sus pacientes y comparar sus diferentes situaciones para tomar decisiones importantes. Además, destaca información relevante como el BI-RADS, eliminando la necesidad de revisar constantemente los informes de los pacientes para obtener esta información.

Esta herramienta resulta sumamente útil en el campo de la salud y es altamente accesible para instituciones de atención médica en general, dado que se basa en información fácilmente obtenible, como los informes de mamografía. Sin embargo, es importante señalar que existe margen para mejoras que permitirían ofrecer resultados aún más precisos. Estas mejoras podrían incluir la consideración de información adicional, como datos del paciente (sexo, antecedentes familiares, reincidencias) o la incorporación de exámenes complementarios.

6.2. Trabajo Futuro

En cuanto al trabajo futuro, hay diversas áreas a considerar para mejorar el rendimiento y obtener resultados más sólidos. Una de ellas es la posibilidad de incorporar más información en el proceso de priorización. A pesar de que la mamografía sirve como base, sería interesante evaluar la inclusión de otros tipos de exámenes complementarios para lograr un diagnóstico aún más preciso de los pacientes.

En relación a la extracción de información, una tarea futura pendiente es la adecuada aplicación de MedspaCy en español. Esta herramienta se reveló como sumamente útil para obtener información adicional a partir del texto de los informes, lo que podría fortalecer el análisis realizado para la predicción de prioridades.

Siguiendo con este punto, consideramos esencial explorar otras alternativas a los clasificadores utilizados en esta ocasión. La restricción principal en esta ocasión fue la disponibilidad de datos, pero con un conjunto de datos de mayor calidad, las oportunidades para buscar resultados más efectivos se ampliarían. Esto incluiría la evaluación de tecnologías como otros modelos de aprendizaje automático o redes neuronales.

Bibliografía

- [1] Organización Panamericana de la Salud. Cáncer de mama.
- [2] Organización Mundial de la Salud. Cáncer de mama, 2023.
- [3] Organización Mundial de la Salud. Cáncer, AÑO.
- [4] Tomorrow: Global cancer observatory.
- [5] Globocan: Cancer today. <https://gco.iarc.fr/today>, 2020.
- [6] Neeraj Agarwal. The ultimate guide to different word embedding techniques in nlp.
- [7] Apache cTAKESTM.
- [8] Sujan Subendran. Using natural language processing to detect breast cancer recurrence in clinical notes: A hierarchical machine learning approach. *UWSpace*, 2021.
- [9] Mostafa Boroumandzadeh, Elham Parvinnia. Automated classification of bi-rads in textual mammography reports. *Turkish Journal of Electrical Engineering Computer Sciences*, 2021.
- [10] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv*, 2020.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2019.
- [12] Rani Horev. Bert explained: State of the art language model for nlp.
- [13] G. McGoldrick, Y. Cao and S. Prince. Understanding xlnet.
- [14] Sicheng Zhou, Liwei Wang, Nan Wang, Hongfang Liu and Rui Zhang. Cancerbert: a bert model for extracting breast cancer phenotypes from electronic health records. *arXiv*, 2022.
- [15] Kexin Huang, Abhishek Singh, Sitong Chen, Edward T. Moseley, Chih-ying Deng, Naomi George and Charlotta Lindvall. Bert explained: State of the art language model for nlp. *arXiv*, 2019.

- [16] C. Friedman, P. Alderson, J. Austin, J. Cimino, S. Johnson. A general natural-language text processor for clinical radiology. 1994.
- [17] National Center for Biomedical Computing I2B2. Health information text extraction.
- [18] Cunningham H, Humphreys K, Gaizauskas R, Wilks Y. General architecture for text engineering.
- [19] Hossein Naseri, Kamran Kafi, Sonia Skamene, Marwan Tolba, Mame Daro Faye, Paul Ramia, Julia Khriuguian, and John Kildea. Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases. *Journal of Biomedical Informatics*, 120:103864, 2021.
- [20] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl_1):D267–D270, 01 2004.
- [21] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 05 2010.
- [22] spaCy 3: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.
- [23] A.E. Johnson, T.J. Pollard, L. Shen, L.W.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark. MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, pages 1–9, 2016.
- [24] Daniel T. Heinze, Mark L. Morsch, Brian C. Potter, and Jr Sheffer, Ronald E. Medical i2b2 NLP Smoking Challenge: The A-Life System Architecture and Methodology. *Journal of the American Medical Informatics Association*, 15(1):40–43, 01 2008.
- [25] Charles E. Kahn Jr, Linda M. Roberts, Katherine A. Shaffer, Peter Haddawy. Construction of a Bayesian network for mammographic diagnosis of breast cancer. *Computers in Biology and Medicine*, 27(1):19–29, 1997.
- [26] Omar D. Castrillón, Eduardo Castaño, Luis F. Castillo. Sistema Predictivo Bayesiano para Detección del Cáncer de Mama. *Información Tecnológica*, 29(3):257–270, 2018.
- [27] Bird S, Klein E, Loper E. *Natural Language Processing with Python*. O’Reilly Media Inc, 2009.
- [28] Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings 2021*, (in press, n.d.).

- [29] Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamára L Box, Scott L DuVall, and Olga V Patterson. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. *AMIA Annu. Symp. Proc.*, 2021:438–447, 2021.
- [30] John Snow Labs: Nlp & ai in healthcare.
- [31] Spark NLP: State of the art natural language processing.
- [32] Detect tumor morphology in spanish texts.
- [33] A Miranda-Escalada, E Farré, and M Krallinger. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*, 2020.