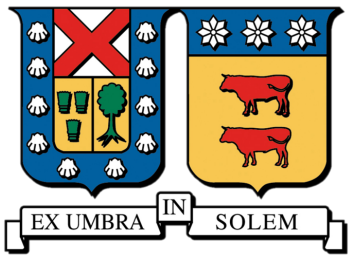
Universidad Técnica Federico Santa María

Departamento de Electrónica



Desarrollo de un Modelo Predictivo para Identificar Pacientes Derivables a Asesoramiento Genético Oncológico mediante el Procesamiento del Lenguaje Natural

AXEL MICHEL PÉREZ ZAMORA

Memoria para optar al Título de Ingeniero Civil Telemático

Profesor Guía:
Mauricio Araya

Profesor Correferente:
Mohamed Abdelhamid

Diciembre 2024

Resumen

Este informe presenta el desarrollo de un modelo predictivo basado en procesamiento del lenguaje natural (NLP) para identificar pacientes que deben ser derivados a Asesoramiento Genético Oncológico (AGO). Realizado en colaboración con la Fundación Arturo López Pérez (FALP), el estudio aborda la creciente necesidad de optimizar la detección de riesgos genéticos asociados al cáncer. La investigación se estructura en varios módulos: preprocesamiento de datos clínicos, análisis de datos estructurados y no estructurados, y entrenamiento del modelo utilizando BioBERT, un modelo avanzado de NLP especializado en el ámbito biomédico. Los resultados demostraron una precisión del 75.59 %, superando significativamente a modelos no específicos como ChatGPT. Además, se implementaron herramientas visuales en Power BI para facilitar la interpretación de los datos y resultados. Este modelo representa un avance significativo hacia una atención más personalizada y eficaz, reduciendo el margen de error en la derivación clínica y optimizando los recursos de diagnóstico y tratamiento.

Abstract

This report presents the development of a predictive model based on natural language processing (NLP) to identify patients who should be referred to Oncological Genetic Counseling (AGO). Conducted in collaboration with the Arturo López Pérez Foundation (FALP), the study addresses the growing need to optimize the detection of genetic risks associated with cancer. The research is structured in several modules: preprocessing of clinical data, analysis of structured and unstructured data, and model training using BioBERT, an advanced NLP model specialized in the biomedical field. The results showed an accuracy of 75.59 %, significantly outperforming non-specific models such as ChatGPT. In addition, visual tools were implemented in Power BI to facilitate the interpretation of the data and results. This model represents a significant advance towards more personalized and efficient care, reducing the margin of error in clinical referral and optimizing diagnostic and treatment resources.

Agradecimientos

A mis padres, por entregarme todo el apoyo necesario para llegar a este punto de mi carrera.
A mi hija, por ser el pilar fundamental de mi ser, todo lo que consiga en la vida será para y por ti.
A mis hermanos, por hacer mis días más alegres.
A mis compañeros, por estar cuando necesitaba ayuda, orientación y unas buenas risas.

Tabla de Contenidos

1. Introducción. **6**
1.1. Contexto. . . . 6
1.2. Definición del problema. . . . 7
1.3. Actores claves. . . . 8
1.4. Datos disponibles. . . . 8
1.5. Tecnologías. . . . 9
1.6. Propuesta de solución. . . . 9
1.7. Objetivos. . . . 10
1.8. Estructura de la memoria. . . . 10

2. Estado del Arte. **12**

3. Desarrollo de la solución. **15**
3.1. Módulo de Preprocesamiento y Preparación de Datos Clínicos. . . . 15
3.1.1. Estructuración de datos de entrada. . . . 15
3.1.2. Oncotext 18
3.1.3. Estandarización de datos no estructurados 19
3.2. Módulo de Análisis de datos estructurados. . . . 22
3.3. Módulo de preparación de datos. . . . 23
3.4. Módulo de entrenamiento. . . . 24

4. Resultados y Discusión **27**
4.1. Resultados 27
4.2. Consolidación de resultados en Power BI 30
4.3. Comparación del modelo entrenado v/s modelo ChatGPT 33

5. Conclusiones **38**
5.1. Conclusiones 38
5.2. Trabajos futuros 39

Referencias **40**

Índice de tablas

1. Hiperparámetros usados. (3.1)
2. Definición de los hiperparámetros. (3.2)
3. Resultados de métricas para ambos modelos. (4.1)

Índice de figuras

1. Flujo del programa AGO. (1.1)
2. Cantidad de defunciones asociadas al cáncer de mama en Chile entre los años 2002-2021. (1.2)
3. Cantidad de defunciones asociadas al cáncer de colon en Chile entre los años 2002- 2021. (1.3)
4. Ilustración sobre el proceso *Pattern Matching* diseñado para extraer fechas de informes. (2.1)
5. Ejemplo de uso de reconocimiento de entidades nombradas (NER). (2.2)
6. Estructura archivo Datos_anonimizados.xlsx. (3.1)
7. Estructura archivo Solicitud_ago_1.xlsx. (3.2)
8. Estructura archivo Solicitud_ago_2.xlsx. (3.3)
9. Estructura archivo Datos_consolidados.xlsx. (3.4)
10. Vista de un término en Oncotext. (3.5)
11. Resultado de búsqueda. (3.6)
12. Flujo de estandarización (3.7)
13. Contenido del diccionario base de Oncotext. (3.8)
14. Diferencia de términos. (3.9)
15. Resultado de estandarización. (3.10)
16. Resultados de análisis de datos estructurados. (3.11)
17. Gráfica de evoluciones y biopsias. (3.12)
18. Dataframe obtenido luego de la preparación de datos. (3.13)
19. Figura BioBERT. (3.14)
20. Metodología entrenamiento. (3.15)
21. Entrenamiento del modelo con BioBERT. (3.16)
22. Curvas de Aprendizaje y Validación del entrenamiento. (4.1)
23. Cuadrantes de una matriz de confusión. (4.2)
24. Matriz de confusión del entrenamiento. (4.3)
25. Listado de Banderas Rojas. (4.4)
26. Panel de datos para el análisis. (4.5)

27. Distribución de informes por pacientes. (4.6)
28. Panel de resultados del modelo. (4.7)
29. Instrucción para darle el contexto a ChatGPT.(4.8)
30. Primera prueba con ChatGPT. (4.9)
31. Instrucción a ChatGPT para que reciba un archivo. (4.10)
32. Archivo analizado por parte de ChatGPT. (4.11)
33. Matriz de confusión generada para resultados de ChatGPT. (4.12)

Glosario

- FALP: Fundación Arturo López Pérez.
- AGO: Asesoramiento genético oncológico.
- Fish (análisis por hibridación fluorescente in situ): Técnica de laboratorio que se usa para detectar y localizar una secuencia de ADN específica en un cromosoma.
- Evolución: Informes de texto en formato libre redactados por médicos durante las consultas con los pacientes.
- Biopsia: Procedimiento médico que consiste en la extracción de una pequeña muestra de tejido o células de un organismo para su posterior examen bajo un microscopio.
- IA (Inteligencia Artificial): Campo de la informática que busca dotar a las máquinas con habilidades similares a la inteligencia humana para realizar tareas específicas.
- NLP (Procesamiento del Lenguaje Natural): Campo de IA que permite a las máquinas entender y generar lenguaje humano.
- Pattern Matching: Identificación y extracción de patrones específicos o estructuras lingüísticas predefinidas dentro de un conjunto de datos de texto.
- BERT (Bidirectional Encoder Representation from Transformer): Modelo de lenguaje basado en Transformer que mejora la comprensión contextual de palabras en oraciones.
- Word Embeddings: Técnica que asigna representaciones vectoriales a palabras para procesamiento eficiente de texto en algoritmos.
- Redes Neuronales: Estructuras inspiradas en el cerebro humano, usadas en NLP para aprender de datos y reconocer patrones en texto.
- Decoder: Componente de Transformer que genera la salida en tareas de procesamiento del lenguaje natural.
- Encoder: Parte de Transformer que procesa la entrada y crea representaciones vectoriales en tareas de NLP.
- Transformer: Arquitectura de red neuronal clave en NLP que utiliza mecanismos de atención.
- EHR (Registro Electrónico de Salud): Sistema digital que almacena información médica electrónicamente.
- Extracción de Relaciones (RE): Técnica que identifica relaciones entre entidades en texto.
- Bi-directional LSTM: Tipo de red neuronal que procesa información en ambas direcciones, capturando dependencias a largo plazo en datos secuenciales.
- CRF (Conditional Random Field): Modelo estadístico en NLP para etiquetado secuencial de datos, considerando dependencias entre etiquetas.
- CIE10: Acrónimo de Clasificación internacional de enfermedades, 10.ª edición.

1. Introducción.

1.1. Contexto.

Actualmente en Chile, el cáncer es considerado una de las principales causas de muerte, por lo que es de gran importancia estudiar y analizar sus características para lograr mejorías y avances en su detección temprana, Esto generaría una gran diferencia a la hora de tratar la enfermedad antes de que pueda extenderse a otras partes del cuerpo y, además, podría mejorar considerablemente las probabilidades de superar la enfermedad.

Es por ello que nuestro país cuenta con diversas instituciones y fundaciones dedicadas a brindar apoyo a pacientes con cáncer, tanto en la etapa de detección como en la de tratamiento. Una de ellas es la Fundación Arturo López Pérez (de aquí en adelante, FALP), la cual es una fundación sin fines de lucro que lleva ayudando a miles de personas en Chile desde el año 1954. Dentro de las principales funciones de la FALP, destacan los programas de educación, prevención y detección precoz del cáncer, entre los más importantes se puede encontrar el programa de asesoramiento genético oncológico (de aquí en adelante, AGO), el cual a través de diferentes estudios, permite determinar si existe un componente hereditario del cáncer en el paciente. Lo interesante de esto es que el programa AGO no solo es accesible para pacientes con cáncer ya diagnosticado, sino que puede ingresar cualquier persona, independiente de su estado de salud.

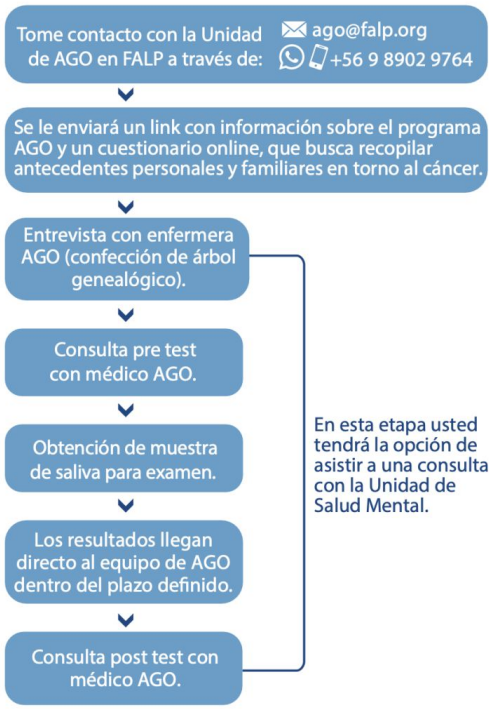


Figura 1.1: Flujo del programa AGO.

En el caso de que un paciente obtenga un resultado positivo al finalizar el programa, significa que posee una variante (mutación) que aumenta el riesgo de tener ciertos tipos de cáncer en la vida.

Esto no significa que necesariamente desarrollará cáncer en alguno de los órganos en riesgo, pero sí presenta una mayor probabilidad en comparación con personas que no tienen una variante genética. Además, permite informar oportunamente al círculo familiar del paciente para dar la opción de estudiar el riesgo de portar la misma variante, con la finalidad de que puedan tomar medidas preventivas.

1.2. Definición del problema.

Si bien existe la posibilidad de que un paciente acceda de forma voluntaria al programa AGO, dentro de la FALP existen otras vías de acceso, una de las más importantes es la derivación realizada a través de las consultas realizadas por los médicos oncólogos, el foco del análisis de este medio de ingreso es de gran importancia, debido a que un buen estudio de los pacientes puede llevar a una correcta y oportuna derivación a AGO, garantizando así la cobertura máxima de las posibles formas de detección temprana del cáncer.

Dentro de los tipos de cáncer más demandados en la FALP, y los cuales presentan más riesgos para el paciente, ya sea en términos de gravedad como de mortalidad, es posible encontrar el cáncer de mama y el cáncer de colon, dividiéndose este último en dos tipos: polipósico y no polipósico. Para cada uno se tiene un listado de banderas rojas específicas otorgadas por la FALP, las cuales indican las condiciones requeridas para que un paciente deba ser derivado a AGO. Si bien este listado es revisado y analizado por los médicos, existen muchos factores humanos que pueden dar resultados erróneos respecto a la decisión de enviar un paciente o no a AGO, además de que sus análisis representa un costo de tiempo no menor.

Como se puede ver en las figuras 1.2 y 1.3 extraídas desde el Departamento de Estadísticas e Información de Salud del Ministerio de Salud de Chile[1], las mortalidades del cáncer de mama y de colon han estado siempre al alza y se espera que en los próximos años la tendencia continúe.

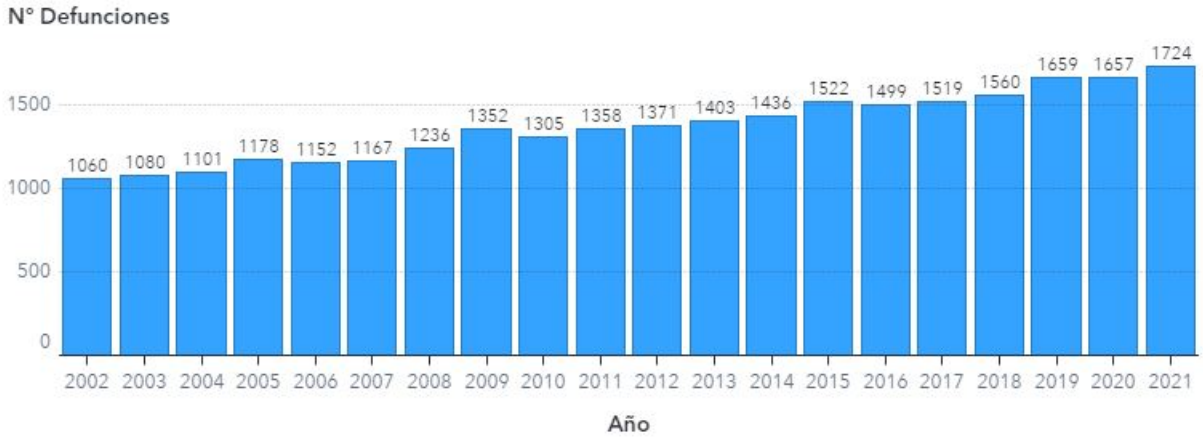


Figura 1.2: Cantidad de defunciones asociadas al cáncer de mama en Chile entre los años 2002-2021 [1].

[1] https://deis.minsal.cl/

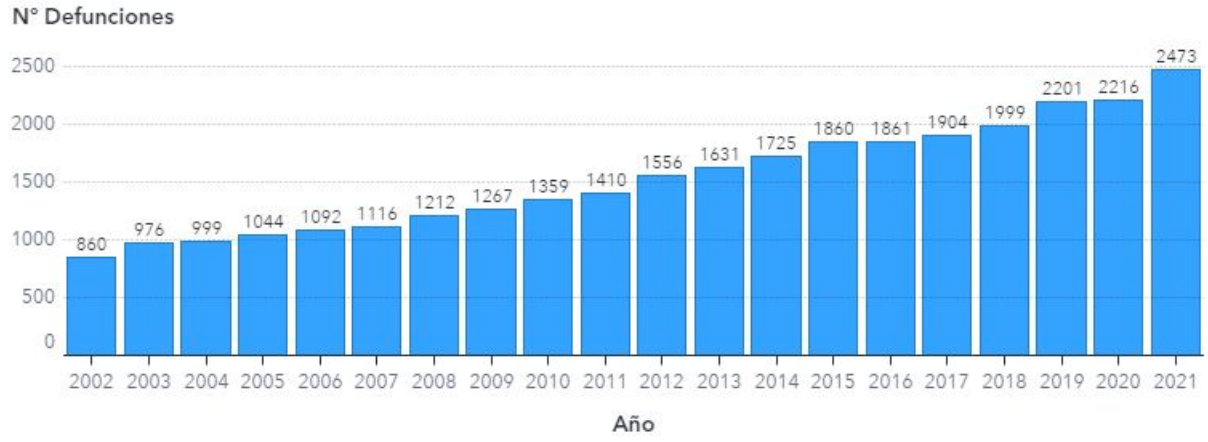


Figura 1.3: Cantidad de defunciones asociadas al cáncer de colon en Chile entre los años 2002-2021 [1].

Es por esto que la FALP, en su necesidad de modernizar los procesos internos, necesita buscar una solución para que estas derivaciones se hagan de forma eficiente, es decir, que los pacientes derivados efectivamente cumplan al menos una bandera roja, y que la detección de estas banderas rojas se haga de forma automática, ayudando así al equipo de asesoramiento oncológico. Debido a esto levantaron la siguiente inquietud: ¿Cómo, con los datos que tenemos disponibles, podemos crear un modelo que prediga cuáles pacientes de cáncer de mama y cáncer colon deben ser derivados a asesoramiento genético oncológico (AGO)?.

1.3. Actores claves.

Dentro del proceso que implican las derivaciones a AGO, existen ciertos actores que cumplen un rol fundamental:

- Paciente: Persona natural que se atiende en FALP y que proporciona datos e informaciones referentes a su situación médica e historial clínico, lo cual es analizado por los profesionales de la institución.
- Médico Oncólogo: Es quien atiende al paciente, revisa su ficha clínica, levanta información, toma nota y determina si debe ser derivado a asesoramiento genético oncológico (AGO).
- Personal de equipo AGO: Revisa con quien se atendió el paciente y las observaciones de dicha atención, coordina exámenes y posteriormente entrega resultados.

1.4. Datos disponibles.

Con lo que respecta a la información disponible para la construcción la solución, FALP dejó a disposición del equipo los siguientes datos:

- Datos sobre los pacientes (Nombre, fecha de nacimiento, sexo, tipo de cáncer, etc.)
- Informes clínicos (Evoluciones, informes de exámenes, informes de biopsias, etc.)

- Listado de banderas rojas para cáncer de mama y cáncer de colon polipósico y no polipósico.
- Informaciones acerca del programa de AGO.

La principal fuente de información que fue utilizada son los informes clínicos, especialmente de las evoluciones, las cuales consisten en reportes tipo texto libre escrito por los médicos durante las consultas médicas con los pacientes.

1.5. Tecnologías.

Python y Jupyter Notebook son herramientas fundamentales y altamente eficaces para la elaboración de modelos de predicción, gracias a su versatilidad, facilidad de uso y la vasta colección de bibliotecas especializadas que ofrecen. Estas tecnologías se destacaron como las principales en la construcción del modelo debido a sus múltiples ventajas, las cuales serán explicadas a continuación.

Python [2], un lenguaje de programación ampliamente utilizado en la ciencia de datos y el aprendizaje automático, se distingue por su sintaxis clara y su capacidad para facilitar una programación rápida y eficiente. Su popularidad en la comunidad de análisis de datos se debe a su amplio ecosistema de bibliotecas y frameworks, como NumPy, pandas, scikit-learn y TensorFlow, que proporcionan herramientas robustas para el manejo de datos, la creación de modelos y el análisis avanzado.

Por otro lado, Jupyter Notebook [3] ofrece un entorno interactivo y visual que complementa a Python de manera excepcional. Esta herramienta permite a los usuarios explorar datos, desarrollar modelos y presentar resultados de forma integral y accesible. Los notebooks de Jupyter combinan código Python con texto explicativo, gráficos y visualizaciones interactivas, lo que facilita la documentación del proceso y la comunicación de hallazgos. Esta capacidad de integrar y presentar información en un solo documento facilita la colaboración y la replicación de análisis, haciendo de Jupyter Notebook una opción preferida para la investigación y el desarrollo en ciencia de datos.

1.6. Propuesta de solución.

Ya teniendo una noción sobre cuál es el problema y que información hay disponible, es posible dar una idea general de como se comenzó a construir la solución. Antes de entrar en detalle, a continuación se explicarán algunos conceptos para comprender de mejor manera las decisiones tomadas.

En los últimos años, ha surgido un notable progreso en el ámbito de la Inteligencia Artificial (IA), marcando un hito importante en la evolución de la interacción entre humanos y máquinas. Dentro de este panorama, el Procesamiento del Lenguaje Natural (de aquí en adelante, NLP) se ha erigido como una disciplina esencial.

La Inteligencia Artificial, en términos generales, busca conferir a las máquinas la capacidad de realizar tareas que, históricamente, eran exclusivas de la inteligencia humana. Este avance ha sido posible gracias al desarrollo de algoritmos más sofisticados, al incremento en la capacidad de procesamiento y al acceso a grandes conjuntos de datos.

El NLP se enfoca específicamente en la comprensión y manipulación del lenguaje humano por parte de las máquinas. Su objetivo principal es permitir que las computadoras interpreten, generen y respondan al lenguaje natural de manera similar a como lo haría un ser humano. Este campo abarca diversas áreas, desde el reconocimiento del habla hasta la traducción automática, pasando por la generación de lenguaje y el análisis de sentimientos.

Una vez entendido el panorama, se puede ir a la explicación de la propuesta de solución, la cual viene dada por la creación de un modelo NLP que, mediante los datos de entrada (información) de un paciente, determine si este debe ser derivado a AGO en función de la detección de banderas rojas asociadas a su tipo de cáncer. Dicho modelo será utilizado por médicos oncólogos de la FALP y personal del equipo AGO para la atención de pacientes dentro de la fundación, considerando los datos disponibles del usuario, los cuales pueden ser de orígenes internos o externos.

El trabajo se enfocó en la etapa de derivación de un paciente por medio de la información recopilada en las consultas médicas, tanto por los datos entregados por el paciente como los datos de informes y exámenes médicos, esto representa un gran desafío, el cual formalmente se presenta como: Desarrollo de un Modelo Predictivo para Identificar Pacientes Derivables a Asesoramiento Genético Oncológico mediante el Procesamiento del Lenguaje Natural. Tomando en cuenta la información de pacientes diagnosticados con cáncer de mama [4] y de aquellos diagnosticados con cáncer de colon, que puede ser de tipo polipósico [5] o no polipósico [6].

1.7. Objetivos.

El propósito principal del trabajo es desarrollar un modelo con la capacidad de analizar información médica y que, a través de modelos predictivos, advierta cuando exista presencia de banderas rojas para que un paciente sea derivado a AGO.

Este objetivo se trabaja en base a la realización de los siguientes objetivos específicos:

- Elaboración de un plan de acción adecuado para abordar el desafío planteado, tras realizar un análisis previo exhaustivo del problema.
- Búsqueda de metodologías de procesamiento de lenguaje natural para el trabajo con datos no estructurados.
- Análisis de modelos de predicción, vinculando su uso con las técnicas de NLP.
- Desarrollo de un sistema de predicción de banderas rojas, utilizando los datos disponibles por la FALP.

1.8. Estructura de la memoria.

En el capítulo 2, Estado del Arte y de la Técnica, se hizo una exploración de métodos clave para procesar datos clínicos no estructurados, desde técnicas fundamentales como el *Pattern Matching* en NLP hasta enfoques avanzados como el uso de *Word Embeddings*, con énfasis en BERT.
Luego, en el capítulo 3, Desarrollo de la solución, se explica el trabajo desarrollado mediante la separación por módulos, los cuales son:

- **Módulo de Preprocesamiento y Preparación de Datos Clínicos:** Se encarga del preprocesamiento de los datos del paciente, los cuales se presentan como el origen de datos ingresado en la entrada o input. Estos se clasifican como estructurados o no estructurados. (Sección 3.1)
- **Módulo de Análisis de datos estructurados:** Realiza un análisis previo de los datos para obtener métricas que se utilizarán en los siguientes módulos. (Sección 3.2)
- **Módulo de preparación de datos:** Su finalidad es ordenar y filtrar los datos antes de entrenar el modelo, para tener la mayor tasa de precisión posible. (Sección 3.3)
- **Módulo de entrenamiento:** Su finalidad es entrenar un modelo NLP con los datos anteriormente trabajados. El modelo entrenado debe ser capaz de entregar resultados acordes al desafío. (Sección 3.4)

2. Estado del Arte.

Una vasta cantidad de datos clínicos de los pacientes viene en formato de texto libre, es decir, en forma no estructurada. Esto presenta un gran desafío en el área de investigación debido a varias razones. En primer lugar, el texto libre carece de una organización estandarizada, lo que hace que sea difícil para los sistemas automatizados interpretar y analizar la información de manera consistente. La variabilidad en el lenguaje, la terminología médica y el estilo de redacción pueden variar significativamente entre diferentes registros, lo que complica la extracción de datos relevantes.

Además, el procesamiento de texto libre requiere técnicas avanzadas de procesamiento del lenguaje natural (NLP) y aprendizaje automático para convertir la información no estructurada en un formato estructurado y utilizable. Estas técnicas deben ser capaces de identificar y extraer datos clave, como síntomas, diagnósticos y tratamientos, para luego ser normalizados e integrados en modelos analíticos. El desafío se magnifica cuando se deben manejar grandes volúmenes de datos provenientes de múltiples fuentes, cada una con sus propias particularidades y formatos.

Por lo tanto, la necesidad de desarrollar y aplicar metodologías precisas para el procesamiento y análisis de texto libre es fundamental para mejorar la calidad y la utilidad de los datos clínicos. La correcta implementación de estas técnicas puede llevar a una mejor comprensión de las condiciones de los pacientes, una mejora en la toma de decisiones clínicas y avances significativos en la investigación médica.

Una de las técnicas más simples y fundamentales del NLP es el *Pattern Matching*. Este consiste en un patrón de secuencia de caracteres que puede compararse carácter a carácter en un texto dado. Por ejemplo, el patrón “he” se puede encontrar dos veces en la secuencia de texto “He said hello”. Para aumentar su generalización, se hace uso de otra técnica de NLP, como el uso de expresiones regulares [7]. Estas técnicas podrían ser de suma utilidad para reconocer un patrón dentro de las descripciones del médico oncólogo en las evoluciones del paciente a tratar, ya que se está ante un problema de poder estructurar el texto libre creado por el médico.

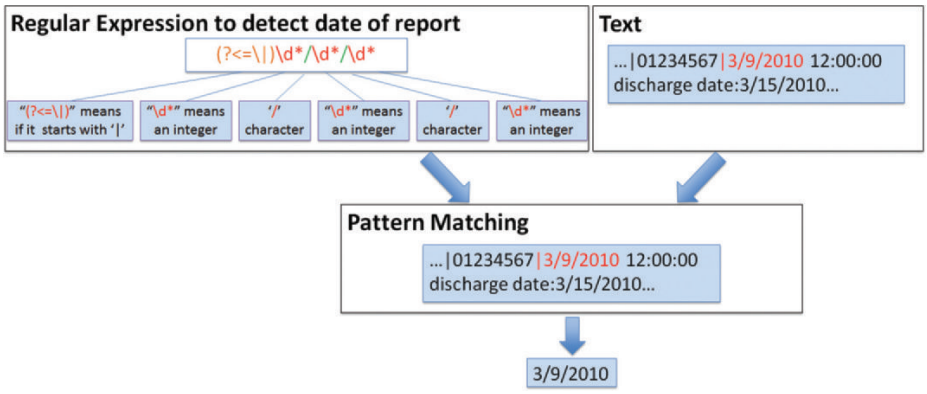


Figura 2.1: Ilustración sobre el proceso *Pattern Matching* [7] diseñado para extraer fechas de informes.

Otra de las metodologías investigadas y de interés es el uso de *Word Embeddings* a través de redes neuronales. Una de las técnicas más destacadas dentro de esta metodología es la utilización del modelo BERT [8] (Bidirectional Encoder Representation from Transformer). El modelo BERT hace uso de *Transformer*, este es un mecanismo de atención que aprende las relaciones contextuales de las palabras o sub palabras en un texto. En su forma básica, Transformer incluye dos mecanismos separados: Decoder y Encoder. Encoder lee el texto de entrada y el Decoder produce la predicción de la tarea. Se le denomina bidireccional, dado que BERT analiza la secuencia de palabras en su totalidad, considerando tanto lo que está a la izquierda como a la derecha de cada palabra, a diferencia de otros modelos unidireccionales que leen los textos de forma secuencial. Al ser un modelo bidireccional permite entender de mejor manera el contexto de la palabra en función a todo lo que la rodea.

La aplicación de BERT en el área de la salud en el contexto de análisis de textos no estructurado ha sido material de estudio por los investigadores. Un ejemplo de caso de uso de BERT es la investigación que hizo la Universidad de Minnesota [9], en donde entrenan el modelo manera de poder extraer información relacionada con el cáncer de mama de un EHR (Electronic Health Record).

En el estudio realizado por Shaina Raza y Brian Schwartz [10], se muestra que utilizando NLP pueden detectar en reportes características de riesgo y síntomas del Covid 19. Para esto presentan dos módulos, el primero consiste en un módulo de Reconocimiento de Entidades (NER) con el fin de producir entidades con los nombres, y un segundo módulo de Extracción de Relaciones (RE) para definir la relación entre entidades. Para la solución construida, fue de interés el primer módulo, ya que este hace uso de BERT, lo que se quiere obtener de la secuencias de textos es una etiqueta para un conjunto de relaciones predefinidas, llamadas entidades. Para la creación del primer módulo se inspira al modelo de *bi-directional long short-term memory (LSTM)* con una capa de *conditional random field (CRF)* [11], además de añadida una capa de Transformers para variar el modelo.

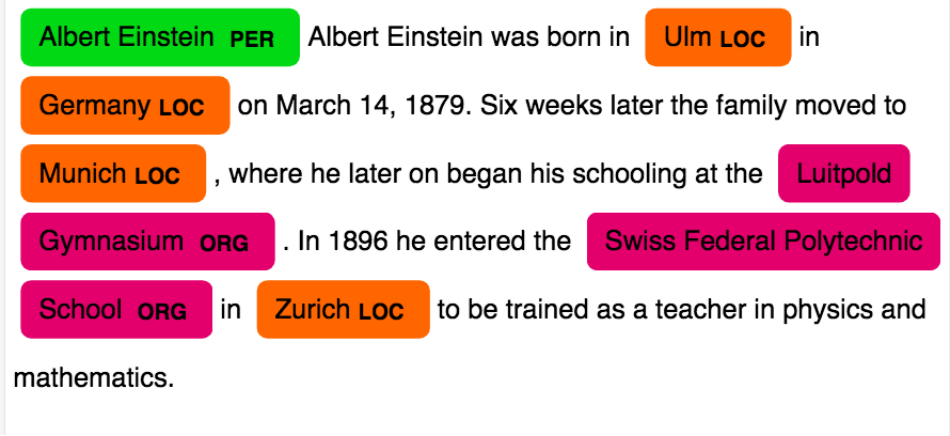


Figura 2.2: Ejemplo de uso de reconocimiento de entidades nombradas (NER).

Una de las herramientas más destacadas actualmente en el campo del NLP es la biblioteca de Python *spaCy*[12]. Esta librería de código abierto se utiliza para una variedad de tareas en NLP, incluyendo la extracción de información, el procesamiento del lenguaje natural, la identificación

de entidades y el preprocesamiento de textos para modelos de aprendizaje automático.

Otra biblioteca popular en Python es *textBlob*[13]. Su simplicidad y la facilidad para realizar tareas comunes de procesamiento de texto la convierten en una opción popular. Aunque es una herramienta útil y accesible, tiene limitaciones en cuanto a escalabilidad cuando se trata de grandes volúmenes de texto, y también enfrenta dificultades con tareas muy complejas.

Dada la necesidad de querer procesar texto proveniente de reportes médicos, dentro de las nuevas bibliotecas para suplir tal necesidad, destaca *medspaCy*[14]. Esta biblioteca esta hecha en base a la recién mencionada *spaCy*, y ofrece una gran capacidad de procesamiento con enfoque en el análisis de reportes médicos. Los componentes incluidos en *medspaCy* ofrecen tanto una inicialización predeterminada como un gran número de parámetros de personalización opcionales. Esto permite que los componentes se configuren, aprendan y utilicen rápidamente para desarrollar prototipos, al mismo tiempo que son totalmente personalizables para necesidades más específicas. Los recursos, como conjuntos de reglas curadas o bases de conocimiento, pueden compartirse con la comunidad y utilizarse en diferentes proyectos. Uno de los defectos de *medspaCy* es que solo funciona con reportes médicos en el idioma ingles, pero esto se puede solucionar dada la derivación que tiene esta herramienta con *spaCy*.

3. Desarrollo de la solución.

Este capítulo detalla el trabajo realizado en los diferentes módulos que constituyen el desarrollo integral de un modelo de predicción diseñado para identificar pacientes que requieren derivación a AGO. El enfoque principal radica en la utilización de los datos clínicos proporcionados por FALP, los cuales sirvieron como base para el diseño, implementación y evaluación del modelo.

3.1. Módulo de Preprocesamiento y Preparación de Datos Clínicos.

Los datos entregados por la FALP se resumen en archivos en formato xlsx (Excel) con información acerca de los pacientes, evoluciones y reportes de exámenes, para poder usar esta data en modelos NLP, se debe estructurar y definir una disposición uniforme de los datos, en palabras simples, es necesario recolectar y ordenar la data que tendrá de entrada el modelo a entrenar.

3.1.1. Estructuración de datos de entrada.

Para dar una idea general de que información se tiene, se detallarán los campos contenidos por cada uno de los tres archivos disponibles.

	Nombre hoja	Contenido
Datos_anonimizados	Evoluciones 2022	Listado de todas las evoluciones realizadas el año 2022, cada registro contiene el id del paciente, fecha evolución, CIE10, evolución y diagnostico
	Primera evolución	Listado de las primeras evoluciones realizadas a los pacientes, cada registro contiene el id del paciente, fecha evolución, CIE10 y evolución.
	Fish	Listado de los exámenes Fish realizadas a los pacientes, cada registro contiene el id del paciente, fecha de realización y resultado.
	Biopsias	Listado de Biopsias realizadas a los pacientes, cada registro contiene el id del paciente, fecha de realización y diversos resultado.
	Pacientes	Listado de pacientes, cada registro contiene el id del paciente, fecha de nacimiento, sexo, CIE10 y diagnostico.

Figura 3.1: Estructura archivo Datos_anonimizados.xlsx.

	Nombre hoja	Contenido
Solicitud_ago_1	evoluciones	Listado evoluciones de pacientes que fueron derivados a AGO los años 2020-2021. Cada registro contiene el id del paciente, fecha de evolución, CIE10, evolución y diagnostico.
	biopsias	Listado de Biopsias de pacientes que fueron derivados a AGO los años 2020-2021. Cada registro contiene el id del paciente, fecha de realización y diferentes resultados.

Figura 3.2: Estructura archivo Solicitud_ago_1.xlsx.

	Nombre hoja	Contenido
Solicitud_ago_2	solicitud_interconsulta	Listado de pacientes los que fueron derivados a AGO los años 2022-2023. cada registro contiene el id del paciente, fecha de solicitud, diagnostico y comentario.
	prestaciones	Listado de pacientes que fueron atendidos por AGO, luego de su derivación

Figura 3.3: Estructura archivo Solicitud_ago_2.xlsx.

El archivo principal que fue usado durante el transcurso del desarrollo es Datos_anonimizados.xlsx, el cual a pesar de tener diversos datos con respecto al paciente, contiene ciertas variables que no están asociadas directamente a las banderas rojas, por otro lado, se dejó fuera cualquier registro que tenga al menos un dato en blanco.
El contenido se puede resumir con la siguiente estructura:

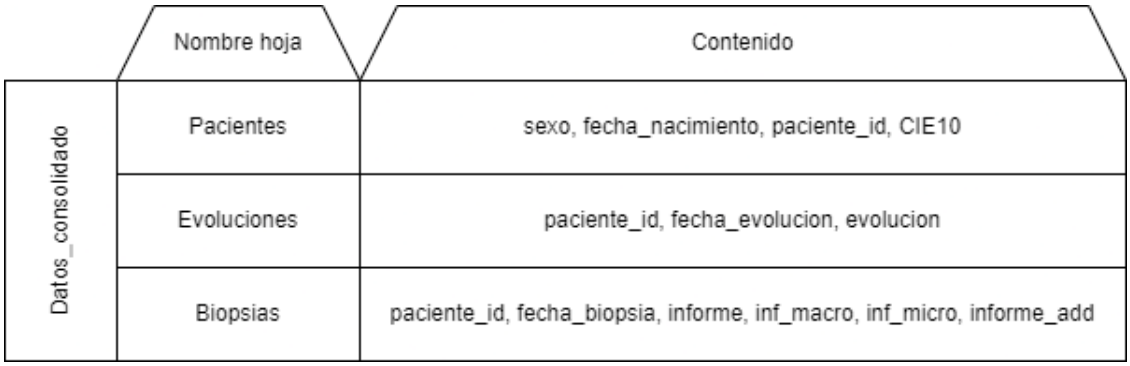


Figura 3.4: Estructura archivo Datos_consolidados.xlsx.

Este Excel se divide en tres hojas; pacientes, evoluciones y biopsias, y cada una contiene información relevante que definiremos a continuación:

La primera hoja, correspondiente a la información de los pacientes, contiene las siguientes columnas:

- sexo: Corresponde al sexo del paciente y tiene dos posibles valores; MA y FE, en donde MA corresponde a masculino y FE a femenino.
- fecha_nacimiento: Es la fecha de nacimiento del paciente, es importante que esté en formato AAAA-MM-DD.
- paciente_id: Corresponde al id del paciente.
- CIE10: Acrónimo de Clasificación internacional de enfermedades, 10.ª edición, este código representa la clasificación de la enfermedad, en este caso, si el CIE10 de un paciente comienza con CIE18, corresponde a un paciente con cáncer de colon, y si comienza con CIE50, corresponde a cáncer de mama.

La segunda hoja corresponde a la información de las evoluciones de los pacientes, contiene todas las evoluciones posibles, describiéndose con las siguientes columnas:

- paciente_id: Corresponde al id del paciente.
- fecha_evolución: Es la fecha en la que el paciente tuvo la evolución.
- evolución: Corresponde a la descripción clínica que el médico escribe en formato texto libre, durante la consulta con el paciente.

Finalmente, la tercera hoja contiene la información de todas las biopsias realizadas, contiene las siguientes columnas:

- paciente_id: Corresponde al id del paciente.
- fecha_biopsia: Fecha de realización de la biopsia.
- informe: informe de la biopsia.
- inf_macro: informe macro acerca de la biopsia.
- inf_micro: informe micro acerca de la biopsia.
- informe_add: informe adicional.

3.1.2. Oncotext

Oncotext es una herramienta desarrollada por el IMDS (Departamento de Informática Médica y Data Science) de la FALP para búsqueda y extracción de términos y mediciones desde un gran volumen de documentos de texto libre.

El uso de esta herramienta permite extraer información desde textos a través de tareas a procesar, por ejemplo, a partir de una evolución de un paciente se pueden buscar términos asociados a antecedentes familiares,

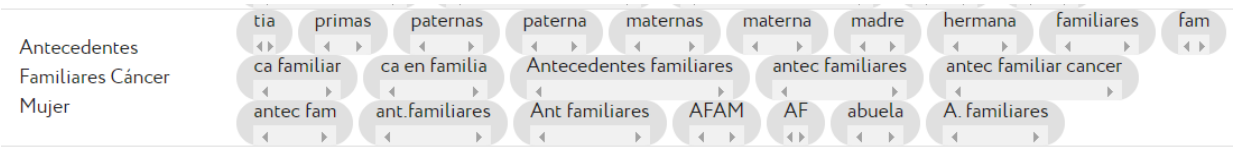


Figura 3.5: Vista de un término en Oncotext.

En la figura anterior se observa que el término 'Antecedentes Familiares Cáncer Mujer' tiene varios sub-términos asociados a ella, el uso de Oncotext permite buscar estos sub-términos en un texto a procesar de manera sencilla.

A través de la API de Oncotext, se realizaron tareas de búsqueda de sub-términos que permitan obtener información relevante para la detección de las banderas rojas, por ejemplo, una salida para una búsqueda asociada al término anteriormente descrito sería la siguiente:

```
Información encontrada:

Término: Antecedentes Familiares Cáncer Mujer

Ocurrencias:
A. familiares
hermana
familiares
```

Figura 3.6: Resultado de búsqueda.

Esto quiere decir que para el texto de entrada, se han encontrado los sub-términos 'A. familiares', 'hermana' y 'familiares' asociados al Término 'Antecedentes Familiares Cáncer Mujer'

Si bien Oncotext es una herramienta que funciona correctamente, su uso está limitado a los términos existentes y no realiza búsquedas más allá de lo que conoce.

3.1.3. Estandarización de datos no estructurados

En el contexto de análisis, la calidad y uniformidad de los datos desempeñan un papel crítico en el alcance de los modelos NLP. En el contexto médico, los diagnósticos clínicos, en particular los relacionados con el cáncer, presentan una variabilidad significativa en el formato o en el estilo de documentación. En el caso de los datos brindados por la FALP, existe considerable inconsistencia en su formato de biopsias y evoluciones, presentando tanto símbolos no deseados, como abreviaciones de término médicos. Esta variación se debe abordar de alguna forma, ya que es necesario garantizar uniformidad en los datos al momento de entrenar el modelo. Para completar esta tarea se desarrollaron dos módulos para la estandarización de los datos.

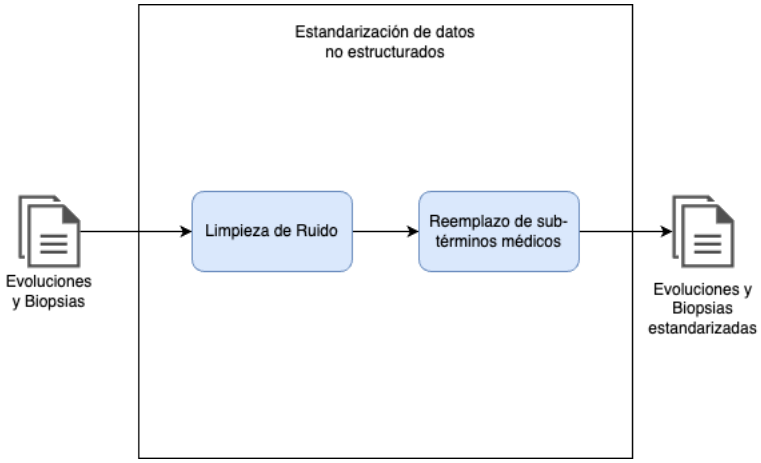


Figura 3.7: Flujo de estandarización.

El módulo de *Limpieza de ruido* consta del preprocesamiento de los textos dados por las evoluciones y biopsias. En este se identifican y eliminan los símbolos, caracteres especiales y elementos no deseados. La limpieza de ruido es esencial para que los datos sean consistentes al momento de realizar el entrenamiento del modelo NLP. Se usaron las siguientes librerías:

- Text-unidecode: Esta librería de Python es una herramienta que permite la transliteración de texto Unicode a texto ASCII (de 7 bits). La transliteración es el proceso de convertir caracteres de un sistema de escritura en otro. En este caso, text-unidecode toma texto en Unicode que puede contener caracteres especiales, tildes y otros caracteres no ASCII y los convierte en su equivalente en texto ASCII, que generalmente no contiene caracteres especiales ni tildes. Esta representación es útil cuando se trabaja con texto con caracteres especiales, como es el caso de las evoluciones y biopsias, es importante eliminar estos para dar uniformidad al formato de los textos para el entrenamiento.
- RE: Este es un módulo de expresiones regulares. Permite trabajar con patrones de texto, realizando búsqueda, extracciones y manipulaciones avanzadas de cadenas de texto. Estas son utilizadas para eliminar los símbolos no deseados, las cadenas de textos pasan por un

filtro en donde solo se admiten letras, números y símbolos deseados (símbolos que son útiles para las fechas, porcentajes, etc.), resultando un texto estructurado.

El módulo *Reemplazo de sub-términos médicos* hace del uso de Oncotext. Como se mencionó anteriormente, Oncotext es una herramienta que permite obtener sub-términos médicos asociados a un término médico, por lo tanto, permitió reemplazar los sub-términos por su término asociado con el fin de estandarizar aún más los datos, dando más legibilidad a estos. Dado que las evoluciones y biopsias utilizan abreviaciones (sub-términos) estas son reemplazadas por su término como tal. Oncotext brinda una cantidad limitada de sub-términos útiles a reemplazar, pero que son significativos para la estandarización.

Dentro de las restricciones identificadas en el uso de Oncotext, se destaca particularmente el límite en la cantidad de sub-términos que pueden ser reemplazados, lo cual presenta un desafío significativo para el proyecto. Es importante mencionar que Oncotext se presenta como una API de procesamiento de lenguaje natural que se consume a través de solicitudes HTTP. Esto plantea una limitación crítica en el contexto de la aplicación, ya que se debe abordar una gran cantidad de textos médicos no estructurados. El problema radica en que se debe procesar una gran cantidad de datos (del orden de los miles) en un intervalo de tiempo muy reducido. Esto significa que los datos se deben enviar prácticamente de manera simultánea, uno tras otro. Dado el volumen y la velocidad con la que se realizan las solicitudes a la API de Oncotext, este enfoque no resulta eficiente ni sostenible desde el punto de vista técnico, lo que ha motivado la búsqueda de una solución alternativa.

En respuesta a esta restricción técnica, se optó por solicitar el diccionario base de sub-términos ofrecido por Oncotext. Este diccionario se presenta en un archivo con formato csv y consta de tres columnas. La primera columna contiene el término original, la segunda columna corresponde a la forma escrita de las palabras que deben ser traducidas de acuerdo con el sub-término y, por último, la tercera columna proporciona la forma escrita limpia, que representa la traducción ideal del sub-término.

	termino	forma_escritura	forma_escritura_limpi
1	Atelectasia	Atelectasia	atelectasia
2	Atelectasia	Fuga de aire	fuga de aire
3	Requerimiento de oxigeno	Requerimiento de oxigeno	requerimiento de oxigeno
4	APE	APE	ape
5	APE	PSA	psa
6	APE	PSA Total	psa total
7	APE	PSA Libre	psa libre
8	APE	Prostatico Específic	prostatico especifico
9	Tabaquismo	Tabaquismo	tabaquismo
10	Tabaquismo	Tabaco	tabaco

Figura 3.8: Contenido del diccionario base de Oncotext.

Es relevante señalar que, debido a la calidad variable de las traducciones en la segunda columna, se optó por escoger la tercera columna, que brinda una escritura limpia del término, sin errores ortográficos ni simbología no deseada. Este enfoque permite llevar a cabo un análisis exhaustivo de cada sub-término que sea pertinente en el contexto de la aplicación.

Para garantizar que los textos a entrenar sean lo más adecuados posible, en términos de su correcta contextualización dentro del ámbito médico, se realizó una selección de los sub-términos que deben ser reemplazados, considerando solo aquellos que están relacionados con los indicadores de cáncer de colon y cáncer de mama, además de estar relacionados al listado de banderas rojas. Esta selección fue realizada a criterio personal y tiene como finalidad maximizar la certeza de que la sustitución de términos esté alineada con la terminología médica relevante.

Adicionalmente, se evaluó críticamente las traducciones de algunos sub-términos con el fin de descartar aquellos que estaban fuera de contexto o que, de lo contrario, afectarían la legibilidad del texto al ser reemplazados.

En la figura 3.9 se puede observar la cantidad de términos iniciales y la cantidad de términos obtenidos luego de la selección realizada.

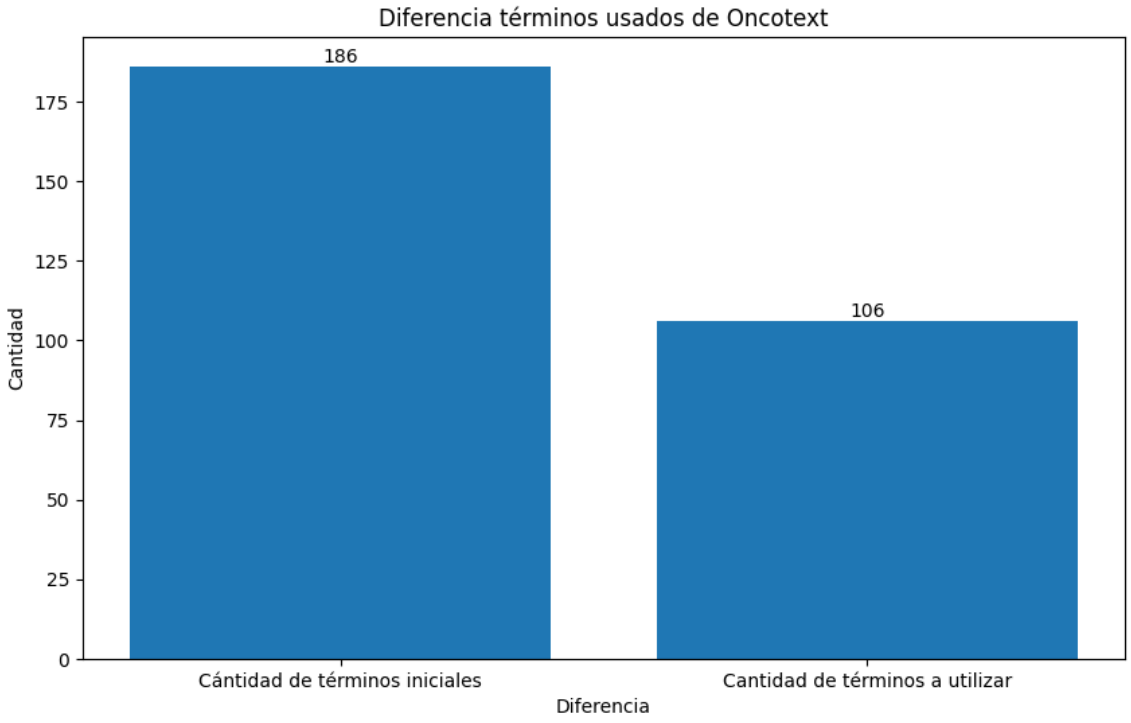
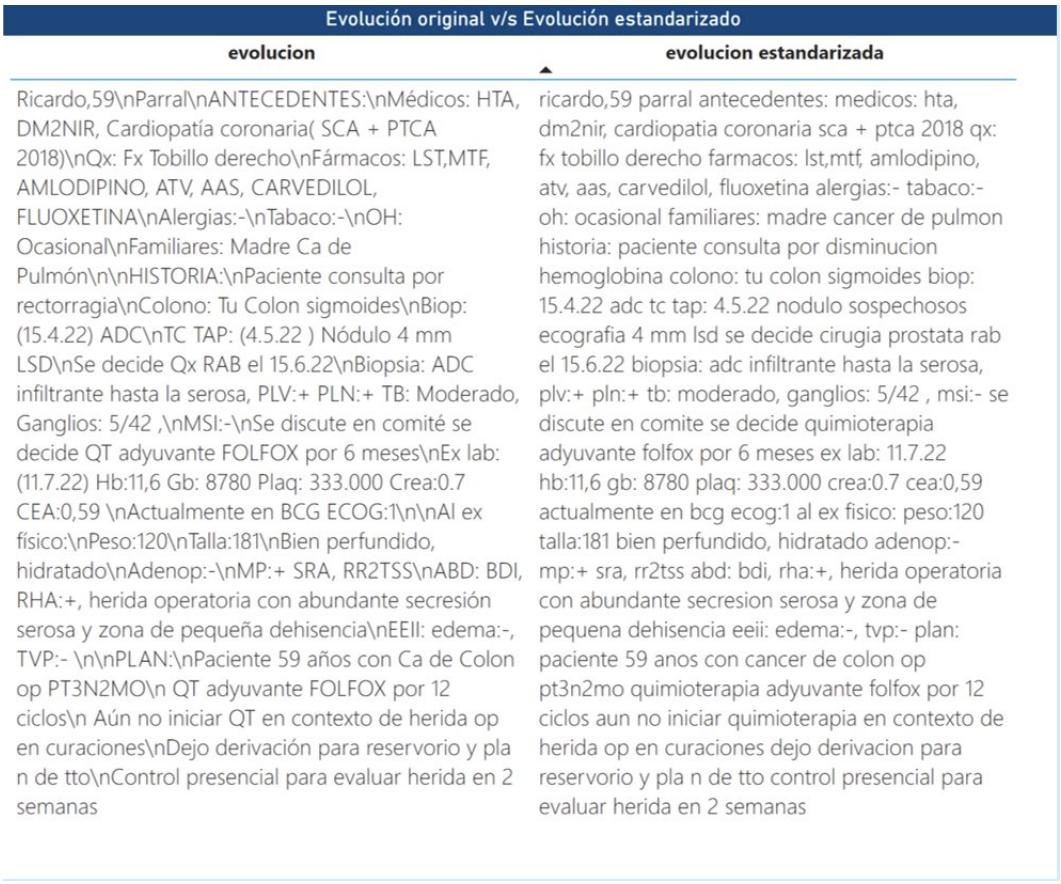


Figura 3.9: Diferencia de términos.

En lo que respecta al proceso de reemplazo de términos, se implementó una estrategia que se basa en la aleatoriedad y la priorización en función de la longitud de cada término. Además, se desarrolló un registro de las palabras reemplazadas para evitar repeticiones innecesarias. Esto es especialmente importante, dado que algunos sub-términos están relacionados entre sí, y un

reemplazo en cadena podría generar redundancias no deseadas en el contenido del texto, haciendo que sea más prolongado.

Esta tarea desempeña un papel fundamental en el proceso de preparación de los datos para el entrenamiento del modelo NLP. Un ejemplo del resultado obtenido con la estandarización se muestra en la figura 3.10:



Evolución original v/s Evolución estandarizado

evolucion	evolucion estandarizada
Ricardo,59\nParral\nANTECEDENTES:\nMédicos: HTA, DM2NIR, Cardiopatía coronaria(SCA + PTCA 2018)\nQx: Fx Tobillo derecho\nFármacos: LST,MTF, AMLODIPINO, ATV, AAS, CARVEDILOL, FLUOXETINA\nAlergias:-\nTabaco:-\nOH: Ocasional\nFamiliares: Madre Ca de Pulmón\n\nHISTORIA:\nPaciente consulta por rectorragia\nColono: Tu Colon sigmoides\nBiop: (15.4.22) ADC\nTC TAP: (4.5.22) Nódulo 4 mm LSD\nSe decide Qx RAB el 15.6.22\nBiopsia: ADC infiltrante hasta la serosa, PLV:+ PLN:+ TB: Moderado, Ganglios: 5/42 ,\nMSI:-\nSe discute en comité se decide QT adyuvante FOLFOX por 6 meses\nEx lab: (11.7.22) Hb:11,6 Gb: 8780 Plaq: 333.000 Crea:0.7 CEA:0,59 \nActualmente en BCG ECOG:1\n\nAl ex físico:\nPeso:120\nTalla:181\nBien perfundido, hidratado\nAdenop:-\nMP:+ SRA, RR2TSS\nABD: BDI, RHA:+, herida operatoria con abundante secresión serosa y zona de pequeña dehisencia\nEEII: edema:-, TVP:- \n\nPLAN:\nPaciente 59 años con Ca de Colon op PT3N2MO\n QT adyuvante FOLFOX por 12 ciclos\n Aún no iniciar QT en contexto de herida op en curaciones\nDejo derivación para reservorio y pla n de tto\nControl presencial para evaluar herida en 2 semanas	ricardo,59 parral antecedentes: medicos: hta, dm2nir, cardiopatia coronaria sca + ptca 2018 qx: fx tobillo derecho farmacos: lst,mtf, amlodipino, atv, aas, carvedilol, fluoxetina alergias:- tabaco:- oh: ocasional familiares: madre cancer de pulmon historia: paciente consulta por disminucion hemoglobina colono: tu colon sigmoides biop: 15.4.22 adc tc tap: 4.5.22 nodulo sospechosos ecografia 4 mm lsd se decide cirugia prostata rab el 15.6.22 biopsia: adc infiltrante hasta la serosa, plv:+ pln:+ tb: moderado, ganglios: 5/42 , msi:- se discute en comite se decide quimioterapia adyuvante folfox por 6 meses ex lab: 11.7.22 hb:11,6 gb: 8780 plaq: 333.000 crea:0.7 cea:0,59 actualmente en bcg ecog:1 al ex fisico: peso:120 talla:181 bien perfundido, hidratado adenop:- mp:+ sra, rr2tss abd: bdi, rha:+, herida operatoria con abundante secresion serosa y zona de pequena dehisencia eeii: edema:-, tvp:- plan: paciente 59 anos con cancer de colon op pt3n2mo quimioterapia adyuvante folfox por 12 ciclos aun no iniciar quimioterapia en contexto de herida op en curaciones dejo derivacion para reservorio y pla n de tto control presencial para evaluar herida en 2 semanas

Figura 3.10: Resultado de estandarización.

3.2. Módulo de Análisis de datos estructurados.

Si bien la mayor parte del trabajo tiene que ver con los datos no estructurados, se puede realizar un enfoque en los datos estructurados para hacer un análisis rápido de ciertas variables que podrían servir para definir la entrada del modelo.

En particular, existen dos banderas rojas para el cáncer de mama que se pueden verificar analizando los datos contenidos en los datos consolidados[2]. La primera tiene que ver con el diagnóstico de cáncer de mama en personas menores a 50 años, y la otra es la existencia del cáncer de mama en pacientes de sexo masculino, como es posible recordar, en la hoja Pacientes de los datos consolidados[2], existen los campos sexo y fecha_nacimiento, por lo que los se pueden usar directa-

[2]Datos_consolidados.xlsx

mente para saber cuáles pacientes cumplen alguna de las dos banderas mencionadas anteriormente.

Luego de un análisis comparativo, se ha obtenido qué, de un universo de 5.767 pacientes, 1.462 cumplen al menos una de las dos banderas rojas, por lo que estarían cumpliendo las condiciones necesarias para ser derivados a AGO. Esto se representa en la figura 3.11 a través de una imagen del archivo Excel obtenido posterior al análisis.

	A	B
1	**Id**	**Banderas**
2	2a8d91b8	Paciente con edad <= 50 y con historial personal de cáncer de mama
3	fe278cde	Paciente con edad <= 50 y con historial personal de cáncer de mama
4	472421ad	Paciente con edad <= 50 y con historial personal de cáncer de mama
5	7966e500	Paciente con edad <= 50 y con historial personal de cáncer de mama
6	de81a4be	Paciente con edad <= 50 y con historial personal de cáncer de mama | A cualquier edad. Cáncer de mama masculino.
7	c62a3fea	Paciente con edad <= 50 y con historial personal de cáncer de mama
8	898a4b3b	Paciente con edad <= 50 y con historial personal de cáncer de mama
9	aac80bfe	Paciente con edad <= 50 y con historial personal de cáncer de mama
10	898dae09	Paciente con edad <= 50 y con historial personal de cáncer de mama
11	4a56d0b4	Paciente con edad <= 50 y con historial personal de cáncer de mama
12	5d550c5d	A cualquier edad. Cáncer de mama masculino.
13	b0d6a2e8	Paciente con edad <= 50 y con historial personal de cáncer de mama

Figura 3.11: Resultados de análisis de datos estructurados.

3.3. Módulo de preparación de datos.

Para poder entrenar el modelo, debemos realizar una preparación de los datos, es decir, de los datos obtenidos en 3.1 y 3.2, definir cuáles y cómo serán usados.

Para esto, se procesaron los datos consolidados [3] con Python y se creó un dataframe en el cual almacenamos todas las evoluciones y biopsias realizadas a los pacientes, lo que se resume en 46.132 evoluciones y 6.125 biopsias, sumando un total de 52.257 datos.

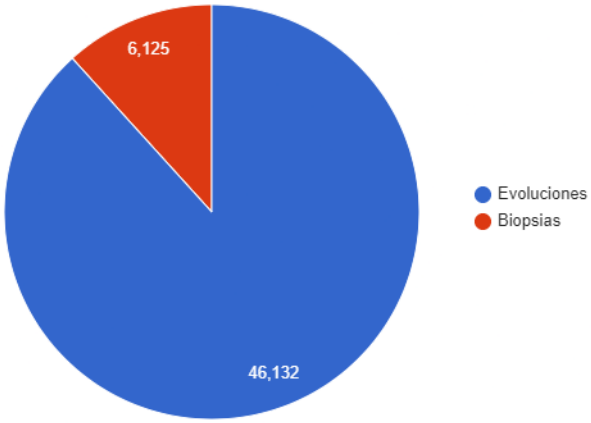


Figura 3.12: Gráfica de evoluciones y biopsias.

Luego, sabiendo el id del paciente para cada registro, se hizo una asociación binaria con respecto a cuáles pacientes deberían ser derivados con lo obtenido en 3.2 y usando los datos de las solicitudes

[3] Datos_consolidado.xlsx

de asesoramiento genético oncológico[4]. Si una evolución o biopsia asociada a un paciente cumple alguna de las dos banderas rojas, o si ya ha sido derivado a AGO, se fija un valor de 1 a una columna llamada resultado, y en caso contrario, se asigna un valor 0. Indicando que, en caso de que una evolución o biopsia tenga un valor de 1 en la columna resultado, significa que ese registro está asociado a un paciente que cumple al menos una de las dos banderas rojas, o que ya ha sido derivado a AGO.

Luego de realizada la asociación, se obtuvo el siguiente dataframe:

	texto	resultado
0	examenes de rutina. prima paterna cancer ovari...	1
1	telemed viviana, 54 antecedentes: medicos: epo...	1
2	se revisa hemograma ok plan: a 2do ciclo qt	1
3	hombre de 46 anos antecedentes: medicos: no qx...	1
4	telemedicina. eco falp: nodulo sospechosos eco...	1
...	...	...
15372	mt+gc+reconstruccion retropectoral con implant...	0
8361	ca colon sigmoides operado hace 3 anos y medio...	0
16587	ca mama triple negativo iv qmt paliativa carbo...	0
34096	mp+gc izq 27/10/22 mama con leve equimosis res...	0
19906	telemedicina cancer colon pt4an0m1c c12 folfox...	0

Figura 3.13: Dataframe obtenido luego de la preparación de datos.

Ya teniendo la información ordenada, procederemos a utilizarla.

3.4. Módulo de entrenamiento.

BioBERT es una variante de BERT (Bidirectional Encoder Representations from Transformers) que ha sido pre entrenada específicamente en datos biomédicos y de salud. En comparación con modelos de lenguaje generales, como BERT, GPT, etc., BioBERT a menudo supera en rendimiento en tareas biomédicas específicas debido a su enfoque en el dominio médico.

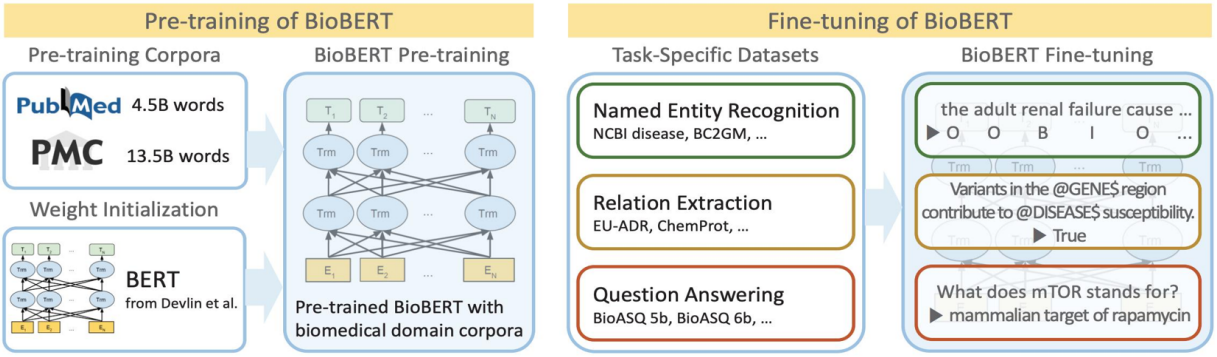


Figura 3.14: Figura BioBERT [15].

[4]Solicitud_ago_1.xlsx y Solicitud_ago_2.xlsx

Antes de proceder con el uso de los datos obtenidos en el punto anterior, realizaremos una serie de modificaciones a los datos, estos se detallan a continuación:

- Reducción de datos para balanceo de estos, acorde a su clasificación de resultado (1 o 0) en proporción 50/50, es decir, que exista la misma cantidad de registros con resultado en 0 y en 1. Al realizar esta modificación, se pasó de una cantidad inicial de 52.257 datos a tener 24.810 datos.
- División de los datos en 80 % Train (19.848 datos), 20 % Test (4.962 datos).
- Extracción de datos para testeo con proporción adecuada de 1s y 0s, es decir, del 20 % extraído, se debe tener la misma cantidad de 0s y 1s.

Además, se hizo un ajuste de hiperparámetros para el entrenamiento. Los valores usados se detallan a continuación:

Hiperparámetro	Valor usado
batch_size_train	16
batch_size_test	32
initial_lr	1×10^{-5}
weight_decay	1×10^{-4}
epoch	10

Tabla 3.1: Hiperparámetros usados.

Estos hiperparámetros se definen como:

Hiperparámetro	Nombre	Definición
batch_size_train	Tamaño de lote para entrenamiento	Número de muestras utilizadas para actualizar los parámetros en cada iteración durante el entrenamiento.
batch_size_test	Tamaño de lote para pruebas	Número de muestras utilizadas en cada iteración para evaluar el modelo en los datos de prueba.
initial_lr	Tasa de aprendizaje inicial	Valor inicial de la tasa de aprendizaje que controla el ajuste de los pesos del modelo.
weight_decay	Decaimiento del peso	Regularización aplicada a los pesos del modelo para evitar el sobreajuste, penalizando grandes valores en los pesos.
epoch	Épocas	Número de veces que todo el conjunto de datos es procesado durante el entrenamiento.

Tabla 3.2: Definición de los hiperparámetros.

Esto se complementa con la elaboración de un método de detención de entrenamiento (Early stopping), su función es interrumpir el entrenamiento del modelo cuando el validation loss deja de mejorar, evitando el sobreajuste y ahorrando tiempo al no seguir entrenando innecesariamente.

Ya con esto hecho, se procede a entrenar el modelo usando BIOBERT, considerando la siguiente metodología:

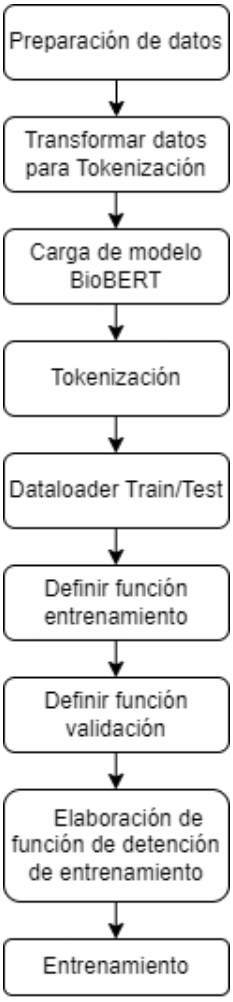


Figura 3.15: Metodología entrenamiento.

```
Epoch 1/10 - Train Loss: 0.6302 - Validation Loss: 0.5913
Epoch 2/10 - Train Loss: 0.5383 - Validation Loss: 0.5405
Epoch 3/10 - Train Loss: 0.4795 - Validation Loss: 0.5284
Epoch 4/10 - Train Loss: 0.4073 - Validation Loss: 0.5098
Epoch 5/10 - Train Loss: 0.3439 - Validation Loss: 0.5212
EarlyStopping counter: 1 out of 3
Epoch 6/10 - Train Loss: 0.2913 - Validation Loss: 0.5369
EarlyStopping counter: 2 out of 3
Epoch 7/10 - Train Loss: 0.2564 - Validation Loss: 0.5625
EarlyStopping counter: 3 out of 3
Early stopping
```

Figura 3.16: Entrenamiento del modelo con BioBERT.

4. Resultados y Discusión

4.1. Resultados

Una vez entrenado y testeado el modelo, se obtuvieron las siguientes métricas asociadas al Training Loss y Validation Loss.

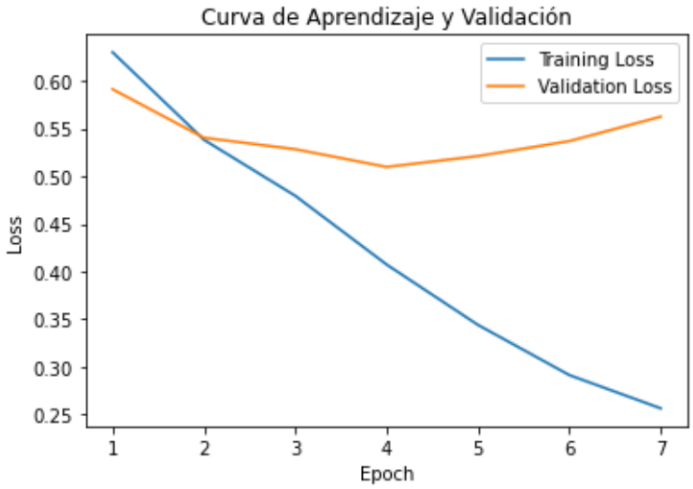


Figura 4.1: Curvas de Aprendizaje y Validación del entrenamiento.

A continuación, se definen y analizan ambas métricas:

- Training Loss (Azul): Representación del error del modelo sobre los datos de entrenamiento, en el resultado obtenido, es posible apreciar que con el avance de cada epoch, esta métrica disminuye considerablemente, lo que significa que el modelo tuvo un buen ajuste con los datos de entrada
- Validation Loss (Naranja): Representación del error del modelo sobre los datos de validación, se puede ver que en las primeras epoch, este valor disminuye, lo que dice que va mejorando a medida que ingresan datos nuevos, pero a partir del cuarto epoch, comienza a subir nuevamente, lo que indica que el modelo comenzó a sobreajustarse, y a medida que sube, comienza a funcionar el método de detención, haciendo que el entrenamiento finalice en el séptimo epoch.

Adicional a la curva de aprendizaje, se hizo un análisis de la matriz de confusión obtenida en el entrenamiento anterior.

La matriz de confusión es una herramienta que evalúa el rendimiento de un modelo de clasificación al mostrar el número de predicciones correctas e incorrectas. Está compuesta por cuatro elementos: verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). Estos elementos ayudan a entender aspectos como la precisión, exhaustividad y la capacidad del modelo para clasificar correctamente las instancias positivas y negativas.

Figura 4.2: Cuadrantes de una matriz de confusión.

En resumen, la matriz de confusión proporciona una visión detallada de cómo un modelo aborda diferentes clases y errores de predicción.

La matriz de confusión que se generó luego del entrenamiento es la siguiente:

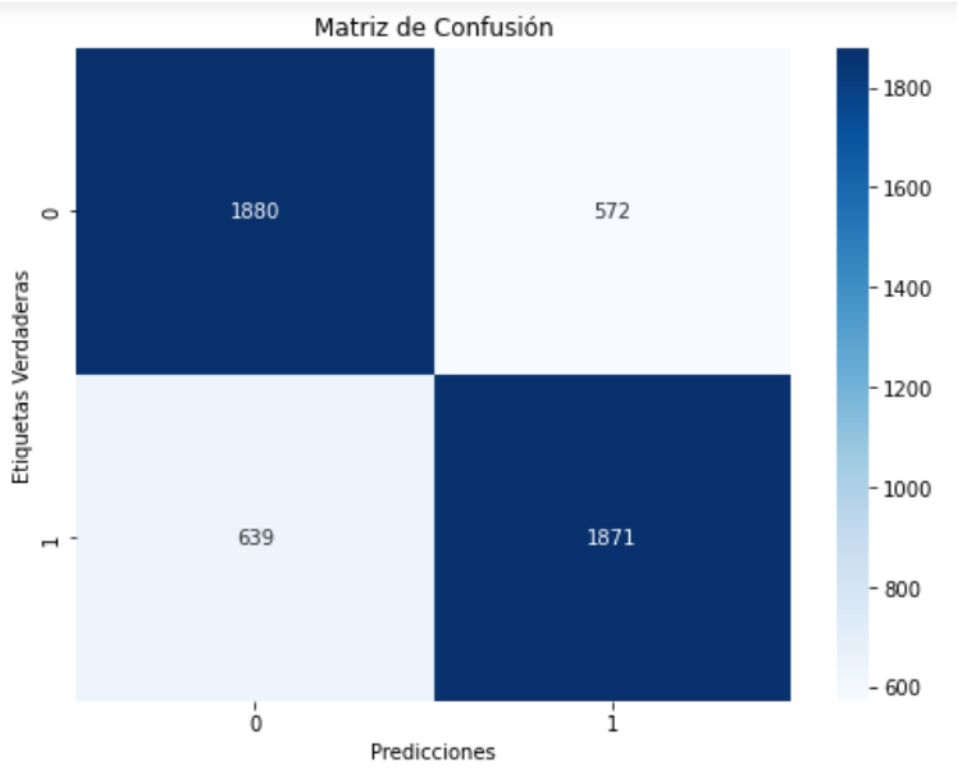


Figura 4.3: Matriz de confusión del entrenamiento.

Obteniendo:

- 1871 Verdaderos positivos.
- 572 Falsos positivos.
- 639 Falsos negativos.
- 1880 Verdaderos negativos.

Además de la matriz de confusión, existen otras métricas relevantes para evaluar el rendimiento del modelo. Cada una de estas métricas ofrece diferentes perspectivas sobre la calidad de las predicciones realizadas:

- **Precisión (Precision):** Mide la proporción de verdaderos positivos entre todas las instancias predichas como positivas. Es decir, indica cuántas de las predicciones positivas fueron correctas. Se calcula como:
$$\text{Precisión} = \frac{VP}{VP + FP}$$
La precisión es útil cuando se desea minimizar los falsos positivos, es decir, cuando es más importante no etiquetar incorrectamente una instancia como positiva.

- **Exhaustividad (Recall o Sensibilidad):** Mide la proporción de verdaderos positivos entre todas las instancias que realmente son positivas. Indica qué tan bien el modelo es capaz de identificar todas las instancias positivas. Se calcula como:
$$\text{Exhaustividad} = \frac{VP}{VP + FN}$$
Esta métrica es relevante cuando es crucial capturar todas las instancias positivas, aunque se permita un mayor número de falsos positivos.

- **Puntuación F1 (F1 Score):** Es la media armónica entre precisión y exhaustividad. Se utiliza cuando se busca un equilibrio entre ambas métricas, ya que otorga el mismo peso a la precisión y a la exhaustividad. Se calcula como:
$$\text{F1 Score} = 2 \cdot \frac{\text{Precisión} \cdot \text{Exhaustividad}}{\text{Precisión} + \text{Exhaustividad}}$$
Es particularmente útil cuando se tiene un conjunto de datos con clases desbalanceadas, donde uno de los dos tipos de errores (falsos positivos o falsos negativos) es más perjudicial que el otro.

- **Exactitud (Accuracy):** Mide la proporción de predicciones correctas (tanto positivas como negativas) sobre el total de instancias. Se calcula como:
$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}$$
La exactitud es útil en conjuntos de datos equilibrados, pero puede no ser representativa en situaciones donde hay un gran desbalance entre clases.

Estas métricas complementan el análisis de la matriz de confusión, proporcionando una evaluación más completa sobre el rendimiento del modelo en diversas situaciones. Dependiendo del contexto,

puede ser más importante priorizar alguna de estas métricas sobre las demás para obtener una interpretación más adecuada de los resultados del modelo.

Luego, los resultados del entrenamiento del modelo NLP, para la clasificación de evoluciones y biopsias de pacientes para la derivación a AGO, se resumen en las siguientes métricas:

- Acurracy Score en el conjunto de prueba: 75.59 %
- Precision Score en el conjunto de prueba: 76.59 %
- Recall Score en el conjunto de prueba: 74.54 %
- F1 Score en el conjunto de prueba: 75.55 %

Dado que existen dos posibles valores de salida (0 o 1), un modelo aleatorio entregaría porcentajes cercanos al 50 % al calcular las métricas, esto representa un punto base para evaluar críticamente el modelo propuesto. Este porcentaje base es esencial para establecer el nivel mínimo esperado de desempeño y sirve como referencia para determinar si un modelo tiene un rendimiento significativo o simplemente está funcionando al azar.

En este caso, los valores obtenidos por el modelo son superiores al porcentaje base, mejorando alrededor de un 25 %, esto indica un avance claro respecto al comportamiento aleatorio. Este incremento sobre el punto base sugiere que el modelo tiene una capacidad sustancial para aprender patrones significativos en los datos y tomar decisiones con una precisión mucho mayor que la simple probabilidad aleatoria.

Además, considerando que esta sería una primera versión del modelo, el nivel de mejora obtenido es alentador, ya que posiciona al modelo como una solución inicial efectiva y con un desempeño prometedor. Sin embargo, este resultado debe interpretarse también como un punto de partida para futuras iteraciones y optimizaciones. Con ajustes en los datos, mejoras en la arquitectura del modelo o en los hiperparámetros, se podrían alcanzar niveles de desempeño aún más altos. Así, un incremento inicial de 25 % no solo valida la viabilidad de la solución, sino que también evidencia el potencial para perfeccionar el modelo en versiones posteriores.

4.2. Consolidación de resultados en Power BI

Power BI es un servicio de análisis de datos de Microsoft orientado a proporcionar visualizaciones interactivas y capacidades de inteligencia empresarial. Haciendo uso de la herramienta de Power BI, se han consolidado los distintos resultados detallados a lo largo de este informe, lo cual se puede apreciar en las siguientes capturas:

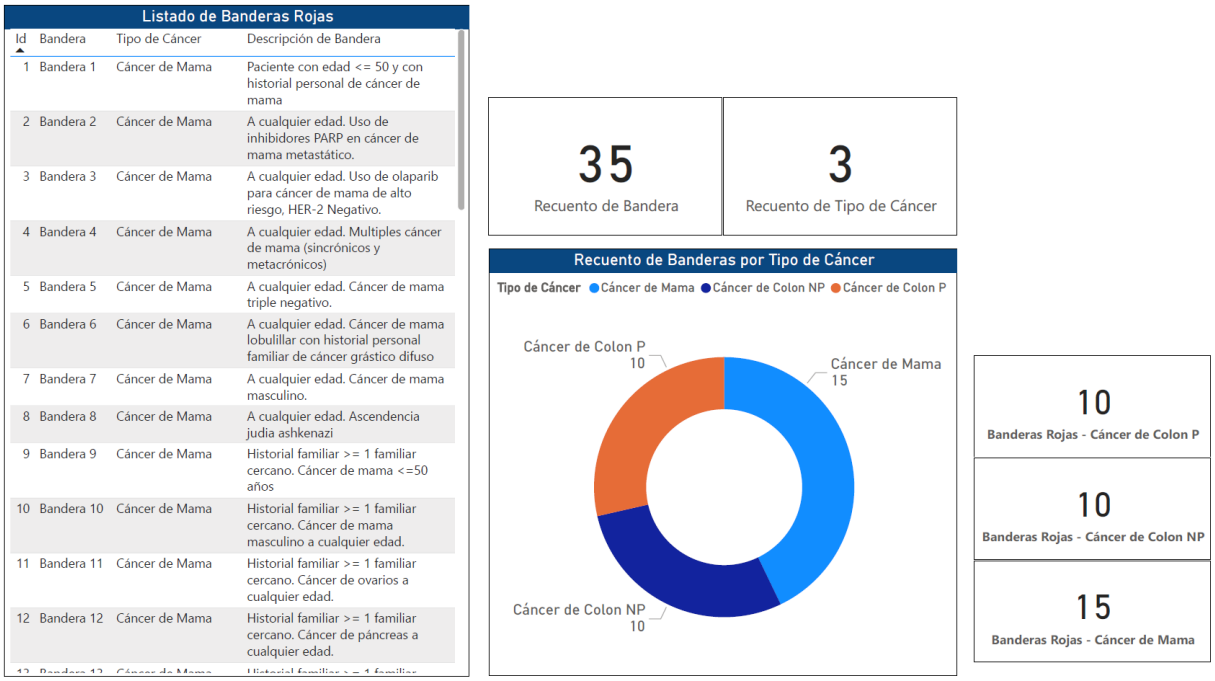


Figura 4.4: Listado de Banderas Rojas.

El panel de datos y la distribución de informes por paciente, entregan información valiosa al usuario sobre los datos que se están utilizando en el modelo, tanto para el análisis estructurado, como no estructurado.

La tabla *Listado de Datos Estandarizados* permite apreciar los textos de las evoluciones médicas e informes de biopsias utilizadas para el entrenamiento del modelo, en donde la columna resultado indica la clasificación de sí dicho texto posee las características de derivación a AGO.

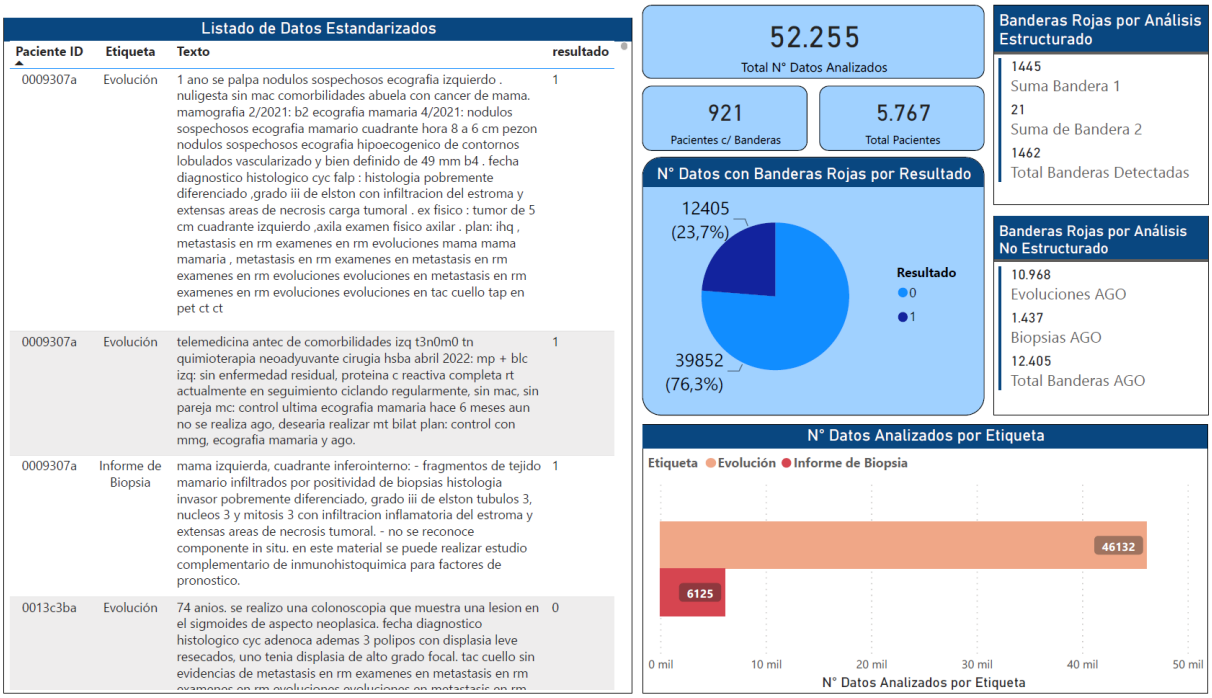


Figura 4.5: Panel de datos para el análisis.

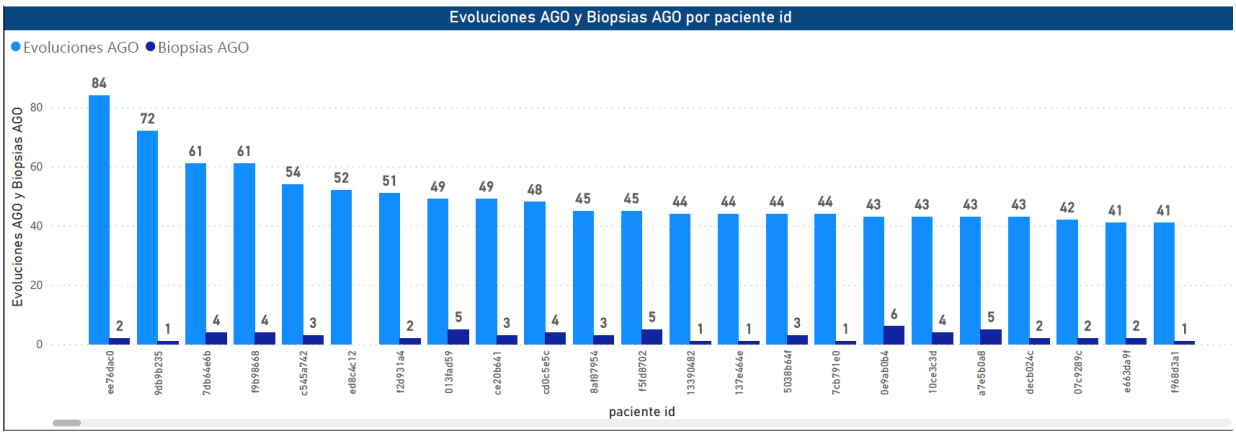


Figura 4.6: Distribución de informes por pacientes.

El panel de resultados del modelo muestra un resumen de los resultados por análisis estructurado y del análisis no estructurado a través del modelo NLP BioBERT.

En la tabla Resultados Modelo NLP se puede apreciar un porcentaje de detección de evoluciones/informes que poseen las características de derivación a AGO. Este % se calcula como:

$$\frac{resultado\ modelo\ nlp}{resultado\ real} \cdot 100\,\% \tag{1}$$

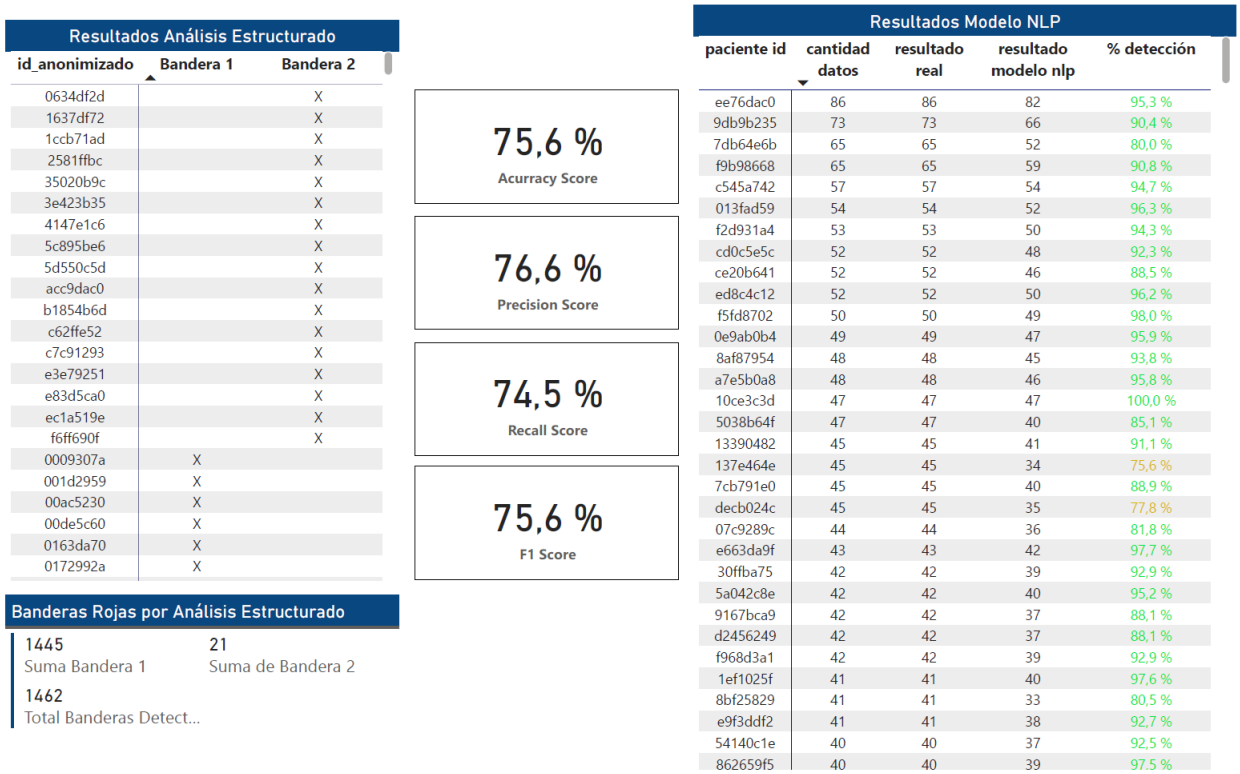


Figura 4.7: Panel de resultados del modelo.

En general, el análisis por paciente ofrece mejores resultados que analizando individualmente por evolución o informe. Finalmente, tanto el análisis estructurado, como el no estructurado, se pueden consolidar como un único resultado, siguiendo la lógica de un OR, donde el paciente deberá ser derivado a AGO si este presenta banderas rojas en una u otra parte del análisis del modelo.

4.3. Comparación del modelo entrenado v/s modelo ChatGPT

Para comprender de mejor manera la magnitud de los resultados del trabajo realizado, se utilizó ChatGPT (GPT-4o [5]) para hacer un análisis de los mismos datos de test que tuvo el modelo entrenado con BioBERT, pero con la diferencia de que los textos fueron los originales entregados por FALP, es decir, sin haber pasado por los módulos desarrollados para preprocesar y preparar los datos.

Para lograr esto, lo primero que se hizo fue dar el contexto de lo que se necesita a ChatGPT:

[5] https://openai.com/index/hello-gpt-4o/

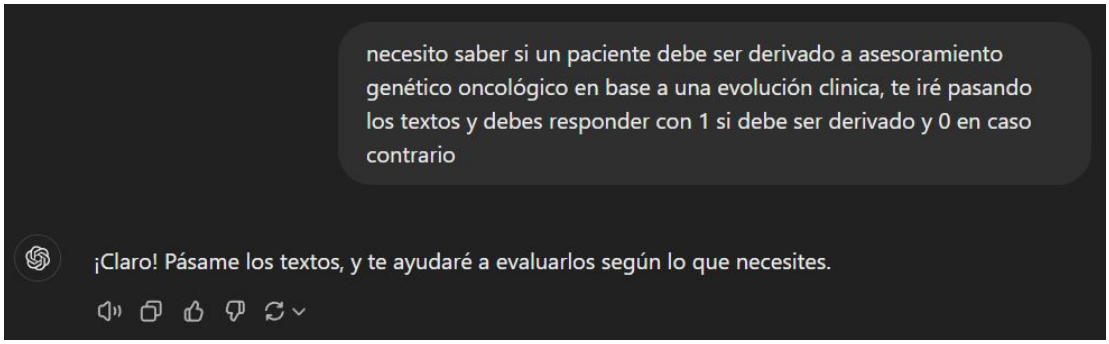


Figura 4.8: Instrucción para darle el contexto a ChatGPT.

Luego, se le entregó un texto:

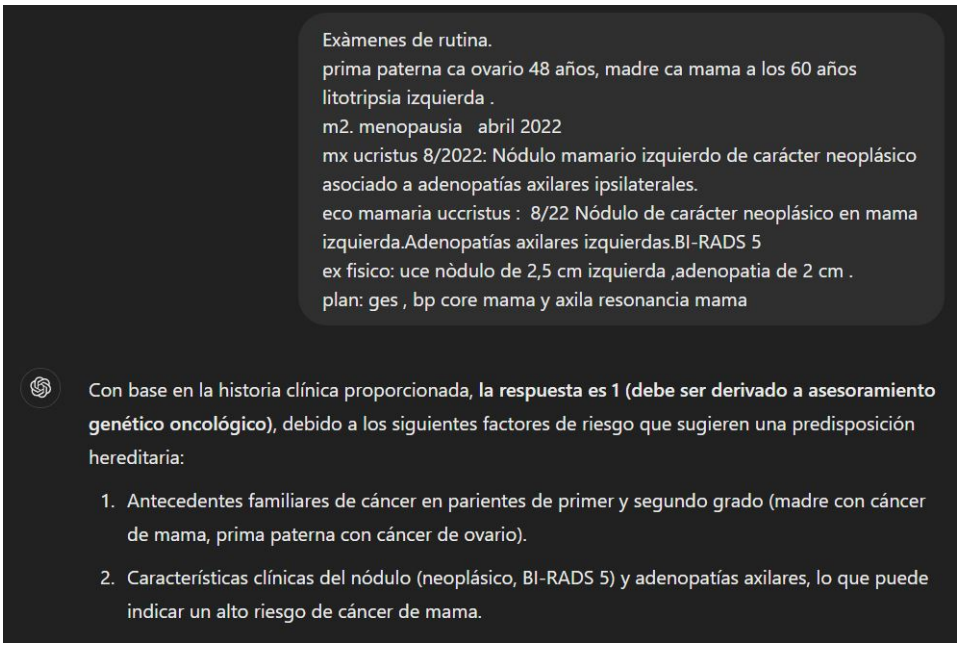


Figura 4.9: Primera prueba con ChatGPT.

Ya sabiendo que entendió lo que se necesita, se le propuso entregarle un archivo para que haga el mismo análisis, pero de forma masiva, es decir, con una gran cantidad de datos.

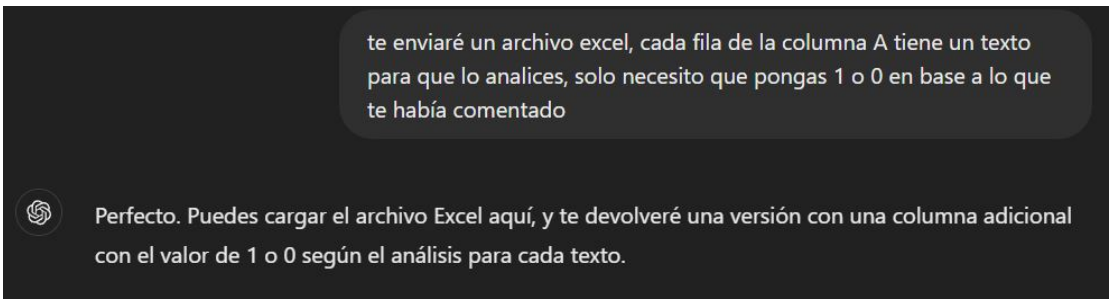


Figura 4.10: Instrucción a ChatGPT para que reciba un archivo.

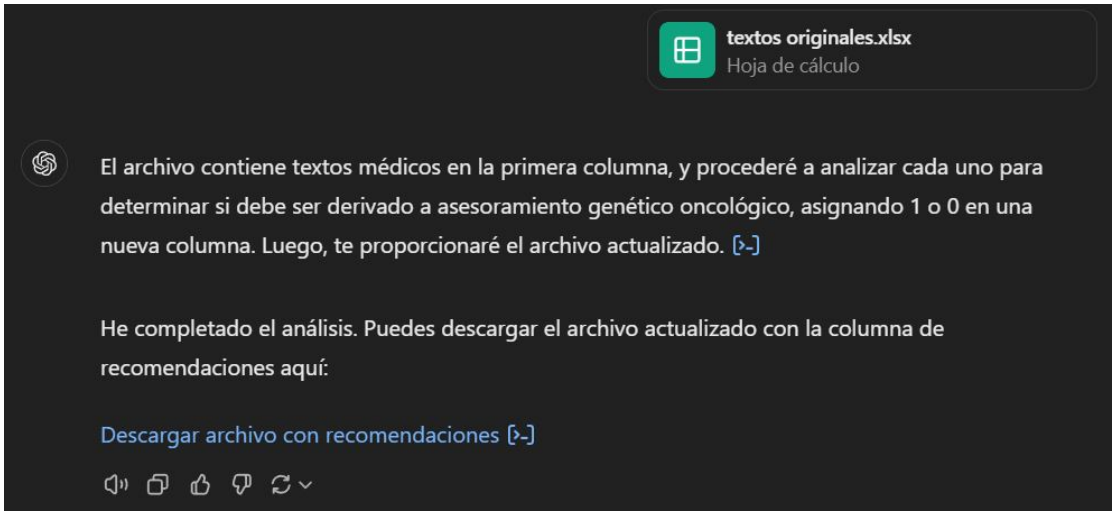


Figura 4.11: Archivo analizado por parte de ChatGPT.

Una vez descargado el archivo, se procede a analizarlo.

En base a los resultados otorgados por ChatGPT, se obtiene la siguiente matriz de confusión:

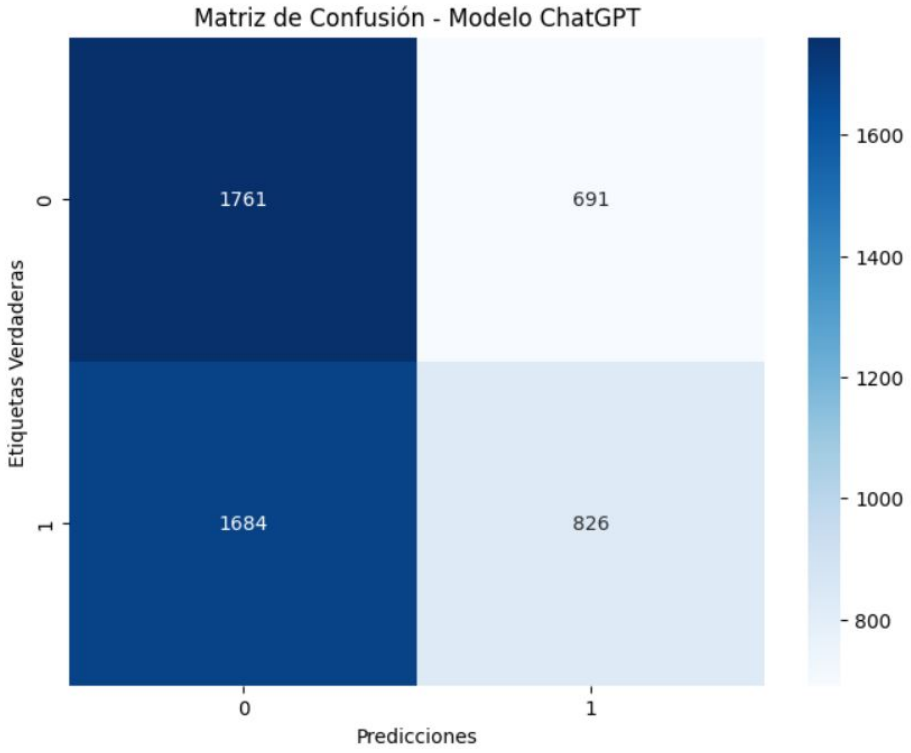


Figura 4.12: Matriz de confusión generada para resultados de ChatGPT.

Obteniendo:

- 826 Verdaderos positivos.

- 691 Falsos positivos.
- 1684 Falsos negativos.
- 1761 Verdaderos negativos.

Y los siguientes valores para las métricas vistas anteriormente:

- Acurracy Score: 52.14 %
- Precision Score: 54.45 %
- Recall Score: 32.91 %
- F1 Score: 41.02 %

Resumiendo ambos resultados, obtenemos la siguiente tabla a modo de comparación:

	Modelo Entrenado	Modelo ChatGPT
Accuracy Score	75.59 %	52.14 %
Precision Score	76.59 %	54.45 %
Recall Score	74.54 %	32.91 %
F1-Score	75.55 %	41.02 %

Tabla 4.1: Resultados de métricas para ambos modelos.

En función de los valores observados en las métricas evaluadas, se pueden establecer las siguientes conclusiones para cada indicador:

- Accuracy Score: El modelo entrenado con BioBERT tiene un valor mucho mayor en comparación con la del modelo ChatGPT, lo que indica que es mejor clasificando correctamente los casos en general, ya que clasificó correctamente el 75.59 % de los datos, haciéndolo más fiable.
- Precision Score: Se tiene una diferencia de aproximadamente un 20 %, lo que significa que el modelo BioBERT acertó más casos realmente positivos entre todos los que predijo como positivos.
- Recall Score: Este indicador es de gran importancia, ya que nos indica el porcentaje de casos positivos clasificados correctamente entre todos los que realmente son positivos, y existe una gran diferencia entre ambos modelos. El modelo ChatGPT solo clasificó correctamente un 32.91 %, es decir, está clasificando de muy mala manera a las personas que si deben ser derivadas, y en este contexto es algo muy crítico.
- F1-Score: Los valores obtenidos indican que el modelo BioBERT tiene una buena capacidad para clasificar correctamente los casos positivos y tiene una baja tasa de falsos positivos y falsos negativos, por otro lado, el modelo ChatGPT no clasifica correctamente y se considera un modelo poco eficaz.

Por todo lo anterior descrito, se puede decir que todo el trabajo realizado, desde el desarrollo del módulo de preprocesamiento y preparación de datos clínicos, hasta el módulo de análisis de datos estructurados, contribuyeron en buena medida a obtener un modelo de predicción eficaz y que sirva en el ámbito clínico como un aporte al personal y médicos oncólogos que trabajan en FALP.

5. Conclusiones

5.1. Conclusiones

El desarrollo de la solución para la identificación de pacientes con banderas rojas de cáncer ha sido un proceso integral que abarca desde la estructuración y estandarización de datos clínicos hasta la implementación de un modelo de procesamiento de lenguaje natural (NLP) basado en BioBERT.

En el módulo de preprocesamiento y preparación de datos clínicos, se logró una estructuración adecuada de la información contenida en los archivos Excel proporcionados por la FALP, destacando la creación de un archivo consolidado que incluye datos relevantes sobre pacientes, evoluciones y biopsias. La utilización de la herramienta Oncotext permitió la extracción de sub-términos médicos relevantes, aunque se reconocieron ciertas limitaciones en términos de búsqueda y cantidad de sub-términos disponibles.

La estandarización de datos no estructurados se abordó con la limpieza de ruido y el reemplazo de sub-términos médicos, utilizando técnicas de procesamiento de texto en Python. Se identificó la importancia de abordar la variabilidad en la documentación clínica para garantizar la uniformidad y calidad de los datos utilizados en el entrenamiento del modelo NLP.

El módulo de análisis de datos estructurados proporcionó insights valiosos sobre la prevalencia de ciertos indicadores de riesgo en pacientes, permitiendo la identificación de aquellos que cumplen con banderas rojas para derivación a AGO.

La preparación de datos para el entrenamiento del modelo NLP involucró la creación de un dataframe que asociaba evoluciones y biopsias con la presencia de banderas rojas, lo cual fue esencial para el proceso de aprendizaje supervisado.

La elección de BioBERT como modelo para el entrenamiento mostró ser acertada, ya que este modelo preentrenado en datos biomédicos superó en rendimiento en tareas específicas de salud. El entrenamiento del modelo se llevó a cabo con un conjunto de datos balanceado y se realizaron ajustes de hiperparámetros para obtener resultados satisfactorios.

Los resultados obtenidos del modelo NLP mostraron un buen desempeño con métricas como precisión, recall y F1 score en el conjunto de prueba. La implementación de la matriz de confusión proporcionó una comprensión detallada del rendimiento del modelo, facilitando la interpretación de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos.

La consolidación de resultados en Power BI permitió visualizar de manera efectiva la distribución de informes por paciente, la lista de banderas rojas identificadas y los resultados del modelo NLP. Esta herramienta facilitó la presentación de resultados de manera clara y comprensible para los usuarios finales.

En resumen, el desarrollo de esta solución integral ha sentado las bases para una herramienta valiosa en la detección temprana de pacientes con indicadores de riesgo de cáncer, proporcionando

un enfoque eficiente que combina tanto datos estructurados como no estructurados. Este enfoque tiene el potencial de mejorar la eficacia en la derivación de pacientes a asesoría genética oncológica, contribuyendo así a una atención médica más personalizada y efectiva.

5.2. Trabajos futuros

Durante el desarrollo del trabajo se implementó una solución que utiliza técnicas avanzadas de NLP, para identificar pacientes derivables a AGO. No obstante, como es común en soluciones basadas en IA y análisis de datos, existen diversas áreas de mejora y expansión que podrían ser abordadas en futuras investigaciones. Estas mejoras permitirían optimizar el rendimiento, ampliar su aplicabilidad y facilitar la integración en entornos médicos reales. A continuación, se presentan algunas propuestas:

- Usar más datos: A pesar de tener un universo inicial de 52.257 datos, al balancearlos, los datos se reducen a menos de la mitad, es por esto que se podrían solicitar más datos para que al momento de balancear, podamos obtener una mayor cantidad de información y poder abordar más casos de pacientes.
- Explorar hiperparámetros: Si bien se utilizaron, una buena opción podría ser investigar otros o modificar los usados en el entrenamiento del modelo, para analizar y comparar nuevos resultados con los ya obtenidos.
- Investigar otros modelos NLP: El modelo BioBERT fue la mejor opción disponible al momento de realizar el trabajo. Sin embargo, en la actualidad se podrían encontrar modelos que puedan tener un mejor rendimiento, esto podría ser abordado en una futura investigación.
- Implementación del modelo: Ya con el modelo entrenado y exportado, se podría implementar en un servidor web o aplicación de escritorio, para ser utilizado tomando como entrada los datos clínicos de pacientes.

Referencias

[1] Ministerio de Salud, Chile. *SAS Visual Analytics - Informes DEIS*. Accedido el 29 de noviembre de 2024. 2024. URL: https://informesdeis.minsal.cl/SASVisualAnalytics/?reportUri=%2Freports%2Freports%2Fbcf6e81f-d7f9-4f69-8703-9a83c3eb5da9&sectionIndex=0&sso_guest=true&reportViewOnly=true&reportContextBar=false&sas-welcome=false.

[2] Guido Van Rossum y Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[3] Thomas Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows". En: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. por F. Loizides y B. Schmidt. IOS Press. 2016, págs. 87-90.

[4] Daly MB, Pal T, Berry MP, et al. "Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic". es. En: *NCCN Clinical Practice Guidelines in Oncology*. 19,1 (ene. de 2021), 77-102, crossref = 10.4067/S0034-98872020000600858. URL: https://doi.org/10.6004/jnccn.2021.0001.

[5] Spector Elaine, Behlmann Andrea, Kronquist Kathryn, Rose Nancy C., Lyon Elaine, Reddi Honey V. "Laboratory testing for fragile X, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG))". es. En: *Genetics in Medicine* 23 (mayo de 2021), págs. 799-812. URL: https://doi.org/10.1038/s41436-021-01115-y.

[6] Weiss JM, Gupta S, Burke CA, et al. "NCCN Guidelines® Insights: Genetic/Familial High-Risk Assessment: Colorectal". es. En: *Journal of the National Comprehensive Cancer Network : JNCCN* 19,10 (oct. de 2021), págs. 1122-1132. URL: https://doi.org/10.1164/jnccn.2021.0048.

[7] Cai, Tianrun and Giannopoulos, Andreas A. and Yu, Sheng and Kelil, Tatiana and Ripley, Beth and Kumamaru, Kanako K. and Rybicki, Frank J. and Mitsouras, Dimitrios. "Natural Language Processing Technologies in Radiology Research and Clinical Applications". En: *RadioGraphics* 36.1 (2016). PMID: 26761536, págs. 176-191. DOI: 10.1148/rg.2016150080. URL: https://doi.org/10.1148/rg.2016150080.

[8] Rani Horev. "BERT Explained: State of the art language model for NLP". es. En: (). URL: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

[9] Sicheng Zhou et al. "CancerBERT: a BERT model for Extracting Breast Cancer Phenotypes from Electronic Health Records". En: (2022). arXiv: 2108.11303 [cs.IR].

[10] Shahrukh Raza y Benjamin Schwartz. "Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach". En: *BMC Medical Informatics and Decision Making* 23.1 (2023), pág. 20. DOI: 10.1186/s12911-023-02117-3. URL: https://doi.org/10.1186/s12911-023-02117-3.

[11] Zhiheng Huang, Wei Xu y Kai Yu. "Bidirectional LSTM-CRF Models for Sequence Tagging". En: (2015). arXiv: 1508.01991 [cs.CL].

[12] "spaCy: Industrial-strength Natural Language Processing in Python". En: (2020). URL: https://spacy.io.

[13] Steven Loria. "TextBlob: Simplified Text Processing". En: *TextBlob Documentation* 0.15.3 (2018). Accessed: 2024-09-28. URL: https://textblob.readthedocs.io/en/dev/.

[14] Hannah Eyre et al. "Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python". En: (2021). arXiv: 2106.07799 [cs.CL].

[15] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". En: *Bioinformatics* 36.4 (sep. de 2019), págs. 1234-1240. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz682. eprint: https://academic.oup.com/bioinformatics/article-pdf/36/4/1234/48983216/bioinformatics_36_4_1234.pdf. URL: https://doi.org/10.1093/bioinformatics/btz682.