UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

Graduado en Ingeniería de la Salud

Clasificación de Neoplasias mediante procesamiento de textos para el estudio de Supervivencia poblacional

Classification of Neoplasms using text processing for the study of population Survival

Realizado por
Alejandro Pascual Mellado

Tutorizado por
José Manuel Jerez Aragonés
Francisco Javier Moreno Barea

Departamento
Lenguajes y Ciencias de la Computación

MÁLAGA, junio de 2024

UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA
UNIVERSIDAD DE MÁLAGA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

GRADUADO EN INGENIERÍA DE LA SALUD

Clasificación de neoplasias mediante procesamiento de textos para el estudio de supervivencia poblacional

Classification of neoplasms using text processing for the study of population survival

Realizado por
Alejandro Pascual Mellado

Tutorizado por
José Manuel Jerez Aragonés
Francisco Javier Moreno Barea

Departamento
Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA
MÁLAGA, JUNIO DE 2024

Fecha defensa: julio de 2024

Abstract

Clinical information in Spanish health systems is mostly stored as unstructured text in electronic medical records. Extracting important data from these documents is crucial, especially in oncology where identifying information such as TNM, hormone receptor values, the possibility of recurrence or the location of neoplasms is vital. In this study, we will develop natural language processing (NLP) methodologies based on transformer models to extract this information.

Another objective of this work is to compare the results with those previously obtained using machine learning techniques, it will briefly explain the techniques in question and the results obtained for the problem.

At the same time, this work clearly explains the development process and the functionalities offered by an application that has been designed to provide statistical information on Real World Data (RWD) in hospitals in association with the Computational Intelligence in Biomedicine (ICB) group and the Spanish Society of Medical Oncology (SEOM). The main statistical analysis to be provided is the Kaplan-Maier survival analysis, although some more general ones will be developed.

Keywords: **Classification, Transformers, Text mining, Survival**

Resumen

La información clínica de los sistemas de salud en España se guarda mayormente como texto no estructurado en historias clínicas electrónicas. Extraer datos importantes de estos documentos es crucial, especialmente en oncología donde identificar información como el TNM, los valores de los receptores hormonales, la posibilidad de una recidiva o la ubicación de las neoplasias es de carácter vital. En este estudio, desarrollaremos metodologías de procesamiento del lenguaje natural (PLN) basadas en modelos transformer para extraer esta información.

Otro de los objetivos de este trabajo es comparar los resultados con los previamente obtenidos usando técnicas de machine learning, se explicaran brevemente las técnicas en cuestión y los resultados que obtuvieron para el problema.

Paralelamente, en este trabajo se expone de forma clara el proceso de desarrollo y las funcionalidades que ofrece una aplicación que se ha diseñado con objetivo aportar información estadística sobre Real World Data (RWD) en los hospitales que se encuentren en asociación con el grupo de Inteligencia Computacional en Biomedicina (ICB) y la Sociedad Española de Oncología Médica (SEOM). El principal análisis estadístico que se pretende proporcionar es el de supervivencia de Kaplan-Maier, aunque se desarrollaran algunos mas generales.

Palabras clave: **Clasificación, Transformers, Minería de textos, Supervivencia**

Índice

Índice de figuras

Índice de cuadros

1 Introducción

1.1. Motivación

Los sistemas sanitarios en España enfrentan desafíos de sostenibilidad y mejora de la experiencia del paciente, además de nuevos retos como la medicina personalizada, el uso de la inteligencia artificial (IA) y la gestión de datos en tiempo real (Real-World Data, RWD). El RWD se ha vuelto cada vez más importante como un complemento a los ensayos clínicos aleatorizados para mejorar la seguridad y eficacia de los estudios [1]. Es crucial transformar los datos de las historias clínicas electrónicas (HCE) en información tabulada para la toma de decisiones clínicas [2].

La implementación de sistemas informáticos en el ámbito sanitario que conviertan los datos de las HCE en información tabulada enfrenta varios desafíos, especialmente debido a la diversa naturaleza de estos datos. Las HCE contienen información no estructurada de distintas fuentes, lo que dificulta la extracción automática de conceptos relevantes y hace que la extracción manual sea costosa y no reutilizable [3].

La investigación llevada a cabo está centrada en la oncología, específicamente en la extracción de la localización de neoplasias a partir de HCE. Esta tarea es particularmente relevante porque el grupo de Inteligencia Computacional en Biomedicina (ICB) lleva más de 15 años colaborando con la Unidad de Gestión Clínica Intercentros de Oncología Médica de Málaga (UGCIO). Juntos desarrollaron el sistema Galén [4, 3], que integra HCE y otra información clínica para la investigación oncológica. Esta colaboración ha llevado a la oportunidad para el grupo de trabajar con informes reales de pacientes de oncología, bajo la supervisión y cooperación de profesionales del sector sanitario

Un problema común en los servicios de oncología es la falta de tiempo del personal clínico para completar la información de los pacientes, como la localización de la neoplasia, que a

menudo se encuentra en formato de texto no estructurado en las HCE. El sistema Galén [4, 3] permite asignar esta información a campos específicos, proporcionando un entorno adecuado para obtener conjuntos de datos etiquetados para desarrollar modelos de IA que automaticen esta tarea.

Además de la extracción de la neoplasia de los informes mediante técnicas de PLN, una de las principales motivaciones de este estudio es el desarrollo de una aplicación que pueda extraer información relevante de distintos hospitales aplicando los modelos desarrollados para inferir la localización de la neoplasia y aportar información estadística mediante distintas herramientas gráficas o de manipulación de la información

1.2. Objetivos

1. Aplicar modelos de vanguardia en PLN para la extracción de la localización de neoplasias según la clasificación CIE-10-ES a partir de HCE en español.
    - Implementar modelos avanzados de PLN, específicamente transformers como BERT, para identificar y extraer información sobre la localización de neoplasias en los registros de HCE.
    - Evaluar la precisión y eficiencia de estos modelos en el contexto específico de la terminología médica y la clasificación CIE-10-ES.
2. Comparar los resultados obtenidos mediante transformers con los resultados obtenidos previamente mediante técnicas de machine learning.
    - Realizar una comparación cuantitativa y cualitativa de los resultados de extracción de información utilizando transformers frente a métodos tradicionales de machine learning.
    - Analizar las ventajas y desventajas de cada enfoque en términos de precisión, tiempo de procesamiento y adaptabilidad a nuevos datos.
3. Desarrollar una aplicación para la extracción y análisis de información de HCE de distintos hospitales, proporcionando información estadística sobre los datos obtenidos, especialmente sobre la supervivencia poblacional.

- Crear una herramienta de software que integre los modelos de PLN para procesar y extraer datos de HCE de diversas instituciones hospitalarias.
- Implementar funcionalidades que permitan el análisis estadístico de los datos extraídos, enfocándose en métricas de supervivencia y otros indicadores clínicos relevantes.
- Facilitar la visualización y el reporte de resultados para apoyar la investigación médica y la toma de decisiones en salud pública.

1.3. Estructura del documento

1. **Introducción**
   - Se presenta el contexto general del TFG, describiendo la motivación, los objetivos principales y la relevancia del tema. Se proporciona una visión clara de la problemática abordada y los beneficios esperados del proyecto.
2. **Estado del Arte y Tecnologías Usadas**
   - Este apartado revisa estudios y desarrollos previos sobre PLN en el ámbito sanitario. Se analizan investigaciones y tecnologías relevantes.
3. **Metodología**
   a) **Conjunto de datos**
      - Se describe el origen y las características de los datos utilizados, incluyendo etiquetado, limpieza y transformaciones necesarias para su análisis.

   b) **Investigación con Modelos transformer**
      - Se detalla la elección y relevancia de los modelos transformer, su implementación y entrenamiento, especificando la arquitectura y los hiperparámetros utilizados.

   c) **Análisis de Estudios Previos en Machine Learning**
      - Se analiza comparativamente estudios previos del grupo de investigación ICB de la Universidad de Málaga, discutiendo metodologías y resultados obtenidos para compararlos con los modelos transformer.

d) **Desarrollo de la Aplicación**

- Se describe el proceso de desarrollo de la aplicación propuesta, incluyendo diseño, implementación, pruebas y validación. Se mencionan desafíos técnicos y sus soluciones.

4. **Resultados**

- Se presentan los hallazgos y resultados obtenidos tras aplicar la metodología, con análisis cuantitativos y cualitativos. Se incluyen gráficos, tablas y otras visualizaciones de datos, discutiendo las implicaciones de los resultados.

5. **Conclusiones**

- Se resumen los hallazgos del TFG, destacando conclusiones relevantes y su impacto en el campo de estudio. Se discuten las limitaciones del trabajo y posibles líneas de investigación futura.

6. **Apéndice de Instalación de la Aplicación**

- Se proporcionan instrucciones detalladas para la instalación y configuración de la aplicación desarrollada.

2

Estado del arte y tecnologías usadas

2.1. Estado del arte

2.1.1. Extracción de información en informes médicos

La extracción automática de información se realiza a través del PLN, una disciplina de la inteligencia artificial dedicada a interpretar y entender el lenguaje humano. La tarea de localizar neoplasias es análoga a la clasificación de textos, donde se asigna el contenido de los informes a categorías basadas en su significado y semántica.

Los métodos principales de clasificación incluyen enfoques basados en reglas [5], aprendizaje automático (ML) [6], y aprendizaje profundo (DL) [7, 8]. Dentro del DL, se destacan las redes neuronales recurrentes (RNN) y los *Large Language Models* (LLM) basados en arquitecturas transformer [9]. Estos modelos han mostrado ser superiores a los sistemas tradicionales obteniendo resultados que establecen el estado del arte (SOTA) en multitud de tareas del PLN, incluida la extracción de información clave, y que tienen una gran relevancia en el ámbito biomédico [10, 11].

2.1.2. Machine learning, Embeddings y Redes Neuronales

Los enfoques de aprendizaje automático (ML) utilizan algoritmos que permiten a las computadoras aprender de los datos sin ser explícitamente programadas. Dentro de ML, los *embeddings* son representaciones vectoriales de palabras o conceptos que capturan sus significados y relaciones semánticas en un espacio continuo. Estos *embeddings* son esenciales para la mayoría de los modelos de PLN, ya que transforman el texto en una forma que las máquinas pueden procesar más fácilmente.

Las redes neuronales (NN) son un componente clave del DL y están inspiradas en la estructura y funcionamiento del cerebro humano. En el contexto del PLN, las RNN, una clase de NN, son especialmente útiles para manejar secuencias de texto debido a su capacidad de mantener información a lo largo de posiciones temporales o secuenciales. Las RNN han demostrado ser efectivas en tareas como la generación de texto y la traducción automática.

2.1.3. Transformers

Los *Transformers* son una arquitectura de red neuronal introducida en 2017 [9]. A diferencia de las RNN, los transformers no procesan datos de forma secuencial, sino que utilizan un mecanismo de atención que permite a la red enfocarse en diferentes partes del texto simultáneamente. Este enfoque paralelo mejora significativamente la eficiencia y la capacidad de los modelos para captar relaciones a largo plazo en los datos.

Un transformer típico consta de una serie de capas de codificación y decodificación, cada una compuesta por una combinación de mecanismos de autoatención y capas feedforward totalmente conectadas. Los *Large Language Models* (LLM) basados en transformers, como BERT y GPT, han establecido nuevos estándares en el PLN, logrando resultados SOTA en diversas tareas, incluida la extracción de información médica.

2.1.4. Aplicación en oncología

La organización de retos y tareas compartidas de PLN, junto con la adopción creciente de HCE en sistemas de salud a nivel mundial, ha impulsado el incremento de estudios sobre extracción automática de información. Un ejemplo notable es la extracción de códigos ICD-10 [12](la versión internacional del CIE-10) para codificar procedimientos y clasificar causas de mortalidad a partir de notas clínicas. Inicialmente, los modelos utilizados para esta tarea eran basados en reglas y aprendizaje automático [5, 13]. Con la llegada de modelos de DL, se desarrollaron RNN con una mayor capacidad para manejar un número mayor de categorías simultáneamente [14], incluyendo información oncológica [15]. En este contexto, cabe destacar el trabajo del grupo ICB en una versión preliminar del problema tratado en este estudio [8], donde se demostró la eficacia de las RNN para extraer la localización de las tres neoplasias más frecuentes en la sociedad española según la SEOM: mama, pulmón y colón/recto [16].

Actualmente, los modelos de DL basados en transformers representan el SOTA en la tarea

de extracción automática de códigos CIE-10 [17, 18], no solo en documentos en inglés, sino también en otros idiomas como francés, italiano y portugués [19, 20, 21]. Un trabajo reciente [22] entrenó un modelo transformer grande para la tarea de reconocimiento de entidades, logrando localizar la topografía e histología de neoplasias en textos clínicos en inglés con resultados prometedores.

En español, varios proyectos han intentado utilizar los datos no estructurados de las HCE. Sin embargo, la disponibilidad limitada de corpus clínicos anotados hace que esta tarea sea desafiante. Un ejemplo es la tarea CodiEsp del CLEF eHealth 2020, que abordaba la codificación automática CIE-10 de términos relacionados con diagnósticos y procedimientos en un corpus artificial de HCE en español. En CodiEsp, se evaluaron diversas estrategias de aprendizaje profundo, logrando modelos basados en transformers resultados prometedores [23, 24]. Otro reto relevante es CANTEMIST (*CANcer TExt MIning SharedTask*), enfocado en la extracción de conceptos de cáncer, específicamente la morfología tumoral, en registros médicos españoles artificiales [25]. En este reto, los trabajos realizados por el grupo ICB, que incluyeron el pre-entrenamiento de modelos transformer con un corpus general de Galén, lograron resultados SOTA en las tareas CodiEsp-D y CANTEMIST.

2.2. Tecnologías usadas

Para la clasificación de textos se han utilizado las siguientes tecnologías y herramientas:

- **Python**: Lenguaje de programación ampliamente utilizado en ciencia de datos e inteligencia artificial. Se ha empleado junto con las bibliotecas clásicas de ciencia de datos e IA.

- **VSCode**: Editor de código fuente utilizado para el desarrollo y depuración del código en Python.

- **Hugging Face**: Plataforma que proporciona modelos de PLN avanzados y herramientas para su implementación [26].

- **Modelos transformers**: Modelos de última generación en el campo del PLN, utilizados para tareas de clasificación de textos.

- **Técnicas de Vectorización de Textos**: Se ha empleado la técnica de Bag of Words, específicamente TF-IDF (*Term Frequency-Inverse Document Frequency*), para la representación de textos.

- **Scikit-Learn**: Biblioteca de aprendizaje automático en Python. Scikit-Learn [27] proporciona herramientas simples y eficientes para el análisis de datos y la construcción de modelos predictivos, incluyendo clasificación, regresión, agrupamiento y reducción de dimensionalidad.

- **Técnicas de Machine Learning**: Se han aplicado diversos algoritmos de aprendizaje automático, como Naive Bayes, Support Vector Machine (SVM) y XGBoost, para la clasificación de textos.

Para el desarrollo de la aplicación se han utilizado las siguientes tecnologías:

- **Lenguaje de Programación R**: Utilizado para el desarrollo de la lógica y funcionalidad de la aplicación.

- **RStudio**: Entorno de desarrollo integrado (IDE) para R, empleado en la escritura y depuración del código.

- **RShiny**: Framework de R utilizado para el desarrollo de aplicaciones web interactivas.

- **Expresiones regulares**: Secuencias de caracteres que forman un patrón de búsqueda. Utilizadas en programación para la coincidencia y manipulación de cadenas de texto, permiten realizar tareas como búsqueda, extracción y reemplazo de patrones específicos en textos.

- **SQLite**: SQLite es una biblioteca de software que ofrece un sistema de gestión de bases de datos relacional sin servidor, ligero y eficiente. Almacena los datos en archivos de disco ordinarios y es ampliamente utilizada en aplicaciones móviles, navegadores web y software de escritorio por su simplicidad y confiabilidad.

- **Anaconda**: Plataforma para la gestión de entornos virtuales y paquetes, facilitando la instalación y administración de dependencias.

- **Git y GitHub Desktop**: Herramientas de control de versiones utilizadas para gestionar el código fuente del proyecto, permitiendo el seguimiento de cambios y la colaboración en el desarrollo.

3
Metodología

3.1. Conjunto de datos

En esta sección se detalla el conjunto de datos utilizado para el entrenamiento de los modelos. Como se mencionó anteriormente, el equipo de investigación pudo obtener información de alta calidad de manera sencilla gracias a la disponibilidad del sistema Galén [4], el cual recopila datos de más de 60,000 pacientes oncológicos de la UGCIO de Málaga. En particular, se seleccionó un corpus compuesto por 23,704 HCEs que incluían una neoplasia primaria asociada y más de 500 palabras. Cada documento contenía información demográfica, de la primera visita y detalles de los episodios y consultas posteriores. Además, expertos autorizados des-identificaron los documentos para cumplir con la Ley Orgánica Española de Protección de Datos Personales y Garantía de Derechos Digitales (LOPD-GDD)[28]. Esto también asegura que los modelos de PLN no puedan asociar nombres de médicos o centros hospitalarios con ciertas neoplasias, lo cual constituiría un comportamiento no deseable..

Dado que la estandarización de conceptos y códigos es fundamental en la práctica clínica, es recomendable usar el sistema de la *International Statistical Classification of Diseases and Related Health Problems 10th edition* (ICD-10) [29]. Este sistema, con su equivalencia en español CIE-10-ES, incluye una codificación para la localización (topografía) y la histología (morfología) de tumores y neoplasias. Aunque cada posición dentro de la codificación proporciona un grado de información, en este caso solo se usan las posiciones referentes a la localización. Específicamente, las correspondientes a tumores malignos de localización primaria, cuyos códigos están comprendidos entre C00 y C97. Dada la distribución de grandes grupos de tumores del CIE-10-ES y la presencia en el sistema de información Galén, las categorías se han agrupado según se describe en la Tabla 1.En ella se muestra cada código, la localización asignada y su presencia en el corpus seleccionado, tanto en número absoluto (abs) de documentos como

en frecuencia relativa (rel).

Es importante destacar la división realizada entre las localizaciones de colon/recto/ano (C18-C21) y otros órganos digestivos (C15-C26), debido a la gran cantidad de documentos asignados a los primeros y su relevancia en la sociedad española [16]. Además, la categoría SARCS incluye tumores en huesos y cartílagos articulares (C40-C41) y sarcomas en tejidos blandos. Finalmente, se agrupan otras localizaciones dentro de la categoría OTROS, debido al bajo número de documentos pertenecientes a estas en el corpus definido. Concretamente, OTROS engloba tumores con código: C97, con una presencia relativa del 0.014 global; C76-C80, con presencia del 0.01; y C00-C14, C69-C72 y C73-C75, con una presencia conjunta de estas tres localizaciones del 0.007 en el corpus.

Cuadro 1: Agrupación ECIE, localización y presencia de documentos en el conjunto de datos.

		Presencia	
Código	Localización	abs	rel
C15-C26	Órganos digestivos	1837	.0775
C18-C21	Colón, recto y ano	4020	.1696
C30-C39	Órganos respiratorios e intratorácicos	3371	.1422
C43-C44	Piel	661	.0279
C45-C49	Tejidos mesoteliales y tejidos blandos	1462	.0617
C50	Mama	6491	.2738
C51-C58	Órganos genitales femeninos	1335	.0563
C60-C63	Órganos genitales masculinos	1187	.0501
C64-C68	Vías urinarias	870	.0367
C81-C96	Tejido linfático y órg. hematopoyéticos	1133	.0478
SARCS	Sarcoma en huesos y tejidos blandos	592	.0250
OTROS	Otras localizaciones	745	.0314
Total		23704	

3.2. Investigación con modelos transformer

En este trabajo se ha realizado un estudio pionero en la aplicación de modelos transformers para la extracción de información sobre la localización de neoplasias a partir de HCE

reales en español. Para este propósito, se han explorado tanto transformers entrenados con corpus multilingües como aquellos específicamente adaptados al dominio biomédico en español, abarcando tanto usos generales como específicos [30].

La metodología de clasificación se basa en arquitecturas transformer, que emplean el mecanismo de autoatención multicabezal (*multi-head self-attention*) [9]. Este enfoque permite crear representaciones numéricas contextuales de cada palabra de entrada y mejora la eficiencia computacional mediante la paralelización. Los transformers han ganado gran popularidad debido a su capacidad para el aprendizaje por transferencia, preentrenándose en corpus de dominio general y afinándose posteriormente en corpus específicos para tareas concretas [31]. Este enfoque ha demostrado obtener resultados de vanguardia (SOTA) en dominios tanto biomédicos como clínicos [10, 32].

En este trabajo, orientado a la clasificación de textos médicos en español, utilizamos seis modelos distintos basados en transformers con soporte para este idioma:

- **XLM-R**: Esta versión multilingüe de la arquitectura RoBERTa se preentrenó en un corpus CommonCrawl de $2{,}4$TB en 100 idiomas, utilizando un vocabulario multilingüe de aproximadamente $250,000$ tokens [33]. Hemos experimentado con la versión Base, que cuenta con alrededor de 277 millones de parámetros entrenables.

- **XLM-R-Galén**: Adaptación específica del modelo XLM-R Base al dominio clínico, preentrenada con textos clínicos reales sin etiquetar provenientes de Galén [34], para adecuar el modelo a las particularidades del español clínico.

- **RoBERTa-BNE**: Versión en español de la arquitectura RoBERTa, preentrenada en un corpus de 570GB de la Biblioteca Nacional de España (BNE), empleando un vocabulario de $50,000$ tokens [35]. Se utilizó la versión Base con alrededor de 124 millones de parámetros entrenables.

- **RoBERTa-Bio**: Modelo de lenguaje biomédico para español, basado en RoBERTa y preentrenado desde cero en varios corpus biomédico-clínicos en español, incluidos corpus clínicos reales [36]. Utiliza un vocabulario específico del dominio con aproximadamente $52,000$ tokens.

- **RoBERTa-Base-Biomed**: Otro modelo biomédico en español basado en RoBERTa, pre-entrenado en un corpus biomédico-clínico en español recolectado de diversas fuentes públicas, empleando un vocabulario de aproximadamente 960 millones de tokens [37].

- **BETO-Galen**: Versión del modelo BETO, adaptada al dominio clínico mediante pre-entrenamiento con textos clínicos reales sin etiquetar de Galén [34], manteniendo la configuración y vocabulario del BETO original [38].

3.2.1. Arquitectura adaptada al problema

Para abordar el problema de clasificación, desarrollamos una metodología integral afinando los modelos sobre el corpus clínico de Galén [4, 3]. Cada documento médico se tokeniza en secuencias de subpalabras, las cuales son procesadas por los transformers para generar representaciones a nivel de subpalabra. Estas representaciones alimentan una capa lineal de neuronas añadida en la definición del modelo que clasifica las secuencias en las categorías del problema. Durante el entrenamiento, se ajusta tanto el transformer como la capa de neuronas para optimizar su rendimiento en la tarea de clasificación. La Fig. 1 ilustra visualmente esta metodología.

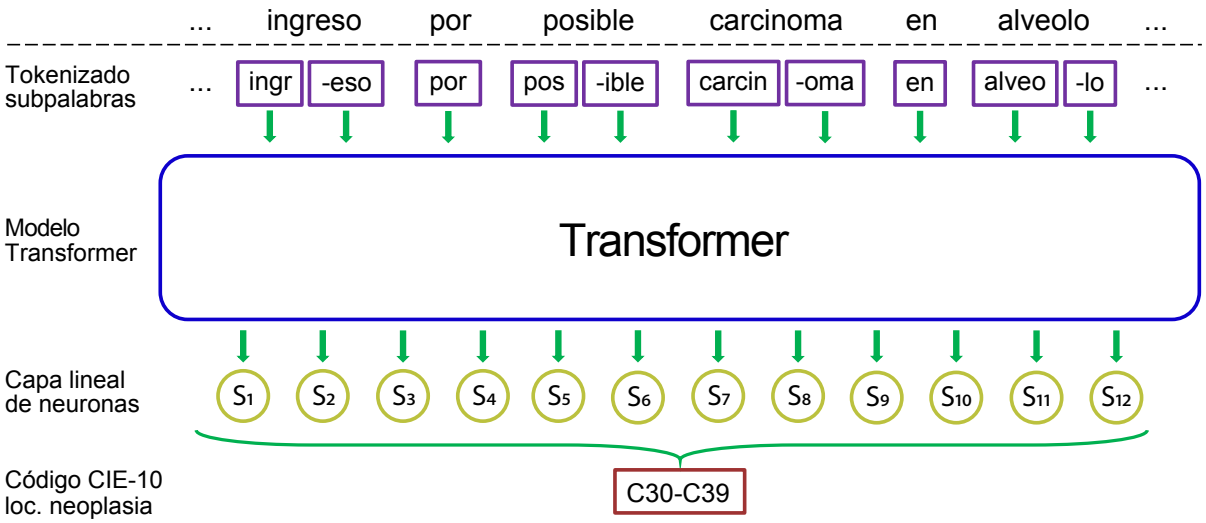


Figura 1: Arquitectura adaptada del transformer, añadiendo una capa lineal de neuronas para el problema de clasificación.

3.2.2. Experimentación

Estrategia de validación

Para llevar a cabo la experimentación, se realizó una división estratificada del corpus en 10 segmentos o *folds*. De estos, 8 segmentos se utilizaron para el entrenamiento, mientras que los 2 restantes se destinaron a la validación y la prueba (*test*) respectivamente. Este enfoque se conoce como cross validation [39] pero en el estudio, por falta de tiempo y recursos, no se realizó un entrenamiento en bucle iterando los segmentos que conforman cada conjunto, si no que se usaron siempre los mismos segmentos para entrenamiento, validación y prueba.

Hiperparametrización

Además, se realizó un ajuste de los hiper parámetros de los modelos, variando el tamaño del lote y la tasa de aprendizaje [40], seleccionando los mejores modelos de cada época basándonos en los resultados obtenidos con el conjunto de validación. Durante el proceso de entrenamiento, se monitorizó el rendimiento en cada época utilizando el conjunto de validación, con el fin de implementar una parada temprana del entrenamiento si era necesario.

Inferencia del conjunto de prueba

Finalmente, se llevó a cabo la inferencia sobre el conjunto de test, obteniendo métricas que reflejan de manera representativa el rendimiento del modelo. Se ha intentado con esta estrategia de experimentación seguir un enfoque completamente honesto, no permitiendo al modelo observar el conjunto de test en ningún momento de su entrenamiento.

Métricas de evaluación

Para evaluar la metodología desarrollada, se emplearon las métricas de porcentaje de aciertos o *accuracy* (Acc), precisión (P), sensibilidad o *recall* (R), y F1-score (F1). La métrica más relevante para este estudio es el F1-score, que es la media armónica de la precisión y la sensibilidad. Esta métrica proporciona una medida confiable del rendimiento de los modelos en problemas donde la sensibilidad es crucial o existe un desbalance significativo entre las clases a predecir.

En este TFG se han utilizado métricas macro para evaluar el rendimiento de los modelos. Las métricas macro son aquellas que calculan la media de las métricas de evaluación para cada clase de manera independiente, sin tener en cuenta el desequilibrio entre clases. Esto es especialmente útil en problemas con clases desbalanceadas, ya que permite obtener una

visión más equilibrada del rendimiento del modelo, asegurando que todas las clases, incluso las menos representadas, sean evaluadas de manera justa.

A continuación se detalla un poco mas en profundidad cada una de las métricas del estudio:

1. **Porcentaje de aciertos o *accuracy* (Acc)**: El porcentaje de aciertos mide la proporción de predicciones correctas sobre el total de predicciones realizadas. Aquí, TP son los verdaderos positivos, TN son los verdaderos negativos, FP son los falsos positivos y FN son los falsos negativos.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

2. **Precisión (P)**: La precisión es la proporción de verdaderos positivos sobre el total de predicciones positivas realizadas por el modelo.

$$P = \frac{TP}{TP + FP} \tag{2}$$

3. **Sensibilidad o *recall* (R)**: La sensibilidad, también conocida como *recall*, es la proporción de verdaderos positivos sobre el total de casos reales positivos.

$$R = \frac{TP}{TP + FN} \tag{3}$$

4. **F1-score (F1)**: El F1-score es la media armónica de la precisión y la sensibilidad. Esta métrica es especialmente útil cuando hay un desbalance entre las clases, ya que proporciona una medida más balanceada del rendimiento del modelo.

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

3.2.3. Tokenizado del conjunto de datos

Procesamiento de Textos con Tokenizadores en Modelos Transformer

Los modelos transformer tienen la capacidad de procesar textos en formato cadena mediante el uso de elementos llamados tokenizadores. Un tokenizador se encarga de dividir los textos en pequeñas unidades llamadas tokens. Estos tokens pueden ser palabras individuales, subpalabras o incluso caracteres, dependiendo del tipo de tokenizador utilizado.

El proceso de tokenización convierte el texto original en una secuencia de tokens numerados que el modelo puede interpretar correctamente. Por ejemplo, la frase "El cáncer de mama

es uno de los tipos de cáncer más comunes" puede ser tokenizada en unidades como ["El", "cáncer", "de", "mam", "a", "es", "uno", "de", "los", "tipos", "de", "cáncer", "más", "comunes"]. Esta secuencia de tokens es esencial para que el modelo transformer entienda y procese el texto, ya que estos modelos no operan directamente sobre las palabras o frases completas, sino sobre estas unidades más pequeñas que luego son convertidas en vectores numéricos en un espacio de alta dimensionalidad.

Límite de Tokens en Modelos transformer

Una de las limitaciones inherentes a los modelos transformer es el número máximo de tokens que pueden procesar por cada documento. Este límite varía entre diferentes modelos, pero en el caso de los modelos que se han utilizado para este trabajo, el límite es de 512 tokens. Esto significa que cualquier documento que, después de ser tokenizado contenga más de 512 tokens, será truncado procesando solo los primeros tokens y descartando el resto.

Esta limitación tiene implicaciones importantes en la precisión de la clasificación de textos médicos según la localización del cáncer, ya que parte de la información puede perderse si el documento original es más largo que el límite permitido. La pérdida de información puede afectar negativamente a la capacidad del modelo para realizar inferencias precisas, sobre todo si hay información relevante para la inferencia en un segmento de texto que ha sido descartado.

Análisis del Impacto del Límite de Tokens

Para evaluar cómo esta limitación afecta al rendimiento del modelo, se registrará el número de tokens de cada documento del conjunto de datos tras su tokenización utilizando el tokenizador del modelo que haya dado mejores resultados en las pruebas preliminares. Este análisis permitirá determinar la distribución de la longitud de los documentos en términos de tokens y observar si hay una correlación entre la longitud del documento y el porcentaje de aciertos del modelo.

El proceso se llevará a cabo de la siguiente manera:

1. **Tokenización del conjunto de datos:** Cada documento del conjunto de datos será tokenizado usando el tokenizador del modelo seleccionado.

2. **Registro del número de tokens:** Se anotará el número de tokens para cada documento.

3. **Evaluación del rendimiento:** Se comparará el rendimiento del modelo (medido como el porcentaje de aciertos en la clasificación) en función de la longitud del documento

tokenizado.

4. **Análisis de correlación:** Se analizará si existe una correlación significativa entre la longitud de los documentos (en tokens) y el rendimiento del modelo.

Este análisis permitirá entender mejor cómo la pérdida de información debido al truncado de tokens afecta la capacidad del modelo para clasificar correctamente los textos médicos en las diferentes localizaciones del cáncer y proporcionará una base para posibles mejoras, como el uso de técnicas de preprocesamiento de texto que reduzcan la longitud de los documentos sin perder información crítica o el reordenamiento de fragmentos de los textos.

3.3. Análisis de estudios previos en Machine learning

En el grupo de investigación ICB, se había llevado a cabo un exhaustivo trabajo previo utilizando el mismo conjunto de datos de HCE para la clasificación y localización de neoplasias. Estos estudios previos emplearon la misma distribución específica de archivos en la validación cruzada para entrenar y evaluar una serie de modelos de *machine learning* clásicos, garantizando así la robustez y la comparabilidad de los resultados obtenidos.

Como parte de este trabajo, se ha estudiado a fondo el funcionamiento y los resultados de estos modelos de machine learning. El análisis incluyó una revisión detallada de las métricas de rendimiento obtenidas en estudios previos. Estos modelos proporcionaron una base sólida para la clasificación de neoplasias en HCE, demostrando capacidades significativas para identificar patrones y características relevantes en los datos.

A pesar de que en la investigación con modelos transformers se han seguido los mismos criterios de experimentación para asegurar la comparabilidad con estos resultados, existen diferencias en el procesamiento de los textos para la interpretabilidad de los modelos (siguiendo un enfoque de vectorización más tradicional) y los parámetros de hiperparametrización (que en este caso dependen del modelo de *machine learning* especifico que se utiliza). Los detalles técnicos específicos serán detallados en las siguientes subsecciones.

3.3.1. Vectorización de Palabras

Para transformar los textos en una representación numérica adecuada para los algoritmos de *machine learning*, se utilizó la vectorización de palabras con un enfoque *bag of word* [41].

En particular, se empleó el esquema TF-IDF (*Term Frequency-Inverse Document Frequency*).

TF-IDF es una técnica que convierte un conjunto de documentos de texto en una matriz de características, donde cada columna representa un término del vocabulario y cada fila representa un documento. La puntuación TF-IDF de un término en un documento se calcula como el producto de dos valores:

- **Term Frequency (TF)**: La frecuencia de un término en un documento específico.
- **Inverse Document Frequency (IDF)**: Una medida de cuán raro es un término en el conjunto de documentos. Se define como el logaritmo del número total de documentos dividido por el número de documentos que contienen el término.

Esta técnica ayuda a reducir el impacto de palabras comunes que no son muy informativas para la clasificación [42].

Los modelos TF-IDF son modelos cuya parametrización depende exclusivamente del máximo número de términos que tiene en cuenta, así que el espacio de hiperparametrización fue un vector numérico con los distintos límites establecidos. Todos los modelos de clasificación que se van a explicar a continuación fueron entrenados con los distintos TF-IDF que se obtuvieron del vector de hiperparametrización, así que este número de términos a tener en cuenta es un parámetro común en los resultados obtenidos por los distintos modelos de clasificación

3.3.2. Algoritmos de Clasificación

Para la clasificación de las historias clínicas electrónicas, se emplearon varios algoritmos de *machine learning*, entre ellos:

- **Naive Bayes**: Un clasificador probabilístico basado en el teorema de Bayes, que asume la independencia entre las características. Es especialmente eficiente para problemas de clasificación de texto debido a su simplicidad y rapidez [43]. Los parámetros con los que se hiperparametrizó son:
  - **Alpha**: Parámetro de suavizado que ayuda a evitar problemas de probabilidad cero en la predicción.
  - **Prior**: Parámetro que permite especificar las probabilidades previas de las clases, si se conocen.

- **Support Vector Machines (SVM)**: Un algoritmo de aprendizaje supervisado que se utiliza tanto para clasificación como para regresión. SVM busca el hiperplano que mejor separa las clases en el espacio de características, maximizando el margen entre las clases más cercanas [44]. Los parámetros con los que se hiperparametrizó son:
    - **True-Kernel-rbf-oc**: Tipo de kernel utilizado, en este caso, *Radial Basis Function* (RBF), que permite manejar relaciones no lineales.
    - **Degree**: Grado del polinomio cuando se usa el kernel polinómico (no aplicable para RBF).
    - **Gamma Scale Coef**: Coeficiente que define cómo una sola muestra afecta el modelo, importante en kernels RBF, polinómicos y sigmoidales.
    - **Weight balanced break ties**: Parámetro que ajusta el peso de las clases para manejar desequilibrios en los datos.
- **XGBoost (*Extreme Gradient Boosting*)**: Un algoritmo de *boosting* que combina la predicción de árboles de decisión y supera significativamente a otros modelos basados en gradientes. Es conocido por su rendimiento y eficiencia en competiciones de *machine learning* y es particularmente útil para conjuntos de datos grandes y complejos [45]. Los parámetros con los que se hiperparametrizó son:
    - **Random Forest**: Método que utiliza XGBoost para crear múltiples árboles de decisión de manera aleatoria.
    - **Num stim**: Número de estimadores, es decir, la cantidad de árboles de decisión en el modelo.
    - **Max depth**: Profundidad máxima de cada árbol de decisión, controlando el sobreajuste.
    - **Min child weigh**: Peso mínimo de la suma de las instancias de un nodo hijo, utilizado para evitar la sobre-segmentación.
    - **Learning rate**: Tasa de aprendizaje que reduce el impacto de cada árbol individual para mejorar la robustez del modelo.
    - **Gamma**: Parámetro de regularización que controla la complejidad del modelo y previene el sobreajuste.

- **Alpha**: Parámetro de regularización L1 que agrega penalización a las características menos importantes.
- **Lambda**: Parámetro de regularización L2 que agrega penalización a los pesos del modelo.
- **Subsample**: Fracción de muestras utilizadas para entrenar cada árbol, ayudando a evitar el sobreajuste.
- **Booster**: Tipo de booster utilizado, puede ser *gbtree*, *gblinear* o *dart*, afectando la forma en que se crean los árboles.

En la sección de resultados se analizaran las mismas métricas que se analizaron en la clasificación mediante modelos transformers, y se hará una comparativa con los mismos teniendo en cuenta que los modelos de *machine learning* son mucho menos costosos computacionalmente.

3.4. Desarrollo de la aplicación

Como último bloque de contenido de este TFG, se ha desarrollado una aplicación para ser instalada en los distintos hospitales que participan en el proyecto de la SEOM.

El grupo de inteligencia computacional en biomedicina tiene un proyecto en marcha financiado por Pfizer y en colaboración con la SEOM, este proyecto consiste en desarrollar una aplicación que pueda ser instalada de forma local en ordenadores de los hospitales que acepten trabajar con el grupo y la SEOM. La principal función de la aplicación es facilitar la tarea administrativa de recopilar y mostrar información que pueda ser relevante en los departamentos de oncología. A continuación se detallaran aspectos importantes sobre la aplicación.

3.4.1. Objetivos generales

Los principales objetivos de la aplicación son los siguiente:

1. **Desarrollar sistema de inicio de sesión por usuarios**: Permitir el acceso a la aplicación solo después de que el usuario haya iniciado sesión con su nombre y contraseña.
2. **Subida de hojas de datos procesadas**: Permitir al personal clínico subir hojas de datos que hayan sido procesadas previamente con las columnas clave que nos comunicaron que serían clave.

3. **Explotar información de HCEs**: Proporcionar un sistema para extraer toda la información que podamos de los HCEs, conociendo previamente la estructura de dichos informes se diseñara un sistema basado en expresiones regulares.

4. **Gestión y análisis de la información**: Facilitar la edición, filtrado y gestión de las entradas de datos subidas por los usuarios.

5. **Generación de gráficas y análisis estadísticos**: Crear visualizaciones gráficas y realizar análisis estadísticos con los datos recopilados que apoyen la práctica clínica y la investigación.

6. **Cálculos de curvas de supervivencia**: Realizar cálculos de curvas de supervivencia de Kaplan-Maier [46] para evaluar la duración de tiempo hasta la ocurrencia de eventos específicos.

7. **Aplicación de métodos de inteligencia artificial para extraer localización de la neoplasia**: Integrar los avanzados modelos de inteligencia artificial descritos en las secciones anteriores para extraer la localización de la neoplasia de cada paciente a raíz de la información textual de los historiales clínicos.

3.4.2. Metodología de desarrollo

Esta aplicación se ha desarrollado siguiendo una metodología ágil, en la cual cada pocas semanas se presentaba el progreso a los clientes para recibir retroalimentación continua. Inicialmente, la aplicación se desarrolla para un solo hospital en asociación con la SEOM, lo que permite un enfoque colaborativo y ajustado a las necesidades reales del entorno clínico.

3.4.3. Aspectos técnicos

El framework seleccionado para el desarrollo de la aplicación fue RShiny. Sin embargo, debido a que la aplicación necesitaba hacer uso de los modelos de clasificación descritos anteriormente, se tuvo que instalar en un entorno virtual de Anaconda que conectaba R y Python mediante la librería reticulate. En este entorno se instalaron también todas las dependencias necesarias para facilitar posteriormente la instalación en cualquier dispositivo. Algunas de

estas dependencias son librerías desarrolladas por el grupo de investigación ICB para el procesamiento de textos y la inferencia de los modelos.

3.4.4. Privacidad y seguridad

La aplicación que se ha desarrollado está diseñada específicamente para ser instalada en un entorno clínico real. Dado que se manejarán datos médicos sensibles, es de vital importancia que la aplicación no tenga ningún tipo de conexión a internet. Está pensada para ser instalada localmente en un ordenador, donde se subirán los datos y se utilizarán todas las herramientas proporcionadas. Es crucial que estos datos no puedan salir del entorno local a través de la aplicación, evitando así cualquier conflicto de privacidad o seguridad. Esto asegura que la información médica se mantenga completamente segura y conforme a las normativas de protección de datos que nos indicó el hospital con el que inicialmente se trabajó.

3.4.5. Principales retos del proyecto

El principal reto del desarrollo de esta aplicación es que está pensada para adaptarse a los diferentes orígenes de datos de los distintos hospitales. Esto es un reto de gran dimensión porque cada hospital tiene los datos organizados de maneras muy diversas. Los HCEs de los departamentos de oncología a veces no siguen una estructura fija y es necesario conocer previamente la estructura de estos informes para poder extraer correctamente los datos que se consideran necesarios para el funcionamiento completo de la aplicación.

El punto positivo del desarrollo actual es que esta aplicación es un prototipo que de momento solo funcionara con uno de los hospitales, por lo que para este TFG se trabajara de una forma mucho mas rígida de lo que en un futuro podrá tolerar el desarrollo de la aplicación.

4

Resultados

En esta sección se presentan los resultados obtenidos con la metodología previamente descrita, así como las principales funciones de la aplicación desarrollada..

4.1. Investigación con modelos transformer

4.1.1. Análisis de las métricas de clasificación

En el Cuadro 2 se muestran los resultados obtenidos por los modelos transformer propuestos para el problema de clasificación. Los mejores valores están resaltados en negrita, y los segundos mejores valores están en cursiva. De los resultados presentados en el Cuadro 2 se puede inferir, por un lado, que los modelos preentrenados con datos biomédicos y clínicos superan en desempeño al modelo de dominio genérico (XLM-R), mejorando el F1-score en casi un 3 %. Por otro lado, los modelos RoBERTa obtienen los resultados más altos en comparación con dos modelos preentrenados en un corpus similar al corpus objetivo (modelos entrenados con datos de Galén). Específicamente, RoBERTa-Base-Biomed logra los mejores valores con un 0.946 en precisión y un 0.908 en F1-score.

Cuadro 2: Métricas de evaluación calculadas en el conjunto de test.Se calculan las métricas de porcentaje de aciertos (Acc), precisión (P), sensitividad (R) y F1-score (F1).

Modelo	Acc	P	R	F1
XLM-R	.9240	.8774	.8787	.8776
XLM-R-Galen	.9350	.8894	.8949	.8913
RoBERTa-Base-bne	.9329	*.8972*	.8887	.8921
RoBERTa-Bio	*.9405*	.8950	**.9061**	*.8987*
RoBERTa-Base-Biomed	**.9456**	**.9204**	*.8982*	**.9080**
BETO-Galen	.9025	.8503	.8421	.8436

Con el objetivo de realizar un análisis detallado del rendimiento de RoBERTa-Base-Biomed, el Cuadro 3 presenta los resultados obtenidos por este modelo para cada una de las 12 categorías CIE-10-ES, agrupadas por clase específica. Además de las métricas de evaluación, se incluye el número de informes en el conjunto de prueba (*Soporte*). El modelo logra un F1-score superior a 0.90 en 8 de las 12 clases. El mejor resultado se obtiene en los documentos correspondientes a la categoría C50 (mama) con un F1-score de 0.995, siendo estos los más abundantes. Las categorías con menor rendimiento son los sarcomas (SARCS) y los agrupados en OTROS, que son de las localizaciones menos frecuentes. Es destacable el buen rendimiento en la categoría C43-C44 (piel) con un F1-score de 0.896, a pesar del reducido número de documentos presentes en el corpus.

Cuadro 3: Métricas de evaluación para cada clase obtenidas por RoBERTa-Base-Biomed

Código	P	R	F1	Soporte
C15-C26	.8698	.9689	.9167	193
C18-C21	.9673	.9698	.9686	397
C30-C39	.9160	.9675	.9410	338
C43-C44	.9667	.9062	.9355	64
C45-C49	.9259	.8681	.8961	144
C50	.9984	.9923	.9954	649
C51-C58	.9362	.9565	.9462	138
C60-C63	.9573	.9492	.9532	118
C64-C68	.9268	.8539	.8889	89
C81-C96	.9640	.9386	.9511	114
SARCS	.8235	.7500	.7850	56
OTROS	.7931	.6571	.7188	70
Promedio	.9204	.8982	.9080	2370

Adicionalmente, se muestra la matriz de confusión representada en la Figura 2 que resultó de la inferencia sobre el conjunto de test. La diagonal claramente marcada muestra que efectivamente el modelo tuvo un buen rendimiento sobre el conjunto de inferencia.

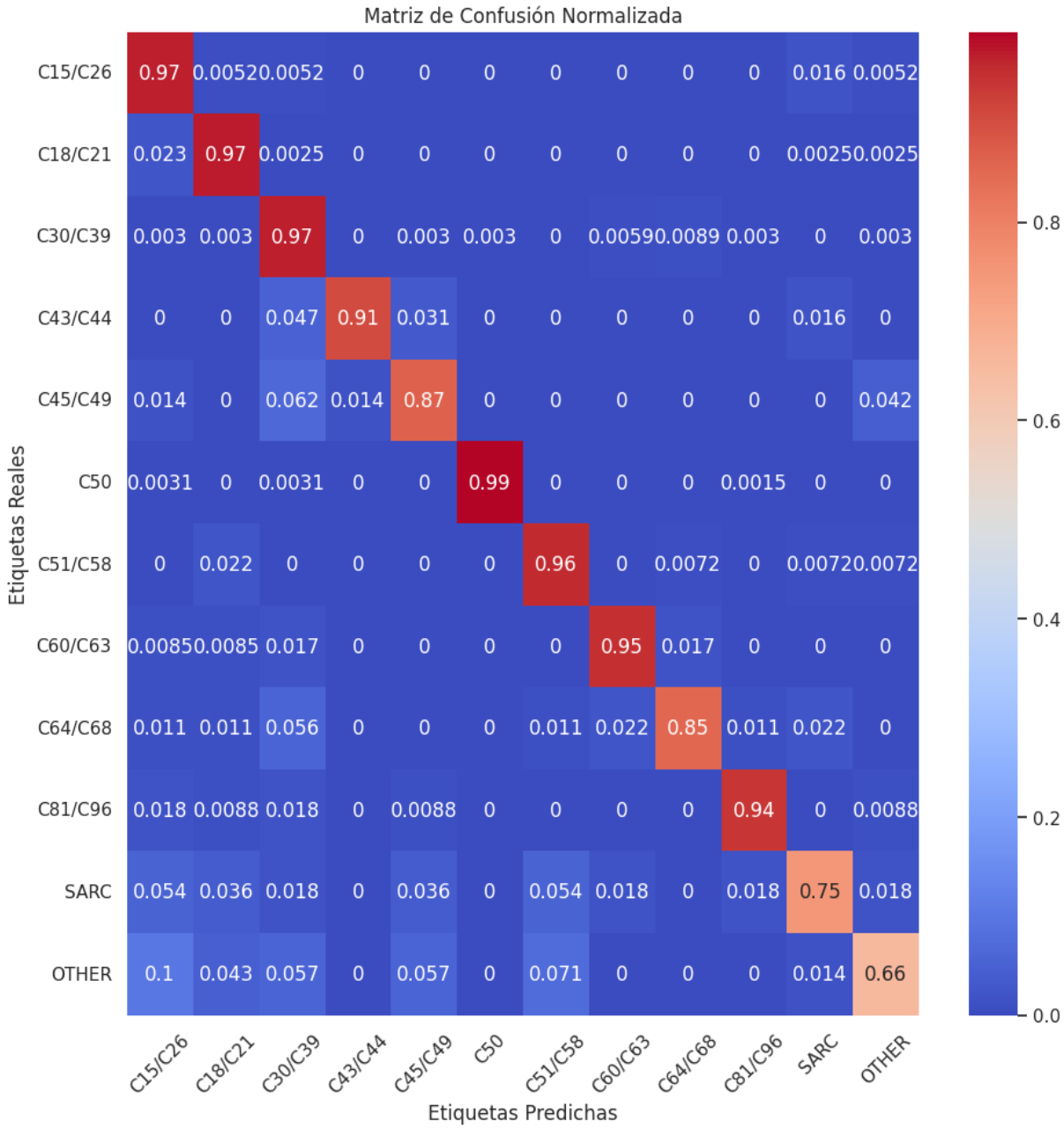


Figura 2: Matriz de confusión obtenida por RoBERTa-Base-Biomed

4.1.2. Análisis del numero de tokens y su impacto en la inferencia

En esta sección se va a realizar un estudio de la cantidad de tokens de los HCE del conjunto de datos. Los HCE serán tokenizados con el tokenizador del RoBERTa-Base-Biomed.

En la Figura 3 podemos observar un histograma con la distribución de longitudes de los archivos y su número de apariciones en el conjunto de datos. La media de longitud medida en

tokens es de 3,199 y la desviación típica es de 3,132, estos son los valores estadísticos para el conjunto de datos en bruto.

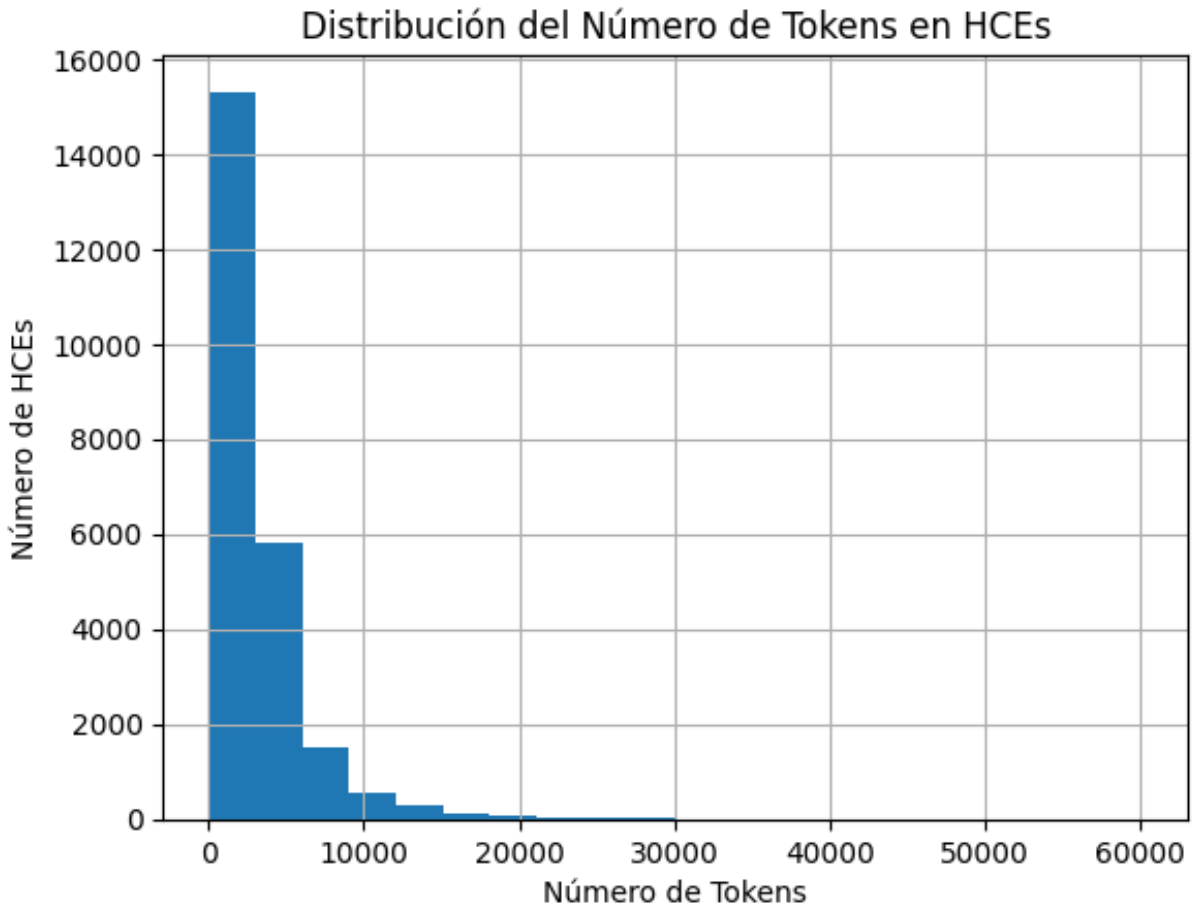


Figura 3: Histograma con la cantidad de historiales por cada longitud de texto en tokens

Sin embargo, podemos observar que en la Figura 3 hay una serie de archivos con gran cantidad de tokens que desbarajustan los resultados a pesar de tener una representación muy escasa en el conjunto de datos, es por esto que también se ha decidido mostrar una imagen similar pero eliminando aquellos historiales que tengan una longitud superior a 10,000 tokens. La Figura 4 muestra este nuevo histograma con el conjunto de datos filtrado. La media se ha reducido a 2,759 tokens y la desviación típica a 1,919.

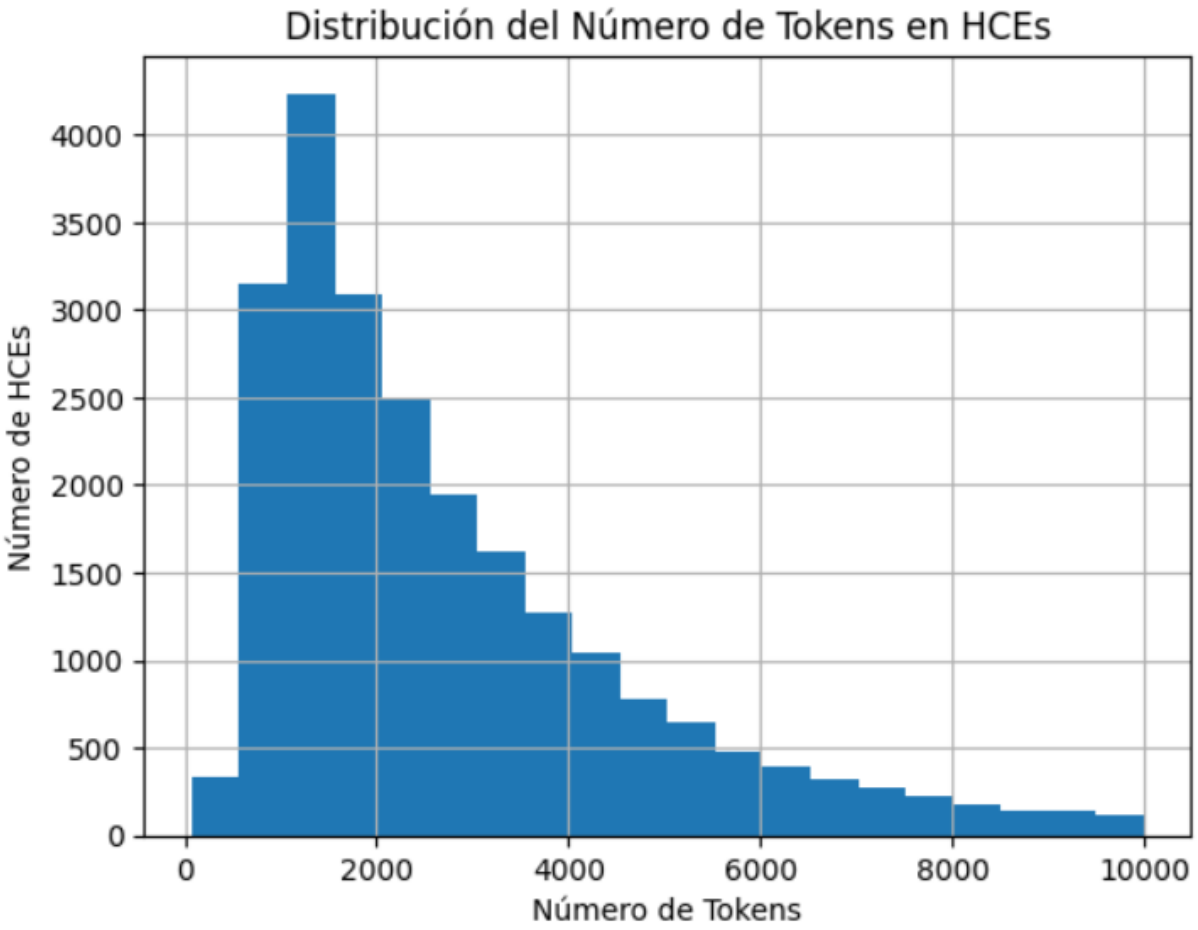


Figura 4: Histograma con la cantidad de historiales por cada longitud de texto en tokens. Se han eliminado los historiales con más de 10,000 tokens.

Observando estas figuras y teniendo en cuenta que el modelo RoBERTa-Base-Biomed solo procesa tanto en el entrenamiento como en la inferencia los primeros 512 tokens, es lógico pensar que se podría aumentar el rendimiento en la clasificación si se le pasase la totalidad de los textos.

Para comprobar si existe una relación entre la longitud de un archivo y el éxito de su inferencia, se ha recopilado el número de tokens de cada informe y una columna binaria que indica si ese informe fue predicho correctamente durante la inferencia. El análisis de correlación que se ha realizado es el de Pearson, este análisis estadístico solo se ha realizado con los informes que componían el conjunto de test, exactamente con 2370 historiales clínicos.

Los resultados de este análisis de correlación han sido esclarecedores pues el valor de co-

rrelación es de *-0.00904* y el p-valor asociado es de *0.65*.

Si la correlación es cercana a 0 y el p-valor es alto (mayor que 0.05), indica que no hay una relación lineal significativa entre la longitud del archivo y el resultado de la predicción. De este modo parece que se puede concluir que no existe una relación entre la longitud del informe y la capacidad del modelo para inferirlo correctamente.

Aún así se añadió a la recopilación previamente citada una columna binaria que indicaba si la longitud del informe era inferior a los 512 tokens que el modelo podía procesar, con el fin de mostrar una gráfica que comparase visualmente la tasa de éxito en la inferencia entre los informes que fueron completamente procesados y los que no.

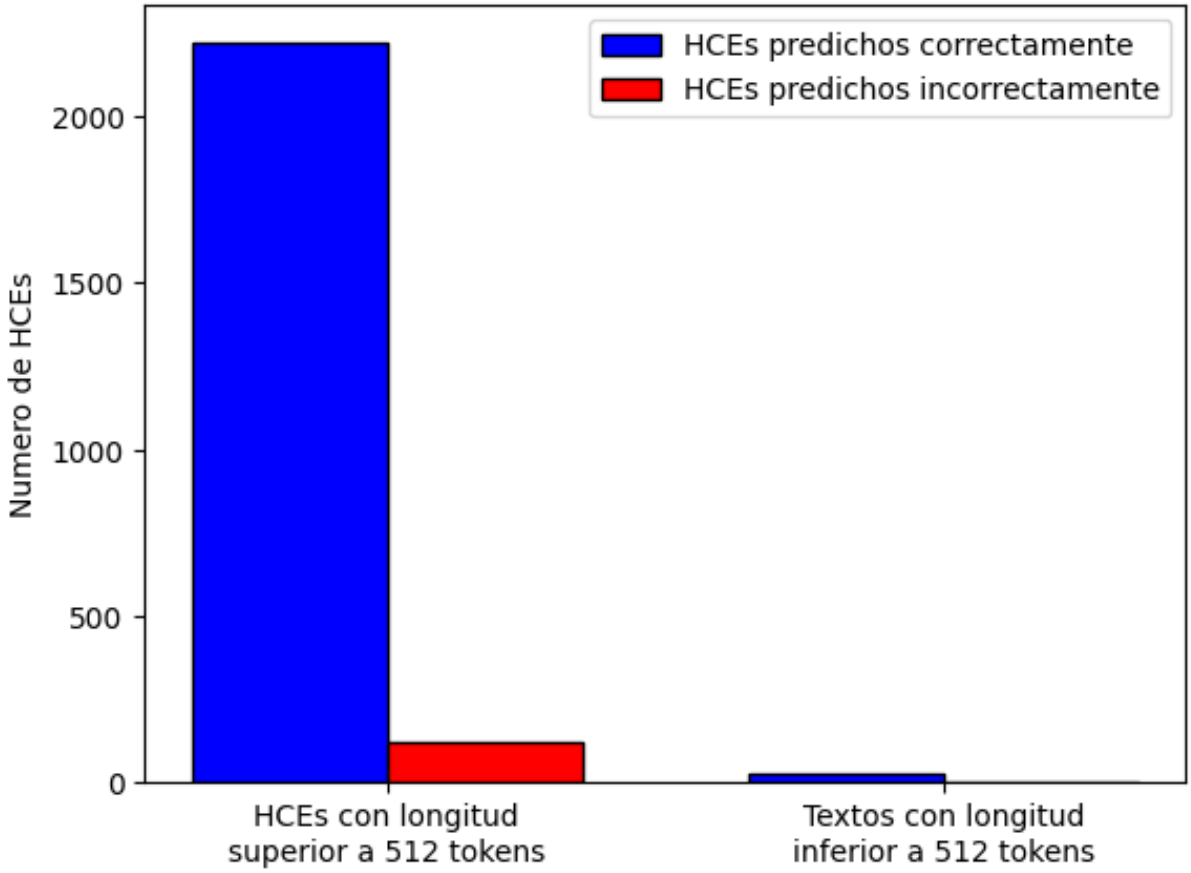


Figura 5: Relación cuantitativa de resultado de la inferencia y el tamaño medido en tokens

En la Figura 5 se observa la gráfica mencionada. A pesar de que solo hay 25 de los 2370 informes del conjunto de test que tienen menos de 512 tokens y que todos fueron predichos correctamente, se puede observar que la proporción de informes con mas de 512 tokens predichos

incorrectamente no es suficientemente alta como para pensar que la perdida de información ha sido relevante.

4.2. Análisis de estudios previos en machine learning

En esta sección se va a realizar una comparativa de los resultados obtenidos en la clasificación de historiales clínicos electrónicos entre las técnicas de machine learning que se han explicado previamente y el modelo RoBERTa-Base-Biomed que hemos concluido que es el mejor modelo de los transformers que se han probado.

Cuadro 4: Métricas de evaluación de las técnicas de *machine learning* y el RoBERTa-Base-Biomed, se calculan las métricas de porcentaje de aciertos (Acc), precisión (P), sensitividad (R) y F1-score (F1).

Modelo	Acc	P	R	F1
Naive bayes	.9409	*.9350*	.8914	.9071
Svm	**.9590**	**.9369**	**.9281**	**.9320**
Xgboost	*.9531*	.9319	*.9157*	*.9219*
RoBERTa-Base-Biomed	.9456	.9204	8982	.9080

El Cuadro 4 muestra claramente que las técnicas de *machine learning*, en las mismas condiciones de experimentación que los modelos de aprendizaje profundo desarrollados para este TFG, han superado al modelo RoBERTa-Base-Biomed y con ello al resto de modelos transformers.

Este mejor rendimiento en las métricas de evaluación sumado al menor coste computacional de los modelos *machine learning* hace a los modelos de machine learning sean mucho mas adecuados que el modelo transformer para la tarea de clasificación de las 12 agrupaciones de neoplasias.

Como se ve en el Cuadro 4 de entre los tres tipos de modelos que el grupo de investigación ICB tenía previamente desarrollados para esta tarea, el mejor en todas las métricas ha sido el sistema SVM.

4.3. Funciones e interfaces de la aplicación

En esta sección, a diferencia de los resultados mostrados previamente, se van a presentar imágenes de las distintas vistas de la aplicación desarrollada. El objetivo es plasmar adecuadamente el resultado final del producto desarrollado mostrando y explicando, una a una, las distintas funcionalidades que proporciona el software. He aquí un listado de respeta el orden en un flujo de uso similar al que debería realizar el usuario.

4.3.1. Interfaz de inicio de sesión

En Figura 6 se muestra la primera imagen que cualquier usuario visualizara al iniciar la aplicación, este sistema de logeo conecta directamente con una base de datos SQLite en la que se almacena una tabla con el nombre de usuario, la contraseña, el hospital al que pertenece dicho usuario y un número de intentos máximos.

Se ha desarrollado un sistema de gestión de errores, que indicará cuando se este intentando entrar con un usuario no registrado o cuando la contraseña proporcionada no sea la correcta. En caso de superar el número de intentos el usuario recibirá un bloqueo y no se le permitirá seguir intentando realizar el inicio de sesión.

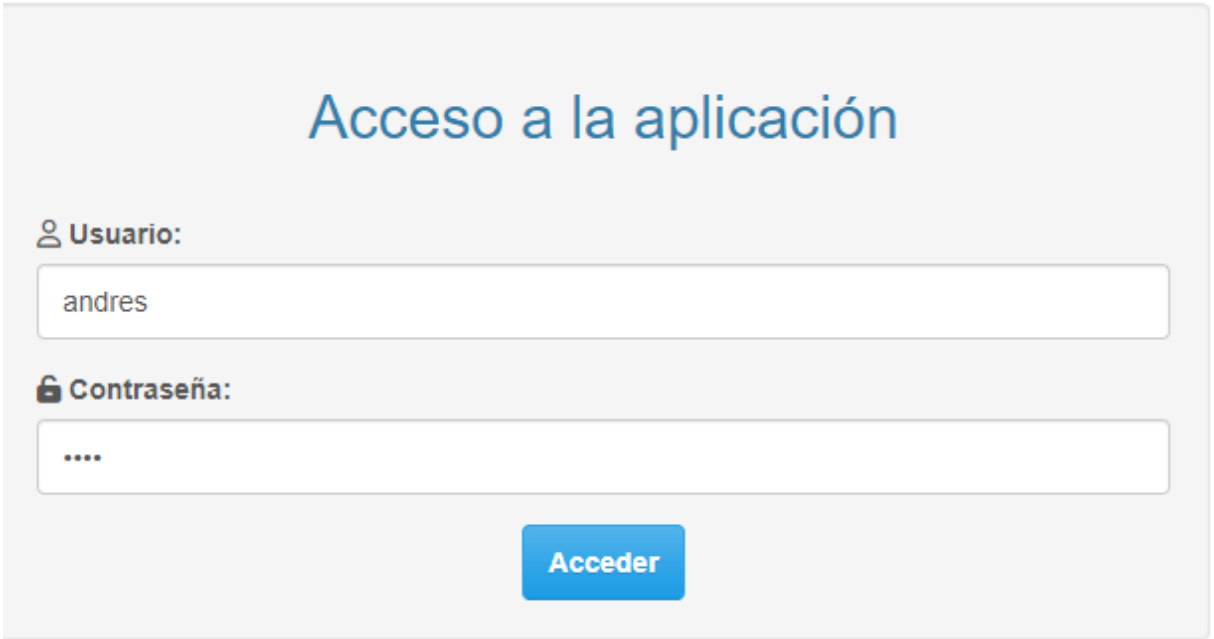


Figura 6: Interfaz de inicio de sesión

4.3.2. Interfaz de subida de hoja de datos

La Figura 7 muestra la interfaz de subida de información vía hoja de datos, dependiendo del hospital asociado al usuario logeado se esperará que esta hoja de datos tenga unas columnas u otras. En esta fase del desarrollo solo hemos establecido contacto con un hospital que nos proporcionó una hoja de datos de ejemplo con la que se ha desarrollado el resto de la aplicación.

Una vez los datos han sido correctamente leídos, se almacenan en un objeto dataframe de R que se quedara guardado de forma local en el directorio donde este instalada la aplicación.

En el caso de volver a subir datos, estos sobrescribirán por completo los anteriormente subidos, esto se le indica al usuario con un modal que aparece cada vez que vuelve a esta interfaz.

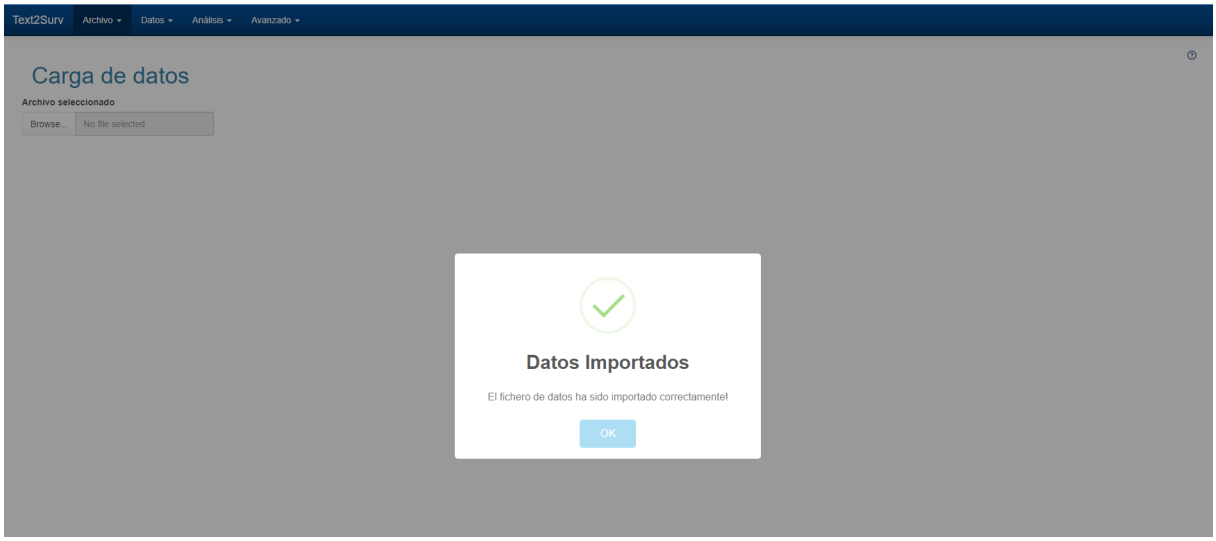


Figura 7: Interfaz de subida de información vía hoja de datos

4.3.3. Interfaz de subida de PDFs

En la Figura 8, se visualiza el flujo que un usuario seguiría para incorporar la información de los pacientes de un archivo zip con historiales clínicos electrónicos en formato PDF. Estos datos se incorporan al volumen de datos que se haya obtenido previamente de la subida de hoja de datos.

Se usan expresiones regulares para capturar campos concretos, esto es posible gracias a que conocemos la estructura de los PDFs que el hospital en cuestión va a subir.

Una vez leídos todos los pdfs y extraída toda la información posible se realiza un mapeo con el número de identificación del paciente o el DNI. Las posibles opciones tras dicho mapeo son que se haya encontrado un paciente con la misma información que estaba en la base de datos (se ha consolidado la información), que se haya encontrado un paciente con diferente información de la que estaba en la base de datos (se sobrescribe dando prioridad a los obtenido del PDF) o que el paciente sea nuevo (se inserta una nueva entrada en la base de datos). Toda esta información se le muestra al usuario en un modal que también indica que se ha finalizado el procesamiento de la información.

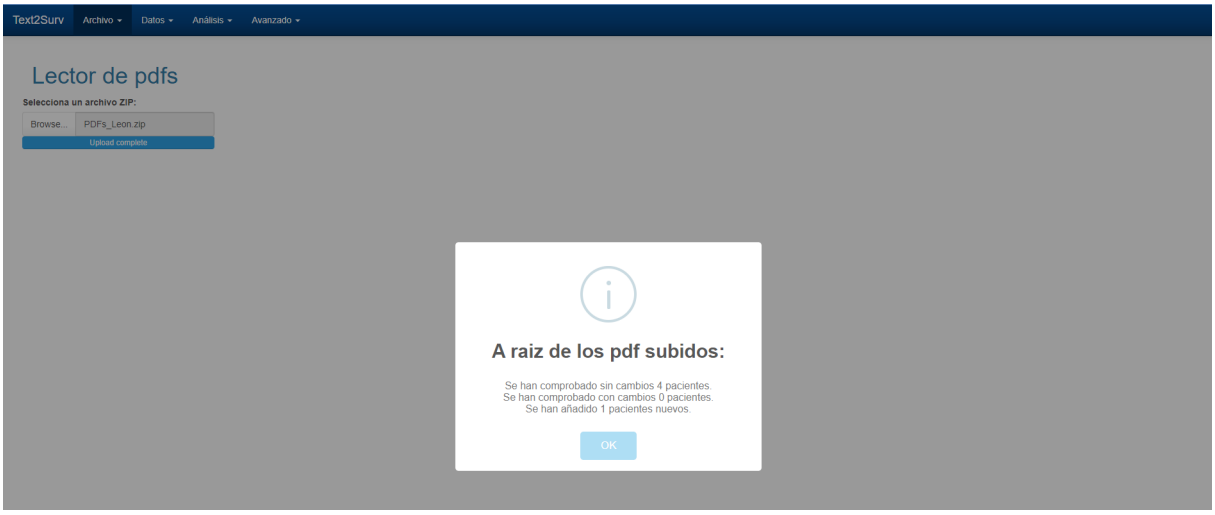


Figura 8: Interfaz de subida de información vía PDF

4.3.4. Interfaz de visualización de datos

Como se observa en la Figura 9, el menú que se ha desarrollado para que el cliente pueda ver la información recopilada consta principalmente de una tabla donde se puede visualizar la información de todos los pacientes por filas.

Los botones que se ven en la imagen 9 son un botón en gris que permite descargar el conjunto en formato Excel y un botón azul que permite modificar la información del paciente seleccionado con la columna que aparece a la derecha de número de orden.

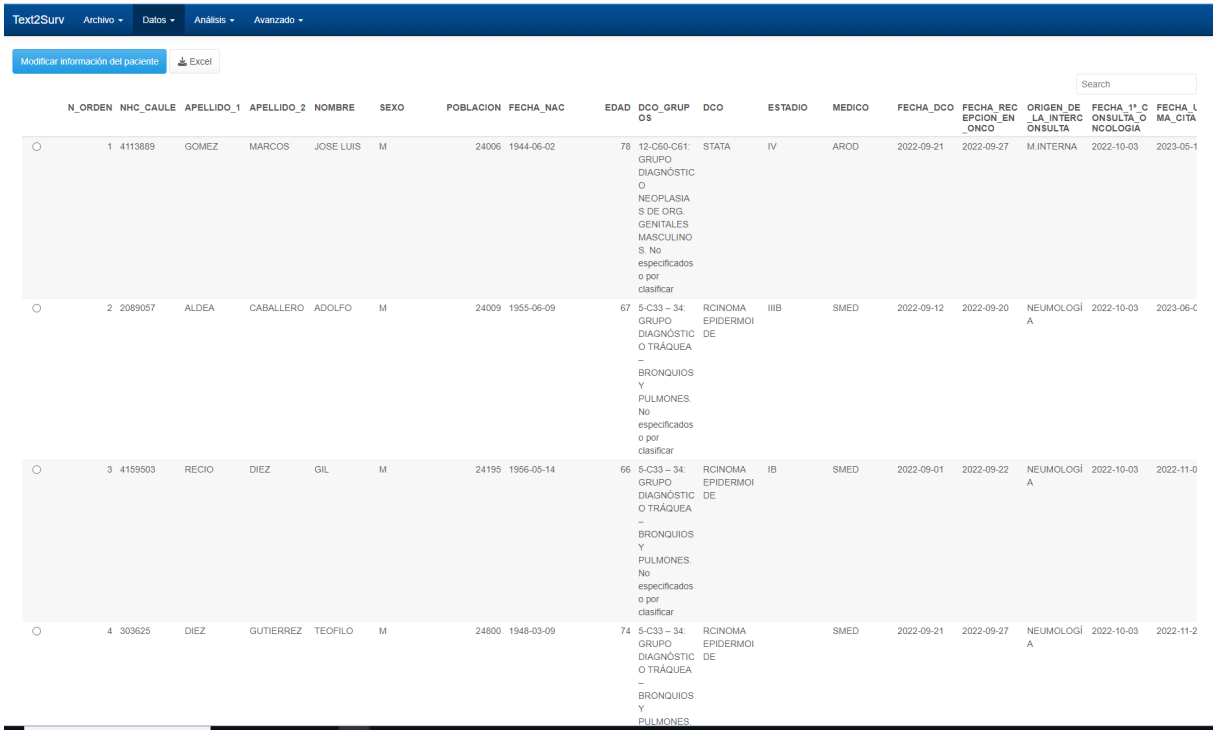


Figura 9: Interfaz de visualización de datos

4.3.5. Interfaz de modificación de datos

La Figura 10 es una imagen del modal para editar una entrada de paciente, en caso de que el usuario quiera modificar la información almacenada de un paciente deberá cambiar el campo deseado y darle al botón de guardar.

Cada campo tiene un widget de Rshiny diferente dependiendo del tipo de columna en cuestión, es decir, los campos de tipo fecha se modificaran con un calendario, los de tipo numérico solo aceptaran números como entrada, los categóricos permitirán seleccionar una de las posibles clases y los de tipo cadena tendrán una entrada de texto.

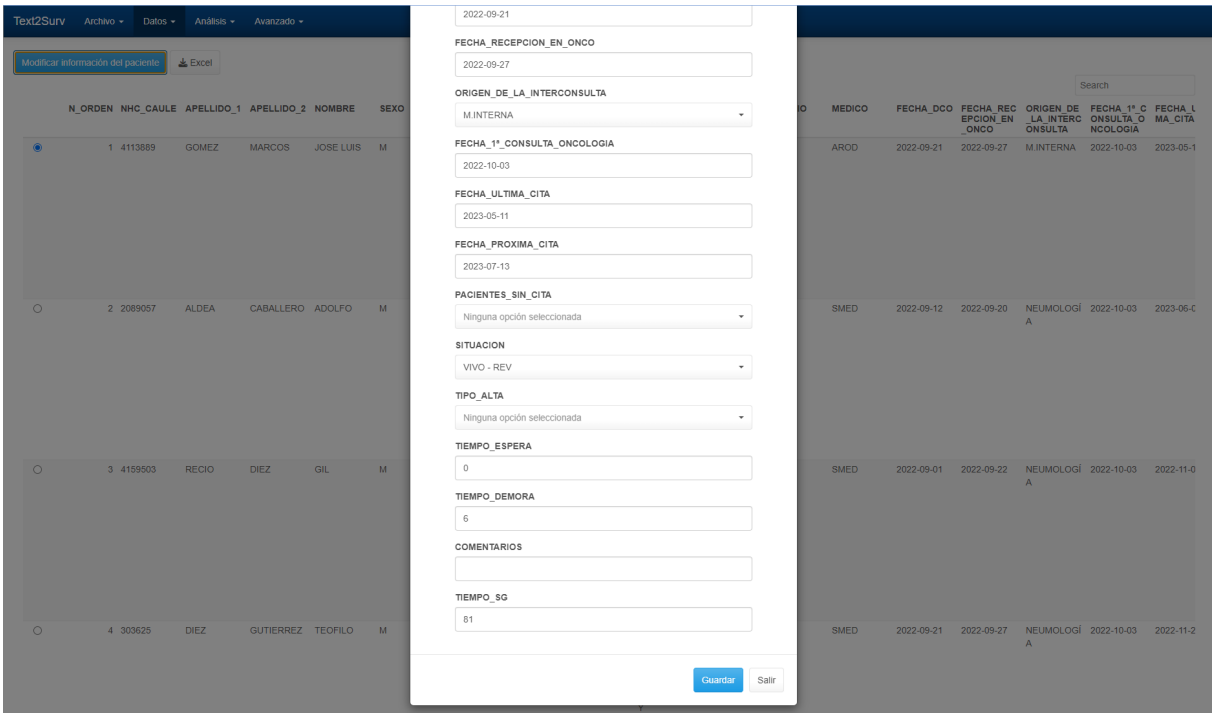


Figura 10: Interfaz del modal de edición

4.3.6. Interfaz de filtrado de datos

En la Figura 11 se puede visualizar el sistema que se ha diseñado para proporcionar la función de filtrado, el usuario podrá seleccionar todos los campos a través de los que quiera filtrar en el widget de la esquina superior izquierda, añadir los parámetros de filtro de cada campo en la división derecha de la ventana (De nuevo dependiendo del tipo de columna se seleccionaran los parámetros de filtrado de una forma u otra) y finalmente aplicar los filtros seleccionados pulsando el botón “Aplicar filtros”.

Se ha desarrollado también un sistema de control de excepciones que evita que el conjunto de datos quede vacio después del filtrado.

Figura 11: Interfaz de filtrado de datos

4.3.7. Interfaz de análisis descriptivo

A continuación se muestra la primera de las opciones de análisis estadístico que han sido desarrolladas, el análisis descriptivo. Esta interfaz consta de un selector de la columna de la que se quiere hacer el análisis descriptivo en un panel lateral situado a la izquierda, en este panel también se pueden cambiar los parámetros de la gráficas (Posteriormente se explicara esta función).

Dependiendo del tipo de la columna seleccionada se mostrará una información u otra en el panel central. He aquí las distintas opciones:

Columnas categóricas y de tipo carácter

En la Figura 12 se muestra una gráficas de barras con las distintas categorías o cadenas de caracteres diferentes en el eje x y su número de apariciones en el conjunto de datos en el eje y.

A la derecha se muestra una tabla que aporta información sobre los valores, su número de apariciones y el porcentaje de representación en el conjunto. Es posible descargar o copiar esta tabla con los botones que aparecen justo debajo de la misma.

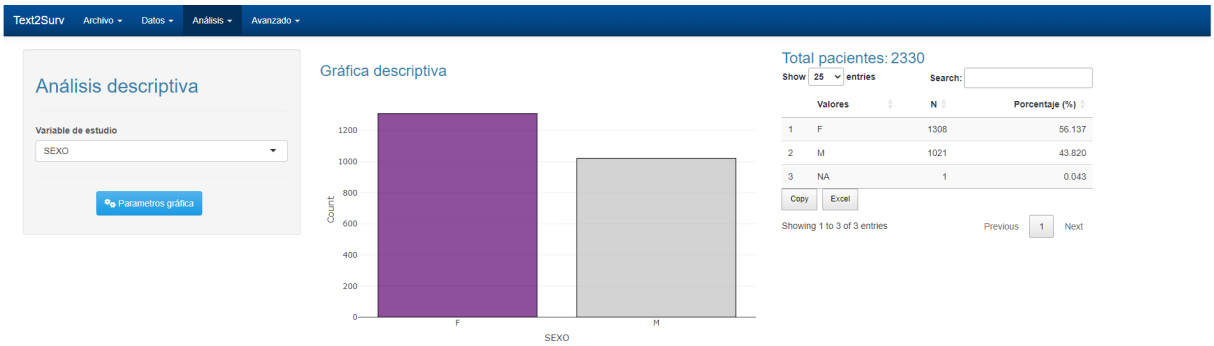


Figura 12: Análisis descriptivo de columnas categóricas o de tipo carácter

Columnas fecha

En la Figura 13 se muestra una linea con una distribución temporal el eje x y su número de apariciones en fechas clave en el eje y.

A la derecha el Cuadro muestra un análisis descriptivo de la columna en cuestión, indicando el valor mínimo, los cuartiles, la media y el máximo . Es posible descargar o copiar esta tabla con los botones que aparecen justo debajo de la misma.

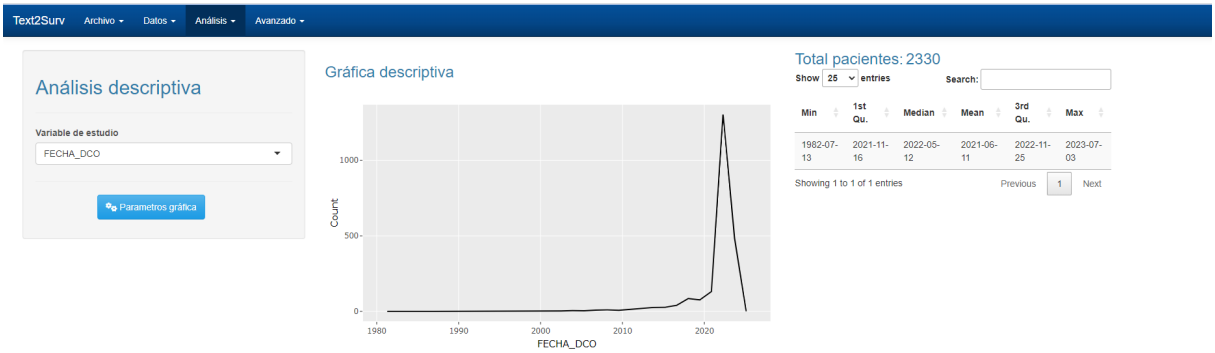


Figura 13: Análisis descriptivo de columnas fecha

Columnas numéricas

En la Figura 14 se visualiza un histograma con la distribución numérica en el eje x y el número de apariciones en el eje y.

A la derecha se muestra una tabla que aporta información sobre los valores, su número

de apariciones y el porcentaje de representación en el conjunto. Es posible descargar o copiar esta tabla con los botones que aparecen justo debajo de la misma.

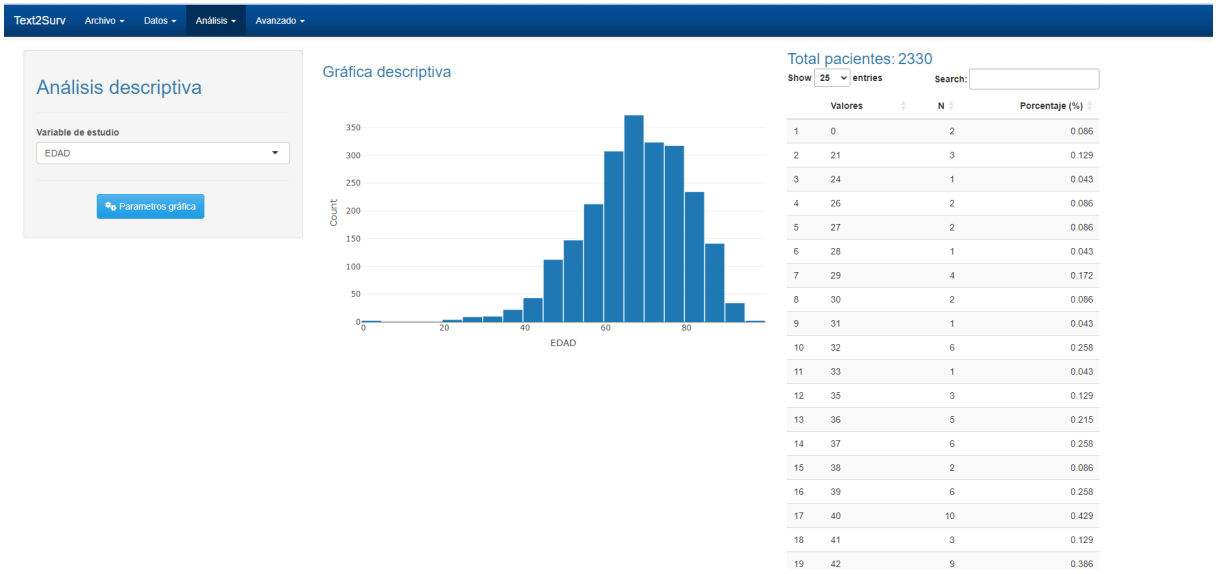


Figura 14: Análisis descriptivo de columnas numéricas

4.3.8. Interfaz de modificación de parámetros de la gráfica

La Figura 15 muestra el modal que aparece en en pantalla tras pulsar el botón parámetros gráfica.

Ofrece la opción de cambiar el titulo de la gráfica o modificar los nombres de los ejes.

Figura 15: Interfaz de modificación de parámetros de la gráfica

4.3.9. Interfaz de análisis de supervivencia

El siguiente análisis estadístico que se ha desarrollado es el análisis de supervivencia con curvas de Kaplan-Maier. En las figuras que se van mostrar a continuación aparece un panel lateral en el que se indicara que variable representa el evento a estudiar (si el paciente ha fallecido o no), que valor de la columna seleccionada representa el evento (fallecimiento) y cual es la variable que representa el tiempo de seguimiento.

Adicionalmente se muestran dos botones que permiten mostrar el intervalo de confianza y el cuadro de pacientes en riesgo. En la Figura 16 se puede observar como se visualizaría el panel principal seleccionando los dos botones previamente citados, además de la curva de supervivencia se muestra una descriptiva estadística del tiempo de seguimiento con los pacientes en riesgo.

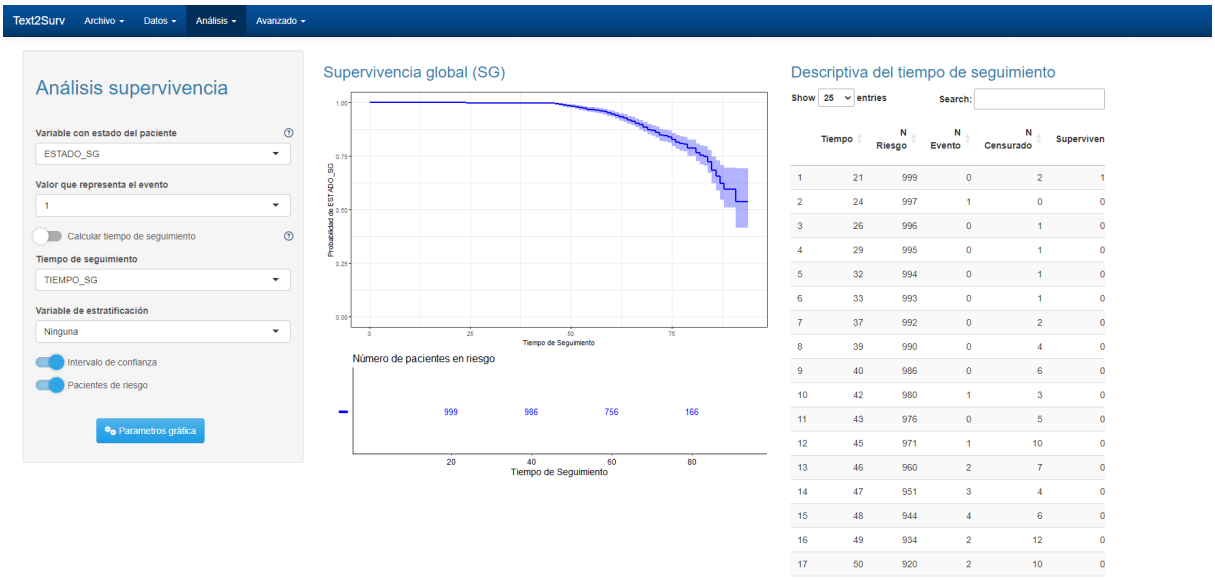
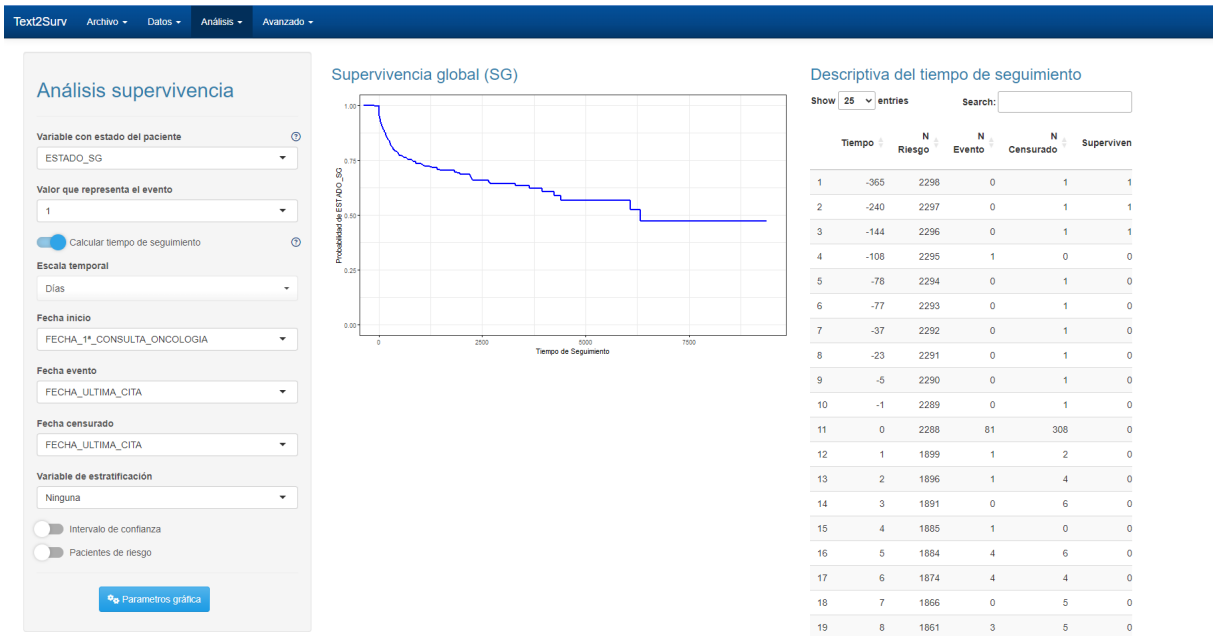


Figura 16: Interfaz de análisis de supervivencia básico

En el caso de que no se dispusiese de una columna que represente al tiempo de estudio se ofrece también la opción de calcularlo solicitando una serie de columnas tipo fecha tras pulsar el botón de calculo del tiempo de supervivencia.

La Figura 17 muestra un ejemplo de esta función de cálculo de tiempo.



	Tiempo	N Riesgo	N Evento	N Censurado	Superviven
1	-365	2298	0	1	1
2	-240	2297	0	1	1
3	-144	2296	0	1	1
4	-108	2295	1	0	0
5	-78	2294	0	1	0
6	-77	2293	0	1	0
7	-37	2292	0	1	0
8	-23	2291	0	1	0
9	-5	2290	0	1	0
10	-1	2289	0	1	0
11	0	2288	81	308	0
12	1	1899	1	2	0
13	2	1896	1	4	0
14	3	1891	0	6	0
15	4	1885	1	0	0
16	5	1884	4	6	0
17	6	1874	4	4	0
18	7	1866	0	5	0
19	8	1861	3	5	0

Figura 17: Interfaz de análisis de supervivencia con cálculo del tiempo

Para finalizar esta función de estudio de la supervivencia, se implementa también en el panel lateral un selector de una columna de estratificación. Esta columna, que debe ser categórica, segmentara el conjunto de datos en tantos subgrupos como categorías existan.

Además de mostrar tantas curvas como categorías estén representadas, también añade un análisis estadístico que proporciona un p-valor para indicar si existen una diferencia significativa en la supervivencia de los distintos subgrupos, esto es de gran importancia para observar que características de un paciente tienen una mayor mortalidad sobre la población de pacientes. La Figura 18 muestra esta última opción

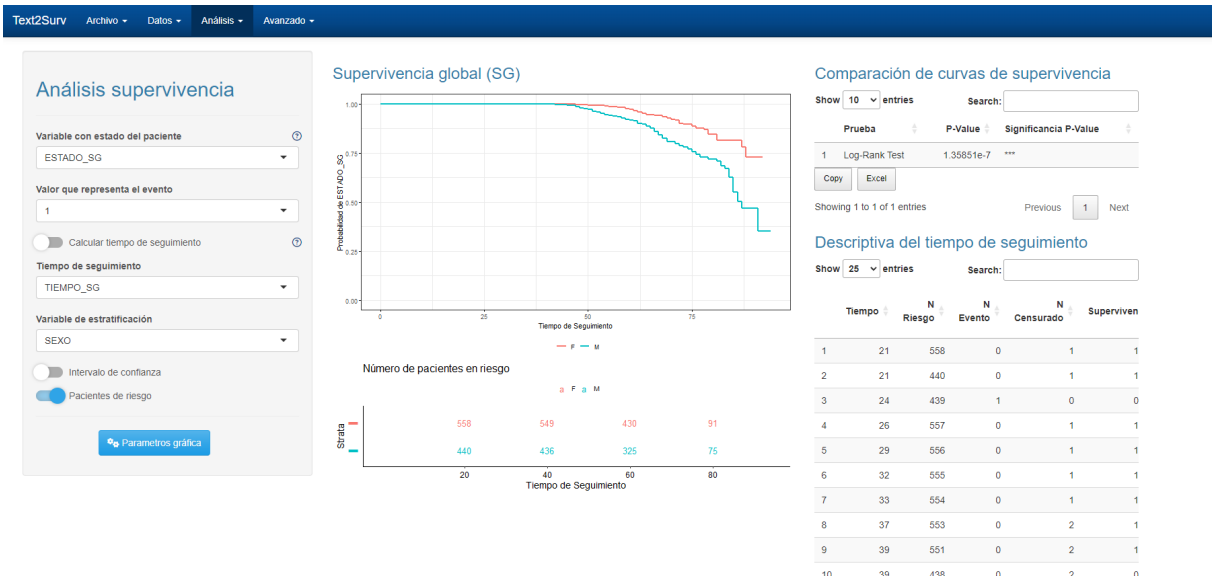


Figura 18: Interfaz de análisis de supervivencia estratificado

4.3.10. Interfaz de inferencia de localización de neoplasias mediante inteligencia artificial

La última funcionalidad que se ha desarrollado es la inferencia de la localización de la neoplasia, que tras pulsar un botón da comienzo al proceso de inferencia que finaliza cuando aparece el modal mostrado en la Figura 19.

Finalmente se ha decidido que esta inferencia se realice con uno de los modelos de Naive Bayes de *machine learning* descritos en la sección anterior. El motivo de usar este modelo y no alguno de los modelos transformers que se han desarrollado para este TFG es que en un entorno portable con un framework como RShiny sería difícil cargar un modelo tan computacionalmente costoso como el RoBERTa, en cambio, el modelo de Naive Bayes es muy ligero y aporta unos resultados mas que aceptables para dejarlo en producción.

Los textos que se usan para inferencia provienen o bien los HCEs subidos vía PDF o bien de algunas columnas de tipo carácter que contienen información sobre el tipo de de enfermedad del paciente, si este no ha sido importado desde archivos PDF sino del excel inicial.

Tras la inferencia se añade una nueva columna categórica llamada Neoplasia al conjunto de datos.

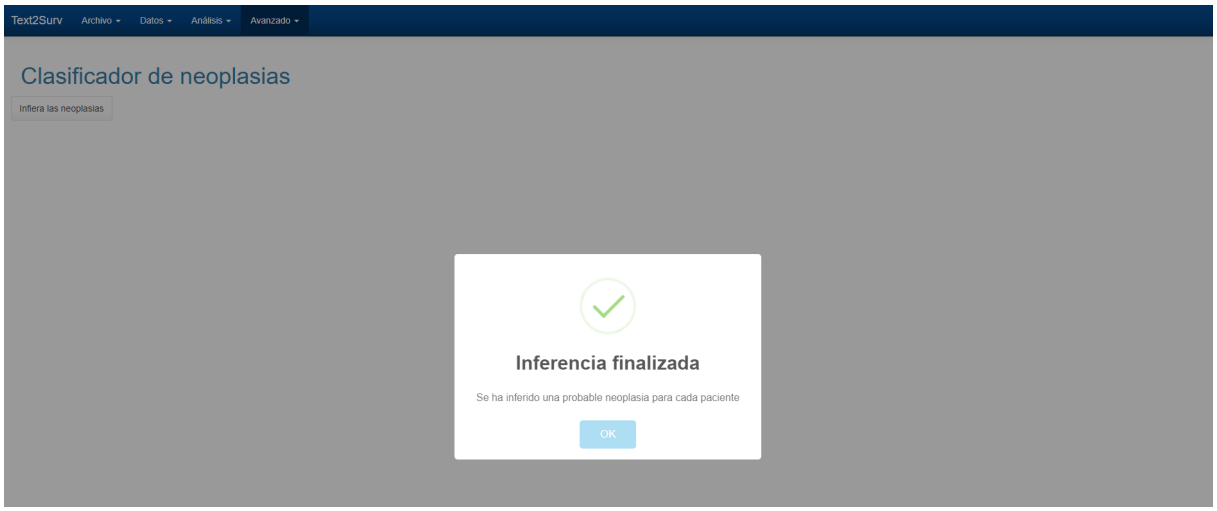


Figura 19: Interfaz de inferencia de localización de neoplasias

5

Conclusiones y Líneas Futuras

5.1. Conclusiones

Investigación con modelos Transfomer

En este trabajo se ha investigado con técnicas de procesamiento de lenguaje natural como extraer la localización de la neoplasia de documentos sanitarios reales redactados en español. El corpus empleado incluye 23,704 historiales clínicos electrónicos del servicio oncológico de Málaga, obtenidos del sistema Galén.

El principal foco de la investigación fue probar los novedososos modelos transformers que están aportando grandes resultados en tantas otras disciplinas del ámbito de la inteligencia artificial.

Se destaca especialmente el rendimiento sobresaliente de los modelos basados en la arquitectura RoBERTa entrenados con textos de dominio clínico y biomédico, en comparación con los modelos de dominio general. En particular, el modelo RoBERTa-Base-Biomed alcanzó una precisión, sensibilidad y F1-score macro-promediadas de 0.920, 0.898 y 0.908, respectivamente. Estos resultados destacan el adecuado rendimiento de los transformers en la extracción de información clínica relevante de los registros médicos.

En cuanto al rendimiento analizado por cada categoría CIE-10-ES, se destaca la clasificación del cáncer de mama (C50), con un F1-score de 0.995. Este rendimiento excepcional podría explicarse por su alta incidencia en el corpus, lo que expone más al modelo a esta localización. Sin embargo, la categoría C43-C44 logró un F1-score de 0.936, a pesar de su menor representación en el corpus ($2{,}8\,\%$). Esto sugiere una capacidad robusta del modelo transformer para extraer correctamente la localización incluso en neoplasias menos frecuentes. Ocho de las doce

codificaciones superaron un valor de 0.9 en sensibilidad y F1-score, mostrando un rendimiento inferior en las categorías SARCS y OTROS. Estas incluyen algunas de las neoplasias más complejas de diagnosticar, como los sarcomas de tejidos blandos y los tumores germinales y de sitios desconocidos.

Uno de los problemas que afrontan los transformers es la perdida de información por truncado de textos, ya que los modelos RoBERTa por ejemplo solo procesan los primeros 512 tokens de un texto. Como la gran mayoría de los textos del conjunto de datos superan esta cantidad se realizó un análisis de correlación de Pearson para evaluar si este problema afectaba significativamente al problema de clasificación.

Finalmente se ha determinado que no existe una relación entre la longitud de l texto y la probabilidad de éxito de la inferencia, esto probablemente se deba a que en muchas de las cabeceras de los informes aparece información relevante sobre la localización de la neoplasia que permite al modelo inferir correctamente su clase.

Análisis de estudios previos en Machine learning

Adicionalmente, se realizó una comparación del modelo transformer con otras técnicas de machine learning. Estos modelos de machine learning que fueron previamente desarrollados siguiendo las mismas estrategias y condiciones que los nuevos transformer han resultado mejorar levemente todas las métricas. En concreto Support Vector Machine obtuvo el mejor resultado en todas las métricas de evaluación, superando hasta en un 3 % algunas de las métricas que obtuvo el RoBERTa-Base-Biomed.

Aunque no se ha detectado un motivo concreto por el que justificar esta derrota de los modelos transformer es muy probable que en problemas de clasificación mas complejos (con un mayor número de clases), los modelos transformer se desenvuelvan mejor que los clásicos modelos de machine learning. Esta hipótesis se debe a que estos modelos transformer tienen la capacidad de crear hiperespacios para alojar los datos procesados con una altísima dimensionalidad, lo que les permite desenvolverse con mucho éxito en tareas mas complejas como reconocimiento de entidades en tareas de calorificación a nivel de token.

Desarrollo de la aplicación

Finalmente, el desarrollo de la aplicación ha resultado en un producto que cumple con las exigencias que el cliente que financia el proyecto solicitaba en un principio.

La aplicación actualmente se ajusta a una entrada de información con formato hoja de

datos que contenga algunas columnas clave, pero es capaz de recopilar el resto de columnas que el usuario quiera añadir y procesarlas de forma automática. Con la subida de informes PDF pasa lo mismo, actualmente siempre que se suba un pdf con una estructura conocida las expresiones regulares desarrolladas serán capaces de capturar la información necesaria.

En cuanto a la inferencia de neoplasias ha resultado éxito pues se extrae de forma automática, lo que permite usar esta información en los análisis estadísticos.

Para finalizar, el desarrollo del análisis de supervivencia ha resultado en una herramienta muy útil que considero que puede ayudar con la gestión de los departamentos de oncología, ya que permite al personal clínico observar características clave sobre la esperanza de vida de sus pacientes de forma estratificada y esto puede desembocar en descubrir departamentos donde invertir mas esfuerzo.

5.2. Líneas Futuras

En futuros trabajos, se planea investigar la extracción de la localización de la neoplasia con un mayor número de categorías, desglosando algunas de las agrupaciones realizadas en este estudio. De este modo se podrá analizar si efectivamente los modelos transformer presentan un rendimiento similar en problemas con mayor número de clases y se podrá comprobar si en este caso superan a los modelos de machine learning

También se intentara afrontar el problema del truncado de información probando modelos longformer, que permiten hasta el procesamiento de 4,000 tokens, o con alguna estrategia de reordenación que permita comprobar si efectivamente la información que viene en la cabecera de los informes de nuestro conjunto de datos es la mas relevante para la extracción de la localización de la neoplasia.

El desarrollo de la aplicación aún debe afrontar su mayor reto, la incorporación de nuevos hospitales al convenio del proyecto y con esto un mayor desarrollo de la adaptabilidad de recogida de datos de la misma.

Apéndice A
Guía de instalación de la aplicación

A.1. Descargar e instalar R y RStudio

- R y RStudio: Instalar R y RStudio

A.2. Descargar e instalar Anaconda

- En Windows: Instalar Anaconda en Windows
- En macOS: Instalar Anaconda en macOS

A.3. Crear entorno Anaconda con Python

- Abrir Anaconda Prompt (Windows) o cmd (macOS).
- Navegar hasta el directorio principal de la aplicación (Text2Surv).
- Crear el entorno con Python 3.9:

```
conda create --prefix=.venv python=3.9
```

A.4. Preparar entorno Anaconda

- Activar el entorno:

```
conda activate "Path"\\.venv
```

- Instalar paquetes necesarios:

```
conda install numpy
conda install pandas
```

- Navegar al directorio donde se encuentren las carpetas con la librería de galen a instalar (`./fragments/`, `./tools/`, etc.) y ejecutar:

```
pip install -e .
```

- Seguir el siguiente orden para las instalaciones:
  - fragments
  - tools
  - brat
  - tokens
  - corpus
  - tnm
- Instalar paquetes adicionales:

```
conda install -c anaconda regex
conda install gensim
conda install scikit-learn
```

A.5. Ejecutar archivo app.R en RStudio

Referencias

[1] H.-G. Eichler, F. Pignatti, B. Schwarzer-Daum, A. Hidalgo-Simon, I. Eichler, P. Arlett, A. Humphreys, S. Vamvakas, N. Brun, and G. Rasi, "Randomized controlled trials versus real world evidence: neither magic nor myth," *Clinical Pharmacology & Therapeutics*, vol. 109, no. 5, pp. 1212–1218, 2021.

[2] J. Concato and J. Corrigan-Curay, "Real-world evidence-where are we now?" *The New England journal of medicine*, vol. 386, no. 18, pp. 1680–1682, 2022.

[3] D. Urda, N. Ribelles, J. L. Subirats, L. Franco, E. Alba, and J. M. Jerez, "Addressing critical issues in the development of an oncology information system," *International journal of medical informatics*, vol. 82, no. 5, pp. 398–407, 2013.

[4] N. Ribelles, J. M. Jerez, D. Urda, J. L. Subirats, A. Márquez, C. Quero, and L. Franco, "Galén: Sistema de información para la gestión y coordinación de procesos en un servicio de oncología," *RevistaeSalud*, vol. 6, no. 21, pp. 1–12, 2010.

[5] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute, "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques," *Journal of the American Medical Informatics Association*, vol. 13, no. 5, pp. 516–525, 2006.

[6] J. St-Maurice, M.-H. Kuo, and P. Gooch, "A proof of concept for assessing emergency room use with primary care data and natural language processing," *Methods of Information in Medicine*, vol. 52, no. 01, pp. 33–42, 2013.

[7] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, "Convolutional recurrent neural networks for text classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–6.

[8] F. J. Moreno-Barea, H. Mesa, N. Ribelles, E. Alba, and J. M. Jerez, "Clinical text classification in cancer real-world data in spanish," in *International Work-Conference on Bioinformatics and Biomedical Engineering*. Springer, 2023, pp. 482–496.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[11] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[12] M. Schuman, Andrew J., "Icd-10: What you need to know," *Contemporary pediatrics*, vol. 32, no. 3, pp. 40–42, Mar 2015 2015/03//, copyright - Copyright Advanstar Communications, Inc. Mar 2015; Última actualización - 2023-09-21. [Online]. Available: https://www.proquest.com/scholarly-journals/icd-10-what-you-need-know/docview/1667196034/se-2

[13] M. Subotin and A. Davis, "A system for predicting icd-10-pcs codes from electronic health records," in *Proceedings of bionlp*, 2014, pp. 59–67.

[14] P.-F. Chen, S.-M. Wang, W.-C. Liao, L.-C. Kuo, K.-C. Chen, Y.-C. Lin, C.-Y. Yang, C.-H. Chiu, S.-C. Chang, F. Lai *et al.*, "Automatic icd-10 coding and training system: deep neural network based on supervised learning," *JMIR Medical Informatics*, vol. 9, no. 8, p. e23230, 2021.

[15] S. Baker, A. Korhonen, and S. Pyysalo, "Cancer hallmark text classification using convolutional neural networks," in *Proceedings of the 5th Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*, 2016, pp. 1–9.

[16] S. E. de Oncología Médica (SEOM), "Las cifras del cáncer en españa 2022," 2022.

[17] Z. Zhang, J. Liu, and N. Razavian, "Bert-xml: Large scale automated icd coding using bert pretraining," *arXiv preprint arXiv:2006.03685*, 2020.

[18] E. Al-Bashabsheh, A. Alaiad, M. Al-Ayyoub, O. Beni-Yonis, R. A. Zitar, and L. Abualigah, "Improving clinical documentation: automatic inference of icd-10 codes from patient notes using bert model," *The Journal of Supercomputing*, pp. 1–25, 2023.

[19] G. Bouzille and N. Grabar, "Supervised learning for the icd-10 coding of french clinical narratives," *Digital Personalized Health and Medicine: Proceedings of MIE 2020*, vol. 270, p. 427, 2020.

[20] S. Silvestri, F. Gargiulo, M. Ciampi, and G. De Pietro, "Exploit multilingual language model at scale for icd-10 clinical text classification," in *2020 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2020, pp. 1–7.

[21] A. D. Reys, D. Silva, D. Severo, S. Pedro, M. M. de Sousa e Sá, and G. A. Salgado, "Predicting multiple icd-10 codes from brazilian-portuguese clinical notes," in *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*. Springer, 2020, pp. 566–580.

[22] M. Chandrashekar, I. Lyngaas, H. A. Hanson, S. Gao, X.-C. Wu, and J. Gounley, "Pathbigbird: An ai-driven transformer approach to classification of cancer pathology reports," *JCO Clinical Cancer Informatics*, vol. 8, p. e2300148, 2024.

[23] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, and M. Krallinger, "Overview of automatic clinical coding: Annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020." *CLEF (Working Notes)*, vol. 2020, 2020.

[24] S. Amin, G. Neumann, K. Dunfield, A. Vechkaeva, K. A. Chapman, and M. K. Wixted, "Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert." in *CLEF (Working Notes)*, 2019, pp. 1–15.

[25] 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." *IberLEF@ SEPLN*, pp. 303–323, 2020.

[26] H. Face, "Hugging face: Natural language processing tools," https://huggingface.co, 2024, accessed: 2024-06-19.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] M. T. I. Miró, "El derecho a la desconexión en la ley orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de los derechos digitales," *Revista de Trabajo y Seguridad Social. CEF*, vol. 432, pp. 61–87, 2019. [Online]. Available: https://doi.org/10.51302/rtss.2019.1350

[29] W. H. Organization, *International Statistical Classification of Diseases and related health problems*. World Health Organization, 2004, vol. 3.

[30] M. Khadhraoui, H. Bellaaj, M. B. Ammar, H. Hamam, and M. Jmaiel, "Survey of bert-base models for scientific text classification: Covid-19 case study," *Applied Sciences*, vol. 12, no. 6, p. 2891, 2022.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[32] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas, "Detection of Tumor Morphology Mentions in Clinical Reports in Spanish Using Transformers," in *Advances in Computational Intelligence*. Cham: Springer International Publishing, 2021, pp. 24–35.

[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Jul. 2020, pp. 8440–8451.

[34] G. López-García, J. M. Jerez, N. Ribelles, E. Alba, and F. J. Veredas, "Transformers for Clinical Coding in Spanish," *IEEE Access*, vol. 9, pp. 72 387–72 397, 2021.

[35] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Vi-

llegas, "MarIA: Spanish Language Models," *Procesamiento del Lenguaje Natural*, vol. 68, no. 0, pp. 39–60, 2022.

[36] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas, "Pretrained biomedical language models for clinical NLP in Spanish," in *Proceedings of the 21st Workshop on Biomedical Language Processing*. Assoc. for Computational Linguistics, 2022, pp. 193–199.

[37] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas, "Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario," 2021.

[38] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez, "Spanish pre-trained bert model and evaluation data," *arXiv preprint arXiv:2308.02976*, 2023.

[39] M. W. Browne, "Cross-validation methods," *Journal of Mathematical Psychology*, vol. 44, pp. 108–132, 3 2000.

[40] F. He, T. Liu, and D. Tao, "Control batch size and learning rate to generalize well: Theoretical and empirical evidence," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a70712a252123c40d2adba6a11d84-Paper.pdf

[41] H. D. Abubakar and M. Umar, "Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec," *SLU Journal of Science and Technology*, vol. 4, pp. 27–33, 8 2022.

[42] S. Qaiser and R. Ali, "Text mining: Use of tf-idf to examine the relevance of words to documents," *International Journal of Computer Applications*, vol. 181, pp. 25–29, 7 2018.

[43] K. P. Murphy, "Naive bayes classifiers."

[44] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," pp. 45–66, 2001.

[45] Z. Qi, "The text classification of theft crime based on tf-idf and xgboost model," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2020, pp. 1241–1246.

[46] "Kaplan–Meier curve," *British Journal of Surgery*, vol. 104, no. 4, pp. 442–442, 02 2017. [Online]. Available: https://doi.org/10.1002/bjs.10238

UNIVERSIDAD DE MÁLAGA | uma.es

E.T.S de Ingeniería Informática
Bulevar Louis Pasteur, 35
Campus de Teatinos
29071 Málaga

E.T.S. DE INGENIERÍA INFORMÁTICA