UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Desarrollo de un sistema automático de extracción eficiente y análisis de variables clínicas radiológicas y patológicas en informes de cáncer de próstata a través de técnicas de procesamiento de lenguaje natural basadas en expresiones regulares

Trabajo Fin de Grado

Grado en Ingeniería Biomédica

AUTOR/A: Castro Anguita, Lourdes

Tutor/a: Ramos Soler, David

Director/a Experimental: MARTINEZ GIRONES, PEDRO MIGUEL

CURSO ACADÉMICO: 2022/2023

RESUMEN

El proyecto ha sido llevado a cabo mediante la colaboración conjunta entre la Universitat Politècnica de València (UPV) y el Grupo de Investigación Biomédica en Imagen (GIBI230), perteneciente al Instituto de Investigación Sanitaria La Fe situado en el Hospital Universitario y Politécnico La Fe de Valencia, junto con el apoyo de los servicios de Anatomía Patológica y Radiología del propio hospital.

El cáncer de próstata es una enfermedad que se origina en la glándula prostática, la cual está ubicada debajo de la vejiga y forma parte del sistema reproductor masculino. Estas lesiones ocurren cuando las células de la próstata comienzan a crecer de manera descontrolada, formando tumores que pueden invadir tejidos cercanos o diseminarse a otras partes del cuerpo. Dicho cáncer se encuentra entre los de mayor prevalencia a nivel mundial, aunque con un diagnóstico y tratamientos correctos se consigue que éste no sea de los de mayor tasa de mortalidad. Se diagnostica a partir del análisis visual de biopsias por medio del patólogo con la correspondiente clasificación de la diferenciación del tejido según la escala Gleason.

Sin embargo, la obtención de biopsias es un proceso invasivo, con riesgos clínicos y que únicamente abarca unas regiones concretas del órgano. Por ello, la vigilancia y búsqueda del diagnóstico es apoyada por medio de pruebas de imagen no invasiva, como es el caso de la Resonancia Magnética Nuclear (RMN). En esta prueba se obtiene una visión global de las características del tumor, sus distintas localizaciones, formas, texturas y la posible afectación a los ganglios linfáticos u otros órganos. Una vez obtenido el diagnóstico, es muy importante determinar qué tipo de tratamiento es el más adecuado para cada paciente. Esto se realiza utilizando toda la información obtenida tanto a nivel de anatomía patológica como a nivel radiológico, la cual está documentada en informes clínicos de texto libre que pueden estar poco estructurados. Por esto, es un reto complejo utilizar esta información para entrenar modelos de Inteligencia Artificial que simulen a partir de estos datos los posibles efectos que tendrían los distintos tipos de tratamiento en cada paciente y así poder facilitar la toma de decisión del tratamiento.

Mediante este trabajo se pretende solucionar este reto de disponer de información estructurada y de la mayor calidad posible, el cual es de vital importancia en el entorno clínico de investigación. Por lo tanto, el objetivo principal es el de extraer y analizar variables clínicas radiológicas y patológicas en cáncer de próstata a través de un sistema automático de procesamiento del lenguaje natural basado en expresiones regulares, dando lugar a bases de datos estructuradas con toda la información de diagnóstico de cáncer de próstata con directa aplicación al flujo de trabajo del entorno clínico real. Como objetivo secundario exploratorio se ha pretendido entrenar un clasificador de *Matching Learning* tipo KNN (*K-Nearest Neighbors*) que logre establecer la relación entre los valores de Gleason del tumor habiendo observado un PIRADS determinado en la imagen radiológica.

Palabras clave: cáncer de próstata; tumores; biopsias; Resonancia Magnética Nuclear; diagnóstico; tratamiento; anatomía patológica; radiología; informes clínicos; Inteligencia Artificial, procesamiento del lenguaje natural; expresiones regulares.

RESUM

L'estudi ha estat realitzat mitjançant la col·laboració entre la Universitat Politècnica de València (UPV) i el Grup d'Investigació Biomèdica en Imatge (GIBI230), que forma part de l'Institut d'Investigació Sanitària La Fe, situat a l'Hospital Universitari i Politècnic La Fe de València, amb el suport dels serveis d'Anatomia Patològica i Radiologia de l'hospital.

El càncer de pròstata és una malaltia que s'inicia a la glàndula prostàtica, ubicada sota la bufeta i integrada en el sistema reproductor masculí. Aquestes anomalies apareixen quan les cèl·lules de la pròstata comencen a créixer de manera descontrolada, formant tumors que poden envair teixits adjacents o estendre's a altres parts del cos. Malgrat la seua alta prevalença global, un diagnòstic i tractament adequats poden reduir la taxa de mortalitat. El diagnòstic es basa en l'anàlisi visual de biòpsies realitzades per patòlegs, utilitzant l'escala Gleason per classificar els teixits.

No obstant això, les biòpsies són procediments invasius amb riscos clínics, que només cobreixen àrees específiques de l'òrgan. Per això, les proves d'imatge no invasives, com la Resonància Magnètica Nuclear (RMN), complementen la vigilància i el diagnòstic. Aquesta prova proporciona una visió global de les característiques tumorals, localitzacions, formes, textures i possibles afectacions als ganglis limfàtics i altres òrgans. Un cop es realitza el diagnòstic, és fonamental determinar el tractament més adequat per a cada pacient, utilitzant informació de l'anatomia patològica i radiologia, encara que sovint no està estructurada en els informes clínics. Per tant, l'ús de la intel·ligència artificial és un repte complex per entrenar models a partir d'aquestes dades i facilitar la presa de decisions sobre el tractament.

Aquest treball busca resoldre aquest repte, proporcionant informació estructurada i de qualitat, crucial en la recerca clínica. L'objectiu principal és extreure i analitzar variables clíniques radiològiques i patològiques en el càncer de pròstata mitjançant un sistema automàtic de processament del llenguatge natural basat en expressions regulars. Això crea bases de dades estructurades amb tota la informació de diagnòstic del càncer de pròstata, amb aplicació directa a la pràctica clínica. Com a objectiu secundari exploratori, s'ha intentat entrenar un classificador de Matching Learning tipus KNN (K-Nearest Neighbors) per establir relacions entre els valors de Gleason del tumor i el PIRADS observat en la imatge radiològica.

Paraules clau: càncer de pròstata; tumors; biòpsies; Resonància Magnètica Nuclear; diagnòstic; tractament; anatomia patològica; radiologia; informes clínics; intel·ligència artificial; processament del llenguatge natural; expressions regulars.

ABSTRACT

The project has been carried out through joint collaboration between the Polytechnic University of Valencia (UPV) and the Biomedical Imaging Research Group (GIBI230), belonging to the La Fe Health Research Institute located at the La Fe University and Polytechnic Hospital in Valencia, along with the support of the Pathology and Radiology services of the hospital itself.

Prostate cancer is a disease that originates in the prostate gland, which is located below the bladder and is part of the male reproductive system. These lesions occur when prostate cells begin to grow uncontrollably, forming tumors that can invade nearby tissues or spread to other parts of the body. This cancer is among the most prevalent worldwide, although with proper diagnosis and treatment it can have lower mortality rates. It is diagnosed by analyzing biopsies visually by the pathologist with the corresponding classification of tissue differentiation according to the Gleason scale.

However, obtaining biopsies is an invasive process with clinical risks and only covers specific regions of the organ. Therefore, surveillance and diagnosis are supported by non-invasive imaging tests, such as Magnetic Resonance Imaging (MRI). In this test, a global view of the characteristics of the tumor, its different locations, shapes, textures, and possible lymph node or other organ involvement is obtained.

Once the diagnosis is obtained, it is essential to determine which type of treatment is most appropriate for each patient. This is done using all the information obtained both at the pathological and radiological levels, which is documented in free-text clinical reports that may be poorly structured. Therefore, it is a complex challenge to use this information to train Artificial Intelligence models that simulate, based on this data, the possible effects that different types of treatment would have on each patient, thus facilitating treatment decision-making.

This work aims to solve this challenge of having structured and high-quality information, which is of vital importance in the clinical research environment. Therefore, according to the main objective, radiological and pathological clinical variables in prostate cancer have been extracted and analyzed through an automatic natural language processing system based on regular expressions, resulting in structured databases with all the diagnostic information of prostate cancer with direct application to the real clinical workflow environment. The secondary and exploratory objective has pretended training a KNN (K-Nearest Neighbors) Matching Learning classifier to achieve a relationship between tumoral Gleason values and PIRADS category shown in the radiologic image.

Key words: Prostate cancer; tumors; biopsies; Nuclear Magnetic Resonance (NMR); diagnosis; treatment; pathological anatomy; radiology; clinical reports; Artificial Intelligence (AI); natural language processing; regular expressions.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCOLA TÈCNICA SUPERIOR ENGINYERIA INDUSTRIAL VALÈNCIA

Grado en Ingeniería Biomédica

TRABAJO FINAL DE GRADO

Desarrollo de un sistema automático de extracción eficiente y análisis de variables clínicas radiológicas y patológicas en informes de cáncer de próstata a través de técnicas de procesamiento de lenguaje natural basadas en expresiones regulares

Autor: Lourdes Castro Anguita

Tutor: David Ramos Soler

Cotutor externo: Pedro Miguel Martínez Gironés (IISLAFE)

GIBI230 GRUPO DE INVESTIGACIÓN BIOMÉDICA EN IMAGEN

La Fe Hospital Universitari i Politècnic

Instituto de Investigación Sanitaria La Fe

CAPÍTULO 1. INTRODUCCIÓN

1.1 CANCER

Según la OMS: "El cáncer es un término genérico que engloba un amplio grupo de enfermedades que pueden afectar a cualquier parte del cuerpo. Se caracteriza por el crecimiento descontrolado y la proliferación de células anormales que tienen la capacidad de invadir tejidos y órganos cercanos, y de propagarse a otras partes del cuerpo a través del sistema linfático o el torrente sanguíneo, proceso conocido como metástasis."

Esta definición resalta las características únicas del cáncer que la diferencia de otras enfermedades: el crecimiento celular anormal e incontrolado, y la metástasis que es la propagación del cáncer a sitios distantes.

El cáncer es la principal causa de muerte en todo el mundo. En España según la Sociedad Española de oncología médica se estima que en 2023 los cánceres más frecuentemente diagnosticados serán, los de colon y recto (42.721 nuevos casos), mama (35.001), pulmón (31.282), próstata (29.002) y vejiga urinaria (21.694). Muy por detrás se encuentran los linfomas no hodgkinianos (9.943), el cáncer de páncreas (9.280), el cáncer de riñón (8.626), el melanoma maligno cutáneo (8.049), los cánceres de cavidad oral y faringe (7.882), y los cánceres de cuerpo uterino (7.171), estómago (6.932) e hígado (6.695). [1]

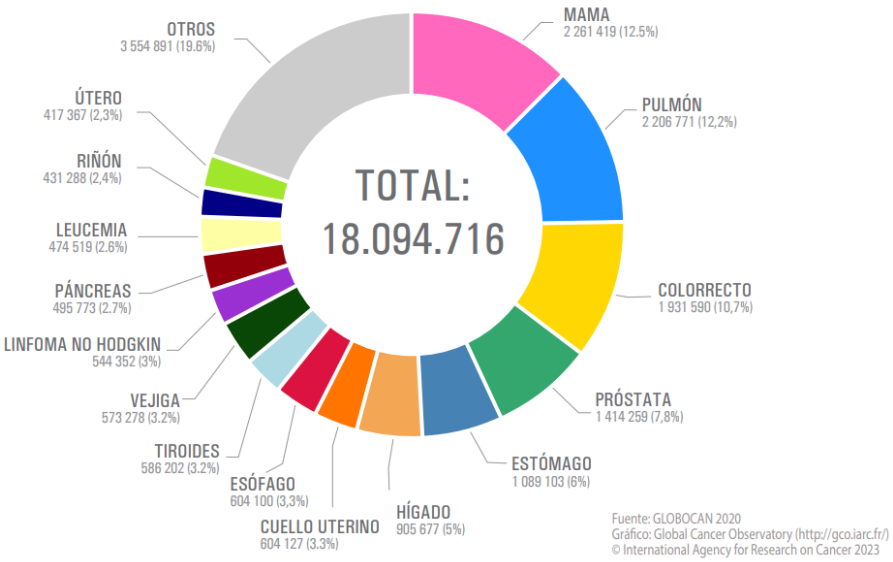

FIGURA 1: CÁNCERES MÁS FRECUENTES DIAGNOSTICADOS EN EL MUNDO

Por sexo, en los hombres, al igual que en 2022, serán mayoritarios los de próstata (29.002), colon y recto (26.357), pulmón (22.266) y vejiga urinaria (17.731). Y, en las mujeres, los de mama (35.001) y los de colon y recto (16.364). [1]

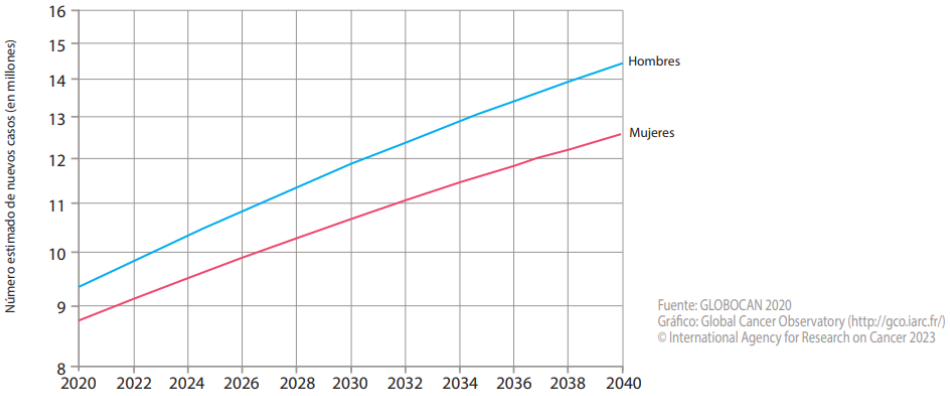

FIGURA 2: INCIDENCIA ESTIMADA DE CÁNCERES EN LA POBLACIÓN MUNDIAL POR SEXO ENTRE LOS AÑOS 2020 Y 2040

1.2 CÁNCER DE PROSTATA

Es el cáncer más común en hombres. Es una de las primeras causas de mortalidad por cáncer en hombres por detrás del cáncer de pulmón y colorrectal. Se diagnostican aproximadamente 1.276.106 casos nuevos en todo el mundo cada año.

Este tipo de cáncer afecta principalmente a hombres mayores, siendo más frecuente en aquellos mayores de 70 años.

Según los datos del Observatorio del cáncer de la AECC, en España se diagnosticaron 33.341 nuevos casos de cáncer de próstata y 6.112 personas fallecieron debido a esta enfermedad en el mismo año. [3]

El cáncer de próstata es una enfermedad que se origina en la glándula prostática, la cual está ubicada debajo de la vejiga y forma parte del sistema reproductor masculino. Estas lesiones ocurren cuando las células de la próstata comienzan a crecer de manera descontrolada, formando tumores que pueden invadir tejidos cercanos o diseminarse a otras partes del cuerpo.

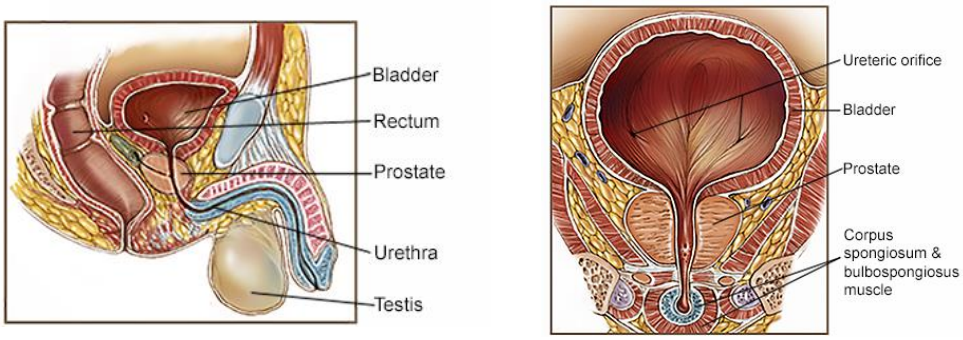

FIGURA 3: ANATOMÍA DE LA PRÓSTATA [2]

1.2.1 DIAGNÓSTICO

Por lo general, si su médico sospecha que puede tener cáncer de próstata lo primero que hará será preguntarle si presenta alguno de estos síntomas:

- Aumento de la frecuencia y la urgencia de orinar.
- Sangre en la orina (hematuria) o en el líquido seminal (hemospermia).
- Dolor y ardor al orinar (disuria).
- Molestias debido al aumento del tamaño de la próstata.
- Sensación de vaciamiento incompleto.

A continuación, realizará un tacto rectal (DRE) y le recomendará que se someta a un análisis de PSA (antígeno prostático específico).

La mayoría de los casos de cáncer de próstata se diagnostican mediante una prueba de sangre que detecte el PSA o mediante un tacto rectal.

1.2.1.1 TACTO RECTAL (DRE)

El tacto rectal es una exploración simple y sencilla en la que el médico introducirá un dedo, protegido por un guante y lubricado, a través del ano para palpar la próstata que se encuentra situada en la parte anterior del recto. Mediante esta prueba se puede detectar:

- Nódulos o irregularidades en la superficie.
- Aumento de la consistencia en una porción de la próstata o de manera difusa.
- Alteración en los bordes.

Aunque la palpación de la próstata sea normal no excluye la presencia de un posible foco de cáncer ya que este puede ser pequeño o no palpable, por este motivo se suele complementar con otras pruebas. [6]

1.2.1.2 PRUEBA DE PSA EN SANGRE (PSA)

El antígeno prostático específico (PSA) es una proteína que se produce en la glándula prostática y se encuentra principalmente en el semen, aunque también se puede encontrar en la sangre.

El PSA en sangre se mide en nanogramos por mililitro (ng/mL). A medida que los niveles de PSA en la sangre aumentan, existe un mayor riesgo de tener cáncer de próstata. Sin embargo, no existe un valor específico que pueda confirmar definitivamente la presencia o ausencia de cáncer de próstata.

- Un nivel de PSA menor a 4 ng/mL de sangre suele ser común en hombres sin cáncer. Sin embargo, un nivel por debajo de 4 no garantiza que un hombre no tenga cáncer.
- Los hombres con un nivel de PSA de 4 a 10 tienen una probabilidad de 1 entre 4 de padecer cáncer de próstata.
- Si el PSA es mayor de 10, la probabilidad de tener cáncer de próstata aumenta y esta es de más del 50%. [5]

Si los niveles de PSA son altos no significa necesariamente que haya cáncer de próstata. Indica que hay una mayor posibilidad de que la persona pueda tenerlo, pero se necesitarán más pruebas para confirmar o descartar el diagnóstico de cáncer de próstata ya que PSA alto también puede deberse a otras causas como:

- Liberación de PSA en la sangre tras un tacto rectal previo a la extracción de la muestra de sangre.
- Masaje prostático.
- Ecografía transrectal previa.
- Procesos infecciosos e inflamatorios de la próstata.
- Retención urinaria.
- Biopsias de próstata. Tras una biopsia los niveles de PSA pueden tardar en volver a sus valores basales hasta un mes.
- Hiperplasia benigna de próstata de gran volumen.
- Colocación de sonda vesical y procedimientos endoscópicos. [6]

1.2.1.3 BIOPSIA DE LA PRÓSTATA

Si los resultados de una prueba de PSA, el DRE u otras pruebas sugieren que podría tener cáncer de próstata se realizará una biopsia de la próstata.

La biopsia es un procedimiento en el cual se extraen pequeñas muestras de tejido de la próstata para examinarlas bajo el microscopio. Existen dos formas de realizar la biopsia, la biopsia tradicional en la que se toman muestras aleatorias de la próstata a través del recto, y la biopsia de mapeo por fusión (técnica más avanzada), también por vía perineal en la que se toman muestras ordenadas de toda la próstata y unas muestras extra de la zona sospechosa en la resonancia.

Durante la biopsia, el médico se ayuda de estudios de imagen como ecografía transrectal (TRUS) o MRI, o una combinación de las dos para poder observar la próstata.

El médico inserta una aguja delgada y hueca en la próstata a través de la pared del recto (una biopsia transrectal) o a través de la piel entre el escroto y el ano (una biopsia transperineal). Se toman varias muestras de tejido de diferentes áreas de la próstata. Por lo general, se toman alrededor de 12 muestras, aunque esto puede variar según el caso. Una vez que se han obtenido todas las muestras necesarias, se retira cuidadosamente la aguja. [5]

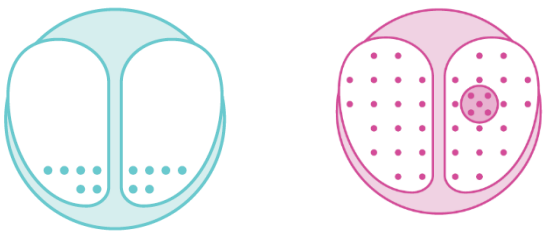

FIGURA 4: BIOPSIA TRADICIONAL Y BIOPSIA DE MAPEO POR FUSIÓN

Las muestras de tejido obtenidas se envían al laboratorio, donde se examinan al microscopio para determinar si hay células cancerosas presentes. Los resultados pueden ser:

- Positivo para cáncer: se observan células cancerosas en las muestras de la biopsia
- Negativo para cáncer: no se observan células cancerosas en las muestras de la biopsia
- Sospechoso: se observó algo anormal, pero puede o no ser cáncer.

RESULTADO NEGATIVO

La probabilidad de que padezca cáncer de próstata es muy bajo por lo que se puede descartar esta patología. Sin embargo, si el médico sospecha que padece dicha enfermedad, le repetirá ciertas pruebas.

RESULTADO SOSPECHOSO

Si obtenemos un resultado sospechoso se puede tratar de:

- Neoplasia prostática intraepitelial (PIN): Se caracteriza por un cambio en la apariencia de las células de la glándula prostática, pero estas no parecen estar invadiendo otras partes de la próstata.

  -PIN de bajo grado: las células de la próstata tienen un aspecto muy similar a las normales.

  -PIN de alto grado: las células suelen tener un aspecto más anormal. Un PIN de alto grado puede implicar una mayor probabilidad de padecer cáncer de próstata con el pasar del tiempo.
- Proliferación microacinar atípica: las células parecen ser cancerosas, pero hay muy pocas como para hacer un diagnóstico con certeza.
- Atrofia inflamatoria proliferativa (PIA): las células de la próstata son más pequeñas de lo normal, y hay signos de inflamación en el área. La PIA no es cáncer, pero a veces puede convertirse en una PIN de alto grado o en cáncer de próstata directamente.

RESULTADO POSITIVO

Si el resultado de la biopsia resulta positivo esto indica la presencia de células cancerosas, y por tanto se determinará el grado y el estadio del cáncer. Los patólogos utilizan la escala de Gleason para evaluar la agresividad del cáncer de próstata. El grado o escala de Gleason asigna una puntuación del 1 al 5 a las dos áreas más comunes de células cancerosas observadas en la muestra. Luego, se suma el valor de ambas áreas para obtener una puntuación final que puede variar de 2 a 10. Cuanto mayor sea la puntuación de Gleason, más agresivo se considera el cáncer.

Según la puntuación de Gleason, los cánceres de próstata pueden considerarse:

-Una puntuación de Gleason de 6 o menos, corresponde con cánceres bien diferenciados o de bajo grado.
-A una puntuación de Gleason 7 se les llama cánceres moderadamente diferenciados o de grado intermedio.
-A los cánceres pobremente diferenciados o de alto grado tienen una puntuación de Gleason de 8 a 10.

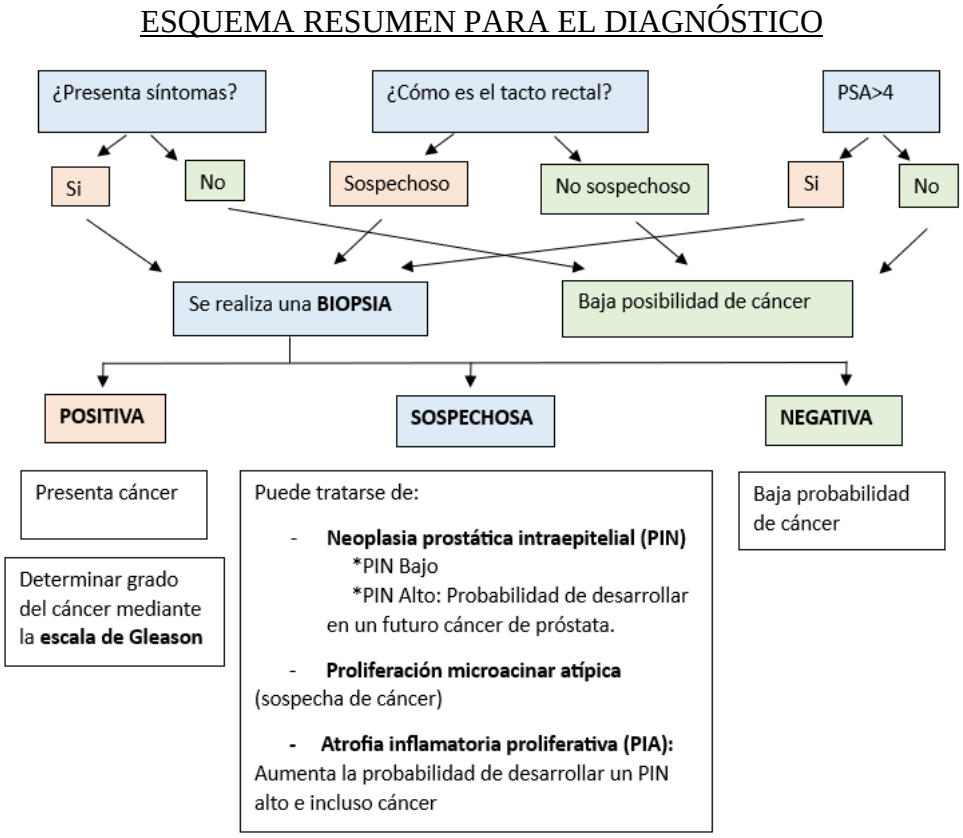

FIGURA 5: ESQUEMA DE DIAGNÓSTICO DEL CÁNCER DE PRÓSTATA EN LA FE

1.2.1.4 ESTUDIO POR IMÁGENES PARA EL CÁNCER DE PRÓSTATA

Los estudios por imágenes utilizan ondas sonoras, rayos X, campos magnéticos o sustancias radiactivas para obtener imágenes del interior del cuerpo. Los estudios por imágenes se suelen utilizar en estos casos para:

- Buscar cáncer en la próstata.
- Ayudar al médico a observar la próstata durante ciertos procedimientos.
- Buscar la propagación del cáncer a otras partes del cuerpo.

Los hombres con un resultado normal en tacto rectal, un nivel de PSA bajo y una puntuación de Gleason baja no necesitan otras pruebas, ya que las probabilidades de que el cáncer se haya propagado son muy bajas [5].

RESONANCIA MAGNÉTICA:

Es una técnica de diagnóstico por imagen en la que se producen imágenes anatómicas tridimensionales detalladas sin la necesidad de radiación. Se crea un potente campo magnético dentro del escáner y se emiten pulsos de radiofrecuencia hacia el área del cuerpo que se desea estudiar. Para construir la imagen, se detecta el cambio generado en la dirección del eje de rotación de los átomos de hidrógeno que se encuentran en el agua que compone los tejidos vivos. La resonancia magnética multiparamétrica (RMmp) es la modalidad que más se utiliza en el cáncer de próstata.

El procedimiento a seguir para realizar una resonancia magnética consiste en:

- Preparación: Como puede ser, evitar comer o beber durante un período de tiempo determinado. Además, de la posibilidad de administrar un agente de contraste intravenoso durante el examen para mejorar la visualización de ciertas estructuras.

- Posicionamiento: El paciente se acostará boca arriba en la camilla de resonancia magnética, con las piernas ligeramente separadas y los pies hacia afuera.

- Adquisición de imágenes: Se realizarán varias secuencias de imágenes utilizando diferentes parámetros y secuencias como:

Secuencia ponderada en T2: Esta secuencia proporciona imágenes estructurales de la próstata y permite visualizar la glándula y sus posibles alteraciones.

Imagen ponderada en difusión (DWI): Esta secuencia evalúa la difusión del agua en los tejidos de la próstata. Las áreas con mayor restricción en la difusión del agua pueden indicar la presencia de tejido tumoral.

Mapas de coeficiente de difusión (ADC): Se generan a partir de la secuencia DWI y proporcionan información cuantitativa sobre la difusión del agua en los tejidos. Las áreas con valores bajos de ADC pueden ser sospechosas de cáncer de próstata.

Secuencia ponderada en T1 con contraste: En algunos casos, se puede administrar un agente de contraste intravenoso para resaltar las áreas de captación anormal en la próstata. Esto puede ayudar a diferenciar entre tejido tumoral y tejido sano.

- **Evaluación e interpretación:** Las imágenes de resonancia magnética se interpretan por radiólogos especializados en cáncer de próstata. Se analizan diversos criterios, como la forma, el tamaño, el patrón de señal y la presencia de lesiones sospechosas dentro de la próstata. Posteriormente se categoriza las lesiones halladas mediante la probabilidad de malignidad mediante el PI-RADS (Prostate Imaging Reporting and Data System)

PI-RADS (PROSTATE IMAGING REPORTING AND DATA SYSTEM)

PI-RADS es un método de valoración de riesgo que predice la probabilidad de CPCS (entre 1, muy improbable, y 5, muy probable) mediante una lectura estandarizada y ponderada por zonas anatómicas de las secuencias que componen la RMmp. Aproximadamente, el 70-75 % de los CP se originan en la ZP y el 20- 30 %, en la ZT. Los cánceres que se originan en la ZC son infrecuentes y habitualmente secundarios a la extensión a partir de los tumores de la ZP.

- Zona periférica (ZP): contiene el 70-80% del tejido glandular, los conductos de estas zona se extienden desde la uretra distal (parte más alejada del verumontanum) en dirección postolateral.
- Zona central (ZC): contiene el 20% del tejido glandular; rodea los conductos eyaculadores con la posición posterosuperior de la glándula, por debajo de las vesículas seminales (VS)
- Zona transicional (ZT): contiene el 5% del tejido glandular y forma dos lóbulos periuretrales proximales al verumontanum con pequeños conductos, que en la edad adulta desarrollan la hiperplasia benigna de próstata (HBP)
- Estroma fibromuscular anterior (EFMA): no posee tejido glandular.

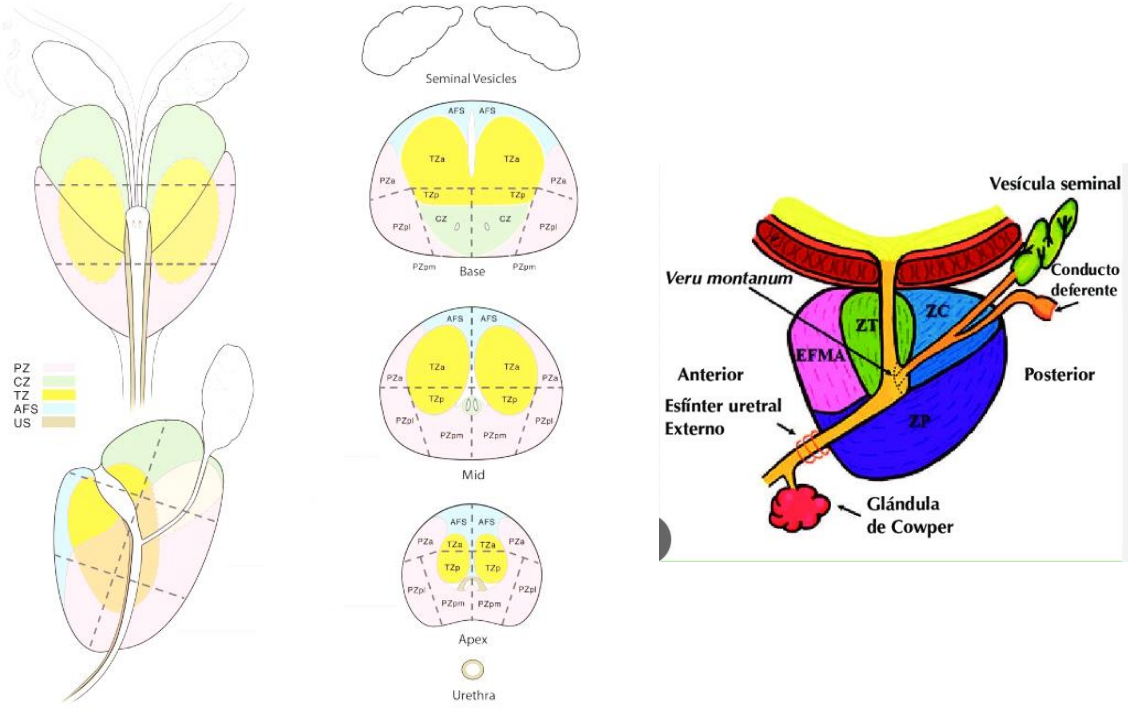

FIGURA 6: ANATOMÍA SECTORIAL DE LA PRÓSTATA SEGÚN EL PI-RADS V2.1

1.2.2 TRATAMIENTO

Existen distintos tipos de tratamientos para el cáncer de próstata en función al estadio que presenta, su localización, si existe o no metástasis…, a veces puede ser necesario también la combinación de varios tipos de tratamientos. Algunos tipos de tratamientos son [7]:

1.2.2.1 RADIOTERAPIA

Consiste en el uso de la radiación ionizante o fotones para dañar el ADN de las células cancerosas y así estas mueran sin crecer o replicarse. La radioterapia actualmente tiene la misma tasa de curación que la cirugía. [11]

Existen distintos tipos de radioterapia utilizados en cáncer de próstata. Entre ellos distinguimos entre:

- Radioterapia externa: es necesario una máquina externa para administrar la radiación al área afectada. La radiación se dirige a la próstata y el área circundante para destruir las células cancerosas. Este tipo de tratamiento se suele administra en sesiones diarias durante varias semanas.

- Braquiterapia: se colocan pequeñas fuentes radiactivas directamente dentro de la próstata. Estas fuentes emiten radiación de baja energía que actúa localmente en la glándula prostática. La braquiterapia puede ser de dos tipos: de baja tasa de dosis (LDR), donde las fuentes permanecen en su lugar de manera permanente, o de alta tasa de dosis (HDR) donde las fuentes se colocan temporalmente y se retiran después de un corto período de tiempo.

- Protonterapia: es una forma de radioterapia que utiliza protones en lugar de fotones o electrones. La deposición de los protones sigue una curva denominada pico de Bragg, lo que significa que pueden depositar la mayor parte de su energía en un punto específico y luego detenerse. Esto permite una dosis de radiación más concentrada en el tumor y menos radiación en los tejidos sanos circundantes. Posee la ventaja de reducir la exposición de los tejidos sanos a la radiación y por tanto se espera que haya menos efectos secundarios a largo plazo. Sin embargo, es más costosa que las anteriores y no todos los centros médicos disponen de ella.

1.2.2.2 CIRUGÍA PARA EL CÁNCER DE PRÓSTATA

Existen distintos tipos de procedimientos quirúrgicos como:

- Cirugía abierta: es la forma clásica en ella el cirujano hace una incisión en la parte inferior del abdomen con el fin de extirpar la próstata.
- Cirugía laparoscópica: Extirpación del tumor mediante pequeñas incisiones. Hay menos sangrado y una recuperación más rápida.
- Cirugía laparoscópica asistida por robot: Se hacen pequeñas incisiones, se insertan en ellas los brazos de un robot quirúrgico que el cirujano controla mediante una interfaz.

1.2.2.3 TERAPIA HORMONAL

La terapia hormonal (también llamada terapia de privación de andrógenos o ADT). El ADT está diseñado para detener la producción de testosterona ya que esta sirve como combustible principal para el crecimiento de células cancerígenas. Aunque la terapia hormonal es eficaz para controlar el crecimiento del cáncer de próstata también tiene efectos secundarios en casi todos los hombres. Estos efectos secundarios pueden ser: desde sofocos y pérdida de densidad ósea hasta cambios de humor, aumento de peso y disfunción eréctil.

Con la terapia hormonal la mayoría de las células cancerosas de próstata mueren o dejan de crecer cuando se les priva de testosterona. Sin embargo, en muchos hombres, algunas células adquieren la capacidad de crecer incluso en un ambiente con niveles bajos de testosterona y continúan creciendo, en estos casos los tratamientos hormonales pierden poco a poco su efecto. A esto también se le conoce como cáncer de próstata resistente a la castración (CRPC).

Se pueden emplear varios tipos de terapia hormonal para tratar el cáncer de próstata:

- Orquiectomía: conocida como castración (los otros son castración química). Aunque es un tipo de cirugía, su principal efecto es como una forma de terapia hormonal. En esta operación, el cirujano extirpa los testículos de forma permanente con el fin de frenar el crecimiento del cáncer o se pararlo por un tiempo.

- Agonistas de LHRH: son medicamentos que reducen la cantidad de testosterona producida por los testículos. Los agonistas de LHRH se inyectan o colocan como implantes pequeños debajo de la piel. Dependiendo del medicamento usado, pueden administrarse desde una vez al mes hasta una vez por año. Cuando se administran por primera vez los agonistas de LHRH, aumentan brevemente los niveles de testosterona antes de disminuir a niveles muy bajos. Este efecto se denomina **exacerbación.**

- Antagonistas de LHRH: El medicamento degarelix (Firmagon) es un antagonista de LHRH que actúa como los agonistas de LHRH, pero reduce los niveles de testosterona más rápidamente y no causa exacerbación del tumor como lo hacen los agonistas de LHRH. Se administra mensualmente en forma de inyección debajo de la piel [11].

1.2.2.4 QUIMIOTERAPIA

La quimioterapia es un tratamiento utilizado en el cáncer de próstata en etapas avanzadas o cuando otros tratamientos no han sido efectivos.

La quimioterapia en el cáncer de próstata funciona atacando las células cancerosas inhibiendo su capacidad para dividirse y crecer. Los medicamentos de quimioterapia en cáncer de próstata son los taxanos como el docetaxel o el cabazitaxel que se administran a través de una vena (intravenosa) y luego se distribuyen por todo el torrente sanguíneo. Llegan a las células cancerosas de la próstata y las células cancerosas en otros lugares donde puedan haberse diseminado.

Aunque la quimioterapia ayuda a frenar el avance del cáncer y reducir su tamaño también pueden causar afectar a células sanas.

1.2.2.5 INMUNOTERAPIA

El objetivo de la inmunoterapia en el cáncer de próstata es activar el sistema inmunológico para mejorar la respuesta inmune del organismo contra las células cancerosas de la próstata. Existen diferentes tipos de inmunoterapia como los inhibidores de puntos de control inmunológicos, terapia celular adoptiva o vacunas terapéuticas. Es importante tener en cuenta que no todos los pacientes responden de la misma manera a la inmunoterapia, y los resultados pueden variar según cada caso individual.

1.3 HISTORIA CLÍNICA

La historia clínica comprende el conjunto de los documentos que recopila información relevante sobre la salud de un paciente.

Estos documentos contienen información como: antecedentes médicos, antecedentes familiares, historial de enfermedades previas, alergias, medicamentos tomados, resultados de pruebas médicas, informes de procedimientos y cirugías, notas de consulta y seguimiento, entre otros.

La función principal de la historia clínica es facilitar el trabajo de los profesionales de la salud que tengan que tratar a un paciente, conociendo de primera mano y de forma inmediata toda la información relativa a su salud.

De esta manera el médico tiene la posibilidad de ofrecer una asistencia personalizada al paciente, aprender y mejorar los aciertos y errores en tratamientos pasados, investigar algunas ramas científicas a partir de la información contenida en el documento, mejorar la calidad de la salud de un paciente, etc.

Es importante destacar que cualquier médico o profesional de la salud que acceda a la información confidencial contenida en una historia clínica debe mantener la privacidad y la confidencialidad de esos datos, de acuerdo con la legalidad y el Código Deontológico, asegurándose de mantener en secreto cualquier información revelada.

El paciente tendrá derecho a que quede constancia escrita de cualquier proceso médico en su historia clínica, este además debe contener el profesional que lo realiza, una fecha y lo que se realiza en el proceso.

El paciente podrá acceder a sus datos siempre que quiera, y a recibir una copia de la misma si la solicita. Además, tendrá derecho a la confidencialidad y privacidad de sus datos, siendo además un delito grave el acceso a la historia clínica sin autorización.

1.3.1 INFORME ESTRUCTURADO

El informe estructurado es una forma específica de organizar y presentar la información en la historia clínica. El informe estructurado utiliza campos predefinidos o elementos definidos para registrar la información. Estos campos pueden ser seleccionados de una lista desplegable o introducidos manualmente por el médico o el personal de salud.

El objetivo del informe estructurado es estandarizar la documentación médica y simplificar la búsqueda, recuperación y análisis de la información clínica.

1.3.2 EXTRACCIÓN DE DATOS

La extracción de datos es un procedimiento clave utilizado por sistemas de información de salud para realizar distintos tipos de investigaciones y análisis clínicos

La extracción de la información se puede realizar de dos maneras:

- Manual: requiere de mucho tiempo y sería casi inviable para un gran tamaño de datos, pero no habría perdida de información.
- Automática: reduce notablemente el tiempo, es utilizada para una gran cantidad de datos, pero puede tener pérdidas de información.

En concreto en este trabajo, buscamos extraer datos de informes clínicos de un gran volumen de pacientes. Una forma de extraer de forma automática esta información es mediante el procesamiento del lenguaje natural (NLP).

1.3.3 INTELIGENCIA ARTIFICIAL

La inteligencia artificial (IA) se define como la capacidad de las máquinas para usar algoritmos y desarrollar sistemas que sean capaces de procesar grandes volúmenes de datos, aprender de ellos y tomar decisiones tal y como lo haría un ser humano de manera autónoma. Sin embargo, a diferencia de las personas, los dispositivos basados en IA no necesitan descansar y pueden analizar grandes volúmenes de información a la vez. Asimismo, la proporción de errores es significativamente menor en las máquinas que en los humanos que realizan las mismas tareas.

Existen varios tipos de inteligencia artificial pero principalmente distinguimos entre dos tipos principales de IA:

- IA basada en reglas: Se centra en la creación de reglas lógicas y algoritmos que guían el comportamiento del sistema. Estas reglas son diseñadas por expertos humanos y limitan el rango de respuestas del sistema.
- Aprendizaje automático (Machine Learning): Se basa en el uso de algoritmos y modelos para permitir que las máquinas aprendan a través de la experiencia y los datos. Los sistemas de aprendizaje automático pueden reconocer patrones, adaptarse y mejorar su rendimiento a medida que se les proporciona más información. La mayoría de los métodos de aprendizaje automático se pueden clasificar en tres categorías según el tipo técnica de aprendizaje utilizada: supervisado, no supervisado y aprendizaje por refuerzo.

En el aprendizaje supervisado inicialmente se entrena a la máquina proporcionando entradas o características (inputs) que están asociados a un resultado conocido o etiqueta (*outputs* o *label*) determinado por expertos humanos. En el caso de un lenguaje no supervisado a diferencia del anterior no se proporciona información previa, se introducen grandes cantidades de datos no etiquetados y el sistema se encarga de encontrar tendencias o patrones para separar la información en grupos de manera automática. Por último, en un aprendizaje por refuerzo se proporcionan tanto datos etiquetados como sin etiquetar, el sistema interactúa con el entorno y recibe recompensas negativas o positivas (*feedback*) de acuerdo con sus acciones, esto le permite desarrollar mejores caracterizaciones y clasificaciones.

Actualmente la inteligencia artificial tiene distintas aplicaciones como automatización de tareas, asistencias virtuales, reconocimientos de imágenes y voz, sistemas de recomendación, análisis del lenguaje natural entre otras.

1.3.3.1 INTELIGENCIA ARTIFICIAL EN MEDICINA

ANTECEDENTES

En los inicios de la medicina los profesionales sanitarios trabajaban en conjunto para resolver problemas que resultaban complejos para un solo individuo, actualmente la medicina es aún más compleja ya que se disponen de muchas más terapias, medicamentos, pruebas…

Es por esto por lo que ha surgido la necesidad de buscar nuevas herramientas con capacidad de integrar grandes cantidades de datos, reconocer patrones y crear modelos que permitan resolver las limitaciones humanas, acelerar la asistencia sanitaria, avanzar en una medicina más personalizada y disminuir los recursos.

La Inteligencia Artificial (IA) tiene sus raíces en la década de 1950 y se ha utilizado en diversos campos, incluido el sector salud. Uno de los primeros sistemas de IA en el ámbito médico fue Mycin, desarrollado en el 1970.

Mycin fue un sistema de IA diseñado para la detección de enfermedades infecciosas de la sangre. El objetivo de Mycin era ofrecer recomendaciones de tratamiento personalizadas para cada paciente.

El sistema Mycin se basaba en una base de conocimientos que incluía reglas específicas relacionadas con la detección y el tratamiento de enfermedades infecciosas. A través de preguntas y respuestas, Mycin recopilaba información sobre los síntomas y los resultados de pruebas médicas para realizar un diagnóstico y recomendar un plan de tratamiento. Sin embargo, también tenía sus limitaciones, ya que su capacidad de diagnóstico y tratamiento estaba restringida a enfermedades infecciosas de la sangre.

Desde entonces, la IA ha ido evolucionando en el sector salud y ha experimentado avances significativos en el análisis de grandes volúmenes de datos médicos, en el diagnóstico por imagen, la medicina de precisión y la atención personalizada. La IA continúa evolucionando y ofreciendo nuevas oportunidades para mejorar la eficiencia y la calidad de la atención médica. [11]

APLICACIONES ACTUALES

Algunas aplicaciones de IA aplicadas actualmente son:

En cardiología existen aplicaciones para realizar predicciones en áreas como la fibrilación y el riesgo de enfermedades cardiovasculares, como el síndrome coronario agudo.

Gastroenterología. Los gastroenterólogos utilizaron redes neuronales convolucionales entre otros modelos de aprendizaje profundo para procesar imágenes de endoscopia y ultrasonido y detectar estructuras anormales.

Radiología. Algunas aplicaciones actuales como la segmentación de imágenes permite analizar regiones de interés en las imágenes radiológicas, lo que facilita la identificación y el estudio de estructuras anatómicas específicas. Otra aplicación es la identificación de patrones y correlaciones que ayudan a predecir resultados clínicos y pronósticos. Esto permite una atención más personalizada y ayuda en la planificación del tratamiento.

Anatomía Patológica. Los sistemas de IA pueden automatizar tareas rutinarias en Anatomía Patológica, como la identificación y clasificación de células y tejidos. También pueden ayudar en el diagnóstico de enfermedades mediante el análisis automatizado de imágenes histopatológicas.

1.3.3.2 EXPRESIONES REGULARES

Una expresión regular es una cadena de caracteres que es utilizada para describir o encontrar patrones dentro de otros *strings*, en base al uso de delimitadores y ciertas reglas de sintaxis.

En medicina se utilizan las expresiones regulares para analizar y procesar grandes volúmenes de texto médico, como informes clínicos, registros electrónicos de pacientes y artículos de investigación. Por ejemplo, se pueden utilizar para encontrar patrones de síntomas, resultados de pruebas de laboratorio, valores de signos vitales, entre otros y luego utilizar estos datos para realizar estudios o desarrollar inteligencias artificiales.

Para trabajar con expresiones regulares en Python se utiliza la librería ‘re’ que contiene funciones y métodos para trabajar con expresiones regulares. Es necesario importar el módulo re en el script Python mediante la declaración "import re”.

La metodología de expresiones regulares se basa en patrones de caracteres que se utilizan para buscar y manipular texto de manera precisa. Estos patrones pueden incluir caracteres literales, clases de caracteres, cuantificadores, metacaracteres y más.

Algunas funciones y métodos útiles proporcionados por la librería re incluyen re.search(), re.findall(), re.sub() y re.split(). La función re.search() busca la primera coincidencia del patrón de expresión regular en el texto y devuelve el resultado. re.findall() busca todas las coincidencias del patrón en el texto y devuelve una lista con todos los resultados encontrados. re.sub() reemplaza todas las coincidencias del patrón con un texto específico. re.split() divide el texto en una lista de subcadenas basándose en el patrón proporcionado [11].

CAPÍTULO 2. OBJETIVOS

Hoy en día para el desarrollo de inteligencias artificiales y estudios en el ámbito médico, se requiere una sólida base de datos estructurada que contenga información relevante. No obstante, uno de los desafíos más significativos es que dicha información se encuentra almacenada en formato de texto libre, lo que dificulta considerablemente su accesibilidad y organización.

En el presente trabajo se tiene como objetivo principal abordar el reto de disponer de información estructurada y de alta calidad en el ámbito clínico de investigación en el cáncer de próstata. Para ello, se propone el desarrollo de un sistema automático de procesamiento del lenguaje natural basado en expresiones regulares, que permita extraer y analizar variables clínicas radiológicas y patológicas relacionadas con el cáncer de próstata.

Primero se implementará un sistema basado en expresiones regulares que permita extraer información sobre variables relevantes de los informes clínicos generados por los servicios de Anatomía Patológica y Radiología del Hospital Universitario y Politécnico La Fe de Valencia. Esto permitirá convertir información expuesta en formato libre en datos concretos y organizados.

Las variables que se pretenden detectar inicialmente son los valores de PI-RADS y Gleason, aunque también nos podría interesar otras variables como el volumen del tumor y su localización. El PI-RADS y el Gleason son dos sistemas de clasificación utilizados en la evaluación del cáncer de próstata. Cada uno proporciona información valiosa sobre la gravedad, extensión y características de la enfermedad.

Objetivo exploratorio

Dado que en el estado del arte se encuentran pocas soluciones que relacionen las variables radiológicas con las descritas por los expertos en anatomía patológica, es de gran relevancia intentar encontrar patrones que conecten ambos diagnósticos y que, por lo tanto, puedan evitar pruebas innecesarias. Por ello, en el presente trabajo también se ha tenido el objetivo exploratorio de buscar estas relaciones mediante un clasificador KNN.

CAPÍTULO 3. MATERIALES Y MÉTODOS

El objetivo principal de este trabajo es la creación de un sistema automático, basado en expresiones regulares, utilizando como entorno de programación Python, con la finalidad de extraer información relevante a partir de informes médicos pertenecientes a los departamentos de Anatomía patológica y radiología en formato de texto libre. Para lograr este propósito, se ha empleado una base de datos que consta de informes procedentes de ambos departamentos pertenecientes al Hospital Clínico La Fe. Cabe destacar que ambas bases de datos han sido previamente anonimizadas para salvaguardar la privacidad de los pacientes.

Mediante el sistema automático implementado, se ha conseguido extraer para cada paciente el valor del Gleason de sus informes de anatomía patológica y su pi-rads de los informes de radiología correspondientes. Tras extraer estas variables, mediante un clasificador de tipo KNN se ha intentado predecir el valor de Gleason a partir de su pi-rads con el fin de arrojar un poco de luz sobre las asociaciones entre ambos valores relevantes en el contexto del estudio del cáncer de próstata.

Este trabajo representa un paso significativo hacia la automatización del análisis de informes médicos, permitiendo una extracción eficiente y precisa de datos clínicos clave a partir de textos libres. Asimismo, las correlaciones encontradas entre variables anatómicas y patológicas aportan información valiosa para la comprensión y tratamiento del cáncer de próstata.

3.1 LISTADO DE MATERIALES

	Componentes	**Características**
Hardware	Ordenador portátil: HP Pavilion x360 2-in-1 14-ek0xx	SO: Windows 11 Home, versión 22H2
		Procesador: 12th Gen Intel(R) Core(TM) i7-1255U, 1.70 GHz
		RAM: 16.0 GB
		Disco duro: 512 GB
		GPU: NVIDIA® GeForce® GTX 1050
Software	Python 3	Lenguaje de programación. Librerías: Pandas, re, numpy, sklearn
		Pycharm
	Microsoft Office 365	Word, Excel, PowerPoint, Teams, OneDrive, Outlook
Sujetos	Base de datos CSV con campos de texto correspondientes a la parte de texto libre de los informes de radiología y de anatomía patológica	- BD radiología: 37.876 informes de 4.548 pacientes - BD anatomía patológica 6.369 informes de 2.386 pacientes

TABLA **1***: MATERIALES UTILIZADOS EN EL TRABAJO*

3.2 BASE DE DATOS

Partimos de dos bases de datos, ambas relacionadas con el cáncer de próstata y proporcionadas por el Hospital Universitario y Politécnico La Fe. La información que contienen estas bases de datos proviene de dos departamentos de este hospital, radiología y anatomía patológica.

En la base de datos del departamento de radiología, se recopilan un total de 37.876 informes correspondientes a 4.548 pacientes.

Por otro lado, la base de datos proveniente del departamento de anatomía patológica consta de 6.369 informes relacionados con 2.386 pacientes.

Seguidamente, para conseguir el propósito principal de este Trabajo Final de Grado (TFG), identificar expresiones regulares a través de un sistema automático de procesamiento de lenguaje natural, se llevó a cabo un proceso de dos fases utilizando el entorno Jupyter Notebook de Python. En la primera etapa, se definieron las palabras clave que se buscaban identificar en los informes. Posteriormente, se utilizó un patrón específico para detectar estas palabras clave en cada informe individual. Como consecuencia de esta detección, se extrajo información relevante, como los valores de Pi-Rads y Gleason.

BASE DE DATOS PERTENECIENTE A RADIOLOGÍA

Para diagnosticar el cáncer de próstata los radiólogos suelen utilizar técnicas de imagen médica como la resonancia magnética. En ella se centran en aspectos claves como:

- Lesiones sospechosas: Áreas anormales en comparación con el resto de tejido normal.
- Textura y morfología de la próstata.
- Intensidad de la señal: Las áreas con una señal más fuertes puede indicar presencia de tumor en la zona.
- Vascularización: Buscan patrones inusuales de vasos sanguíneos ya que esto podría indicar presencia de cáncer.
- Determinan si el tumor ha invadido otros tejidos. En el caso de invasión de otros tejidos indica una mayor agresividad. También analizan las distintas zonas de la próstata afectadas.
- Evaluación multifuncional: Utilizan diversas secuencias de imágenes para obtener una visión completa y detallada de la próstata y las posibles lesiones.

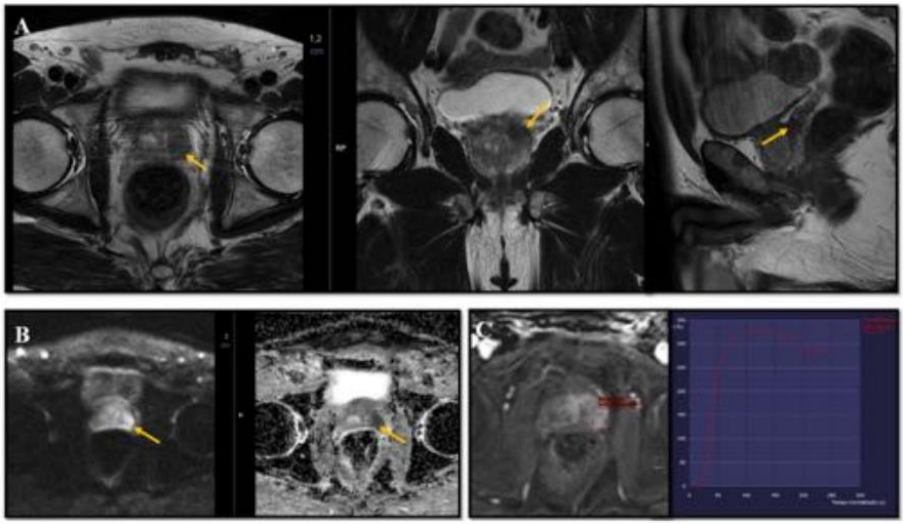

FIGURA 7: EJEMPLO IMAGEN DE RESONANCIA MAGNÉTICA. FUENTE: SOCIEDAD ESPAÑOLA DE RADIOLOGÍA MÉDICA

En la imagen A) se muestra un estudio morfológico T2 de alta resolución. En la imagen B) un estudio de difusión. Imagen C) Curvas de captación tipo III tras la administración de CIV. [15]

En relación con los informes radiológicos, nuestro enfoque fue obtener el valor de Pi-RADS.

Inicialmente partimos de texto libre proveniente de informes realizados por radiólogos del hospital la Fe con una estructura mas o menos parecida:

informe.txt

Archivo Editar Ver

```
Se obtienen planos de proyección axial, y coronal.
El paciente deniega el consentimiento para la inyección de gadolinio.

En el presente estudio se observa la glandula prostática que mide  36x47.5x38.8          mm, sin  condicionar      impronta sobre el trígono vesical .
Observamos unas imágenes de baja señal en las secuencias en T2 y Stir, con incremento de la restricción en la secuencia de difusión, en la adquisición con
coeficiente de 1000, y con señal negativa en el mapa ADC, localizadas en:
1:  Próstata periférica del tercio apical y medio posterior y lateral derecho (25,3 mm) PZpl, PZpm

 glandula periferica: 5
 DWI: 5
 Curva perfusión: -
 - PI-RADS 5

2:  Próstata transicional del tercio basal paramedial posterior derecho (6,6 mm) TZp

 glandula  central: 4
 DWI: 4
 Curva perfusión:   -
 - PI-RADS 4
.
Vesículas seminales libres.
La lesión rotulada con el número 1   muestra imagen de rotura capsular con mínima extensión extracapsular de 2,7mm (imagen 12  de la serie 5  ).
Aanillos neurovasculares  normales .
No se observan imágenes de crecimiento adenopático periprostático ni en las cadenas pélvicas.
No existe líquido libre en fondo de la pelvis menor.
No existe ureterohidronefrosis.
En el interior de la vejiga no se observan imágenes de defectos de repleción.
En el esqueleto óseo visualizado en el presente estudio no se visualizan imágenes que sugieran diseminación metastática ósea, identificando unas pequeñas
áreas de edema óseo en las palas ilíacas, en la vertiente posterior de las mismas, junto a la articulación sacroilíaca.

Por la características descritas esta lesión se encuentra en estadío    T3a  N0  Mx
```

FIGURA 8: EJEMPLO INFORME RADIOLÓGICO

A continuación, para poder obtener nuestro objetivo propuesto, el valor de Pi-Rads, empleamos el conjunto de palabras a identificar (palabras_buscadas): ['rads', 'Rads', 'RADS', 'rad', 'RAD'] y el siguiente patrón:

```
patron = r'\b(?:no\s*)?' + palabra_buscada + r'\s*(\d+)\b|\b(\d+)\s*(?:no\s*)?' + palabra_buscada + r'\b'
```

Si nos fijamos en el patrón utilizado no solo nos centramos en la búsqueda de la palabra si no también en el número más cercano a esta ya que corresponde con el valor Pi-RADS del informe.

En nuestro algoritmo para poder extraer esta información empleamos un bucle principal que recorre cada celda dentro de la columna de texto, la cual almacena los informes. En esta secuencia, establecemos dos listas vacías: palabras_encontradas y valores_encontrados. Estas listas contendrán las palabras clave identificadas en la celda y los valores numéricos correspondientes a dichas palabras.

En el proceso de búsqueda de las palabras clave, hemos empleado la función re.findall(), la cual rastrea todas las coincidencias del patrón dentro de la celda actual. Como resultado, obtenemos una lista de tuplas, donde cada tupla contiene dos posibles números: num1 y num2, encontrados en relación con la palabra clave.

Para cada tupla de coincidencias (num1, num2), evaluamos cuál de los dos números se encuentra más próximo a la palabra clave que estamos buscando. Además, tomamos en consideración que el número detectado se encuentre en el rango de 1 a 5, ya que esto concuerda con los valores del Pi-Rads, limitándonos a aquellos que son relevantes para nuestro análisis.

ID	fecha	Palabras	Valor Rads
167385	2015-10-27 00:00:00	[‘rads’,’rads’,’Rads’,’RADS’,’RADS’]	[‘5’,’4’,’5’,’4’,’5’,’4’]

FIGURA 9: INFORMACIÓN EXTRAÍDA DEL INFORME RADIOLÓGICO MOSTRADO DE EJEMPLO

Una vez localizada la palabra clave y el número más cercano, nos enfrentamos al desafío de que un solo informe puede presentar varias instancias de la palabra clave, incluso con diferentes valores de Pi-Rads. Esto se debe a la posibilidad de que distintas áreas de la próstata estén afectadas. Para abordar esta variabilidad, optamos por tomar el valor más elevado, dado que representa el nivel de agresividad más alto en relación con el cáncer.

ID	fecha	Palabras	Valor Rads
167385	2015-10-27 00:00:00	rads	5

FIGURA 10: RESULTADO OBTENIDO PARA EL EJEMPLO

En situaciones en las cuales no se detecta ninguna coincidencia en un informe, decidimos excluirlo, ya que carece de la información relevante que estamos buscando.

Al aplicar todas estas operaciones, logramos generar una nueva base de datos que comprende un total de 729 informes. Este proceso refinado nos permite concentrarnos en los informes más significativos y relevantes para nuestro análisis.

BASE DE DATOS PERTENECIENTE A ANATOMÍA PATOLÓGICA

El diagnóstico definitivo del cáncer de próstata se realiza a través de la anatomía patológica, que implica el análisis microscópico de muestras de tejido tomadas en una biopsia.

Los patólogos para analizar estas muestras de tejido primero preparan la muestra, posteriormente se realizan una serie de cortes histológicos, se tiñen estas muestras y finalmente se analizan en el microscopio.

En estos cortes teñidos el patólogo se va a centrar en analizar:

- Arquitectura celular: Observar cómo las células cancerosas se organizan en comparación con el tejido normal.
- Tamaño y forma de las células. Un núcleo agrandado puede indicar que esta célula sea cancerosa
- Tinciones especiales. En ocasiones se utilizan tintes especiales para resaltar características de las células y así poder diferenciar en función de estas características entre varios tipos de cáncer.
- Invasión en tejidos cercanos
- Si hay necrosis
- Vascularización anormal

La figura muestra distintos tipos de indicadores tumorales de cáncer de próstata.

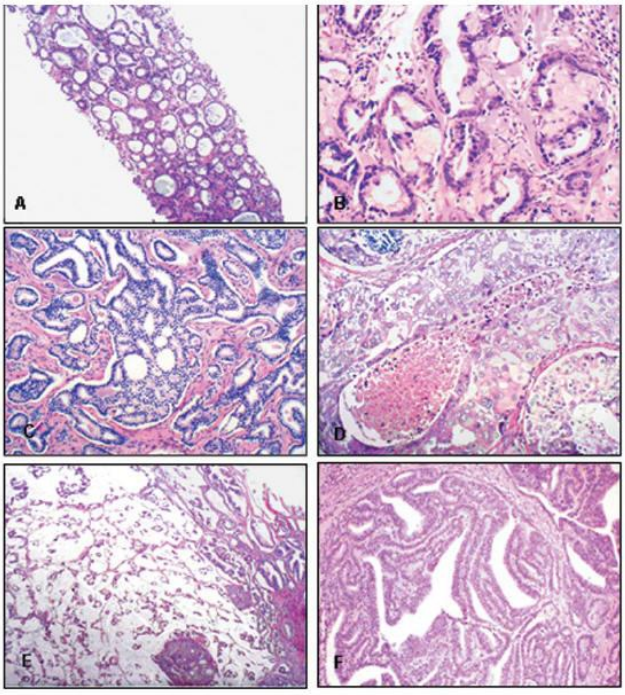

FIGURA 11: EJEMPLO MUESTRA HISTOLÓGICA BIOPSIA CÁNCER DE PRÓSTATA [16]

De manera similar a los Pi-Rads, hemos logrado identificar los valores de Gleason asociados a cada informe suministrado por el departamento de anatomía patológica.

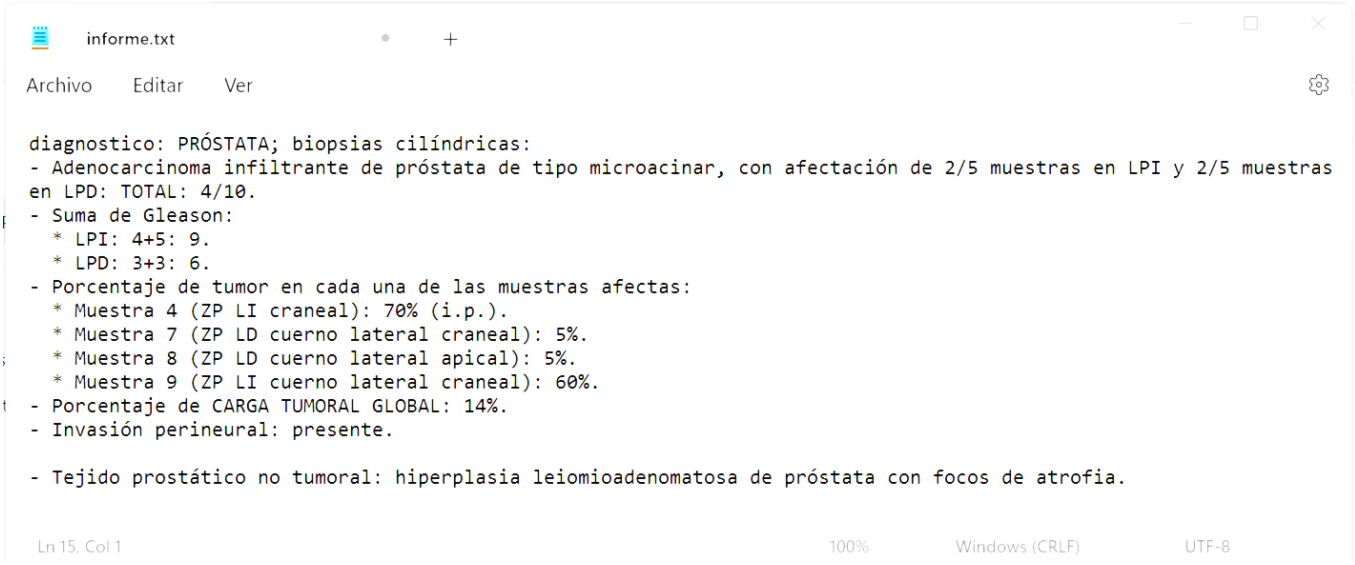

FIGURA 12: INFORME EJEMPLO ANATOMÍA PATOLÓGICA

En este contexto, el método más eficiente para realizar esta detección fue mediante la búsqueda de la expresión regular '+', ya que todos los informes seguían una estructura uniforme para presentar este valor: valor primario + valor secundario = puntuación Gleason.

El patrón utilizado para estos tipos de informe ha sido:

```
patron = r'(\d+)\s*\+\s*(\d+)'
```

En contraste con los informes radiológicos, en el contexto de los informes de anatomía patológica hemos logrado extraer tres valores de importancia:

-El número primario, correspondiente con el patrón más predominante dentro de la biopsia.

-El número secundario, correspondiente al segundo patrón más predominante.

-La suma Gleason, es la suma de los dos anteriores y nos indicara como de agresivo es el cáncer de próstata.

Mediante la implementación de estas operaciones, hemos creado una base de datos con un total de 1215 pacientes.

ID	fecha	Número Primario	Número Secundario	Suma Gleason
3239657	2015-09-22 00:00:00	[4,3]	[5,3]	[9,6]

FIGURA 13: INFORMACIÓN EXTRAÍDA DEL EJEMPLO

ID	fecha	Número Primario	Número Secundario	Suma Gleason
3239657	2015-09-22 00:00:00	4	5	9

FIGURA 14: RESULTADO OBTENIDO PARA EL EJEMPLO

BASE DE DATOS UNIFICADA

Con el fin de obtener una única base de datos con toda esta información se ha realizado otro algoritmo capaz de leer ambos archivos obtenidos anteriormente, compararlos por su person_id y generar así una nueva base de datos con cada person_id, su valor Gleason y su valor Pi-Rads.

Al unificar los datos nos encontrarnos con el problema de que había varios valores de Pi-Rads correspondientes para un mismo paciente en algunos casos por lo que se intentó resolver calculando cual de todos estos informes era el más cercano a su Gleason:

Fecha_ Pi_Rads _Seleccionada= min (Fecha_Pi_Rads – Fecha_Gleason)

Sin embargo, no conseguíamos que funcionase correctamente por lo que optamos por escoger el de fecha más antigua ya que solía coincidir con el más cercano a su Gleason en la mayoría de los casos. Primero, ordenamos de manera ascendente los informes dentro de cada paciente y finalmente seleccionábamos el primero de ellos que correspondía con el más antiguo.

En situaciones en las cuales no se detecta ninguna coincidencia en un informe, decidimos excluirlo, ya que carece de la información relevante que estamos buscando.

Tras realizar todas estas operaciones obtenemos una nueva base de datos con un total de 474 pacientes.

ID	fecha	Número Primario	Número Secundario	Suma Gleason	fecha	Palabra	Valor Rads
53950	2016-02-08 00:00:00	3	4	7	2016-10-27 00:00:00	rads	5

FIGURA 15: EJEMPLO PACIENTE BASE DE DATOS FINAL

APLICACIÓN PRÁCTICA

Posteriormente, para ver la utilidad que tiene extraer datos de este tipo mediante expresiones regulares hemos querido aplicar un clasificador de tipo KNN. El objetivo de este KNN es predecir qué tipo de Gleason tiene un paciente a partir de su Pi-Rads detectado en imágenes de resonancia magnética. Predecir el Gleason puede llegar a disminuir la necesidad de realizar una biopsia al paciente. Esto es beneficioso ya que se trata de una prueba invasiva.

El K-Nearest Neighbors (K-Vecinos Más Cercanos o KNN, por sus siglas en inglés) es un algoritmo de aprendizaje automático supervisado que clasifica o predice nuevos datos según la mayoría de las etiquetas de los K vecinos más cercanos en el espacio de características, en nuestro caso a partir del Pi-Rads nos detectará el Gleason.

Primero de todo hemos tenido que dividir nuestra base de datos en datos de entrenamiento y datos test con una proporción de 70% para el entrenamiento y 30% para los de test ya que vimos que era lo que más se utilizaba en estos casos, aunque también se suele utilizar la proporción 80-20%.

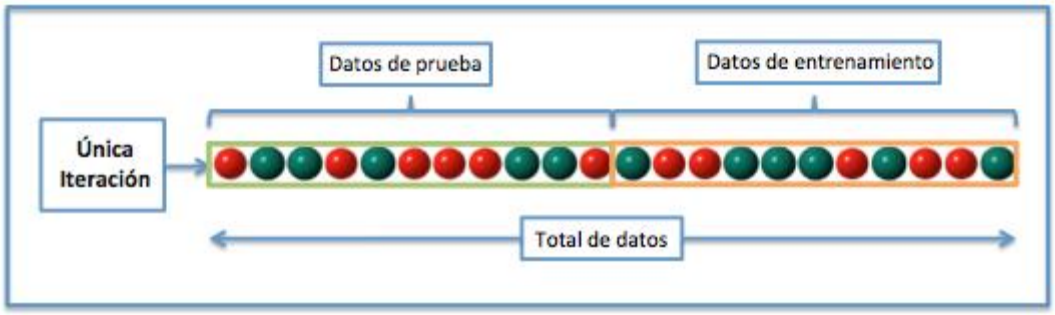

FIGURA 16: DESCRIPCIÓN GRÁFICA DE DIVISIÓN DE DATOS [16]

Para abordar el problema de la variabilidad en la partición de datos en conjuntos de entrenamiento y prueba, y proporcionar así una evaluación más sólida del desempeño del modelo utilizamos la validación cruzada.

La validación cruzada divide los datos en n partes iguales. Luego, el modelo se entrena y evalúa n veces, cada vez utilizando n-1 de los subconjuntos como datos de entrenamiento y el restante como conjunto de prueba en cada iteración.

Las partes en las que se suele dividir el conjunto (cv) suele ser 3, 5 o 10 dependiendo el volumen de los datos, como nuestra base de datos es de pequeño tamaño la opción que mejor se adapta es la de cv=3.

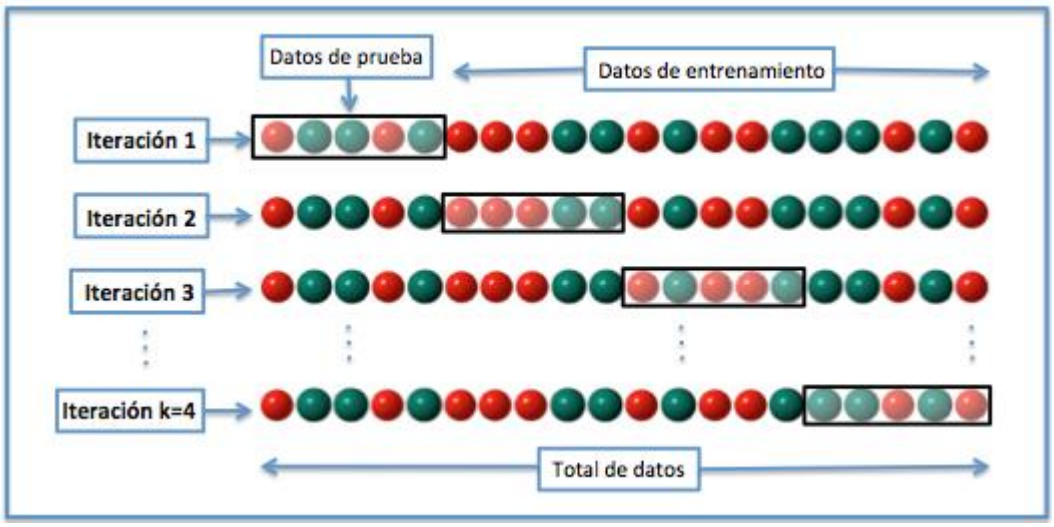

FIGURA 17: DESCRIPCIÓN GRÁFICA DE VALIDACIÓN CRUZADA CON CV=4 *[16]*

CAPÍTULO 4. RESULTADOS Y DISCUSIÓN

4.1 RESULTADOS OBTENIDOS MEDIANTE EL PROCESAMIENTO DEL LENGUAJE NATURAL

Tras aplicar los algoritmos anteriormente explicados hemos ido reduciendo el volumen de nuestros datos y quedándonos únicamente con los datos más relevantes para nuestro caso.

A continuación, se muestra un esquema resumen de cómo se ha ido modificando el volumen de estos datos y lo que se ha conseguido con ello:

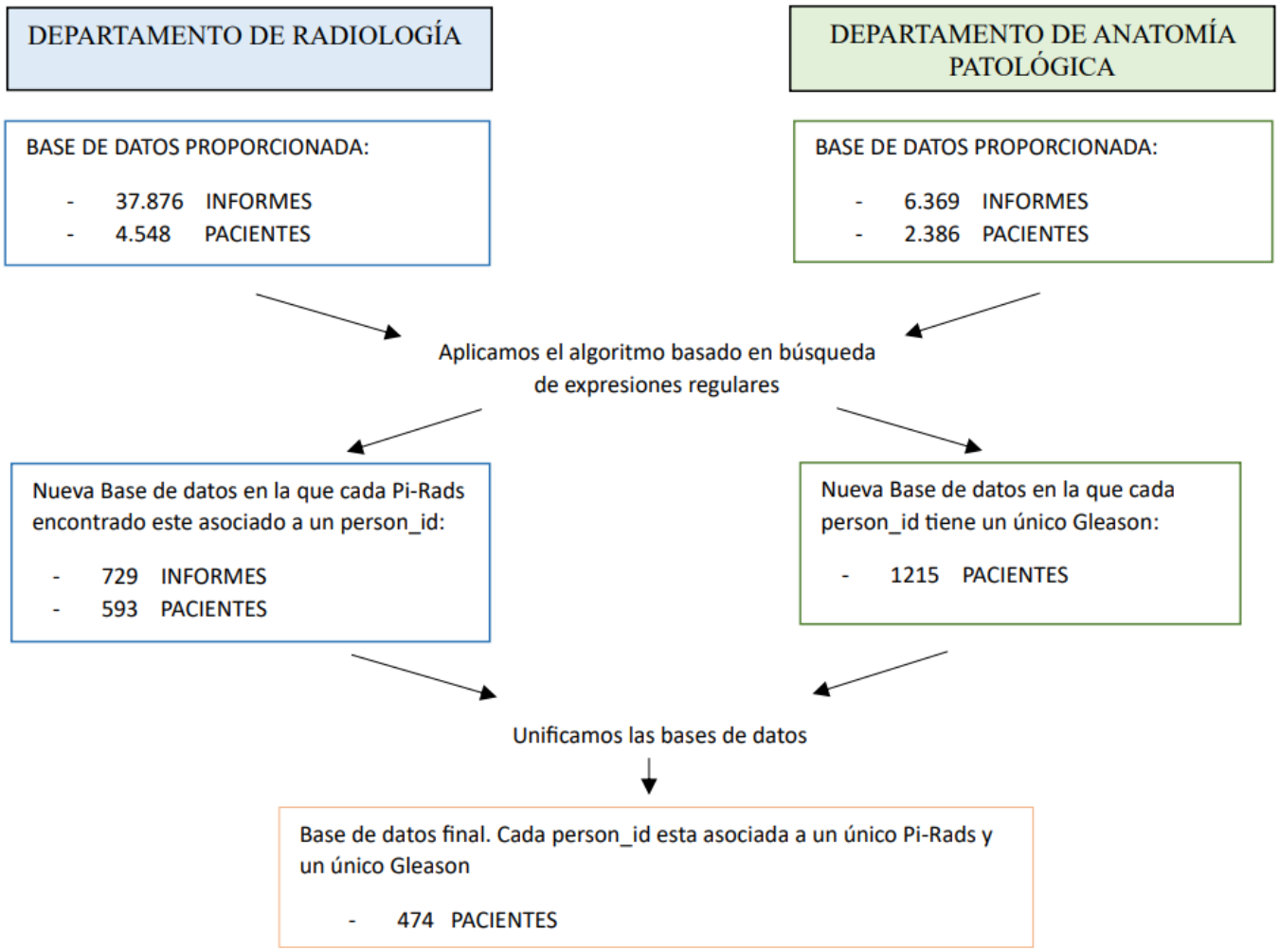

FIGURA 18: ESQUEMA RESUMEN DE ACTUACIÓN

Que se haya ido reduciendo el volumen de los datos tiene sentido ya que en un inicio no todos los informes proporcionados contenían la información de interés que buscamos. También se seleccionó únicamente el dato por paciente con fecha más relevante, por lo que ahí también eliminamos una gran cantidad de datos. Finalmente, al unificar las bases de datos se tuvo que descartar aquellos pacientes de los que no se tenía toda la información, por lo que la base de datos resultante quedo con un volumen mucha más pequeño que las que teníamos inicialmente.

BASE DE DATOS PERTENECIENTE A RADIOLOGÍA

Base de datos inicial

Column1	note_id	person_id	date	text	t_stage	n_stage	cTNM	pi_rads
0	1912498	1084282	01/01/2015		No match pa	No match pa	No match pa	No match pa
1	4499730	8907517	01/01/2015	JUICIO	No match pa	No match pa	No match pa	No match pa
2	4499728	8907517	01/01/2015	Sepsis	No match pa	No match pa	No match pa	No match pa
3	3860952	6734868	01/01/2015	JUICIO	No match pa	No match pa	No match pa	No match pa
4	4021877	7284281	01/01/2015		No match pa	No match pa	No match pa	No match pa
5	3968094	7097343	02/01/2015	JC : colico	No match pa	No match pa	No match pa	No match pa
6	2566874	2369223	02/01/2015	JUICIO	No match pa	No match pa	No match pa	No match pa
7	2997304	3807234	02/01/2015	DATOS	No match pa	No match pa	No match pa	No match pa
8	3655116	6039206	02/01/2015	no se observ	No match pa	No match pa	No match pa	No match pa
9	3231928	4612710	02/01/2015	Protocolo.	T2,	No match pa	No match pa	No match pa
10	2020833	1266007	02/01/2015	No	No match pa	No match pa	No match pa	No match pa

TABLA 2: BASE DE DATOS RADIOLOGÍA

Base de datos obtenida

ID	fecha	Palabra	Valor Rads
1739258	2015-09-24 00:00:00	rads	1
8743856	2016-02-04 00:00:00	rads	1
1380715	2016-03-31 00:00:00	rads	1
7298175	2020-05-14 00:00:00	rads	1
1307510	2020-10-06 00:00:00	rads	1
1266620	2020-12-18 00:00:00	rads	1
48500661	2015-03-26 00:00:00	rads	2
5759082	2015-09-03 00:00:00	rads	2
5732532	2015-12-17 00:00:00	rads	2
1013232	2016-01-13 00:00:00	rads	2

TABLA 3: BASE DE DATOS RADIOLOGÍA OBTENIDA

BASE DE DATOS PERTENECIENTE A ANATOMÍA PATOLÓGICA

Base de datos inicial

person_id	date	text
4529117	08/01/2015	diagnostico: 14B-25137 hasta 14B-25150:PRÓSTATA.-Pieza de prostatectomía radical laparoscópica (incluye adheridas, las dos vesículas s
6482392	09/01/2015	diagnostico: PRÓSTATA; fragmentos de RTU:- Adenocarcinoma infiltrante de próstata grado 8 de Gleason (3+5) y que ocupa un 15% de la r
2615795	15/01/2015	diagnostico: PRÓSTATA.-Pieza de prostatectomía radical (incluye adheridas, las dos vesículas seminales):-Adenocarcinoma microacinar pro
9989636	19/01/2015	diagnostico: PRÓSTATA; biopsias cilíndricas (10 biopsias):- Adenocarcinoma microacinar prostático infiltrante que afecta a las muestras 1,
8862369	20/01/2015	diagnostico: PRÓSTATA; biopsias cilíndricas:- Adenocarcinoma prostático de tipo microacinar con zonas cribiformes de apariencia ductal, c
5922958	20/01/2015	diagnostico: PRÓSTATA; biopsias cilíndricas:- Adenocarcinoma microacinar infiltrante de próstata, con afectación de ambos lóbulos de la gl
5635064	21/01/2015	diagnostico: PRÓSTATA; biopsias cilíndricas:- Adenocarcinoma infiltrante microacinar prostático con afectación de ambos lóbulos (aunque
1280129	21/01/2015	diagnostico: PRÓSTATA; biopsias cilíndricas:- Adenocarcinoma prostático infiltrante de tipo microacinar, con afectación exclusiva de LPD.- S
4887503	27/01/2015	diagnostico: PRÓSTATA; prostatectomía radical laparoscópica:- Adenocarcinoma microacinar prostático infiltrante que se extiende de man
7389868	28/01/2015	diagnostico: PRÓSTATA; biopsias cilíndricas:- Adenocarcinoma infiltrante de próstata de tipo microacinar, con afectación exclusiva de LPI.-

TABLA 4. BASE DE DATOS ANATOMÍA PATOLÓGICA

Base de datos obtenida

ID	Fecha	Número Primario	Número Secundario	Suma Gleason
24972	2022-01-18 00:00:00	3	4	7
53950	2016-02-08 00:00:00	3	4	7
89874	2019-01-12 00:00:00	3	4	7
113471	2016-06-24 00:00:00	3	4	7
136158	2019-03-04 00:00:00	3	4	7
197260	2016-08-30 00:00:00	3	3	6
202327	2016-05-16 00:00:00	3	4	7
206176	2021-02-17 00:00:00	3	4	7
260268	2021-03-01 00:00:00	3	4	7
276593	2016-11-10 00:00:00	3	5	8

TABLA 5: BASE DE DATOS ANATOMÍA PATOLÓGICA OBTENIDA

BASE DE DATOS UNIFICADA

ID	Fecha	Número Primario	Número Secundario	Suma Gleason	fecha2	Palabra	Valor Rads
53950	2016-02-08 00:00:00	3	4	7	2016-10-27 00:00:00	rads	5
136158	2019-03-04 00:00:00	3	4	7	2017-08-08 00:00:00	rads	4
202327	2016-05-16 00:00:00	3	4	7	2015-12-23 00:00:00	rads	5
325524	2017-02-02 00:00:00	4	5	9	2017-04-06 00:00:00	rads	5
368817	2020-03-23 00:00:00	3	3	6	2021-09-21 00:00:00	rads	4
542307	2016-03-03 00:00:00	3	3	6	2015-12-15 00:00:00	rads	3
668715	2018-01-24 00:00:00	4	5	9	2017-08-08 00:00:00	rads	5
687326	2017-04-06 00:00:00	4	3	7	2016-09-27 00:00:00	rads	4
728367	2015-09-08 00:00:00	3	4	7	2015-08-24 00:00:00	rads	4
753367	2021-10-22 00:00:00	3	3	6	2021-07-23 00:00:00	rads	4

TABLA 6: BASE DE DATOS UNIFICADA

4.2 RESULTADOS OBTENIDOS TRAS LA APLICACIÓN DEL CLASIFICADOR KNN

Para mostrar una de las posibles aplicaciones que puede tener la extracción de datos de interés mediante expresiones regulares, hemos aplicado un clasificador de tipo KNN para predecir el valor de Gleason a partir del Pi-Rads con el fin de poder evitar las biopsias, ya que se trata de una prueba invasiva. Los resultados obtenidos nos muestran una probabilidad de acierto de:

- % de acierto Número Primario: 77.43%
- % de acierto Número Secundario: 52.53%
- % de acierto Suma Gleason: 50.00%

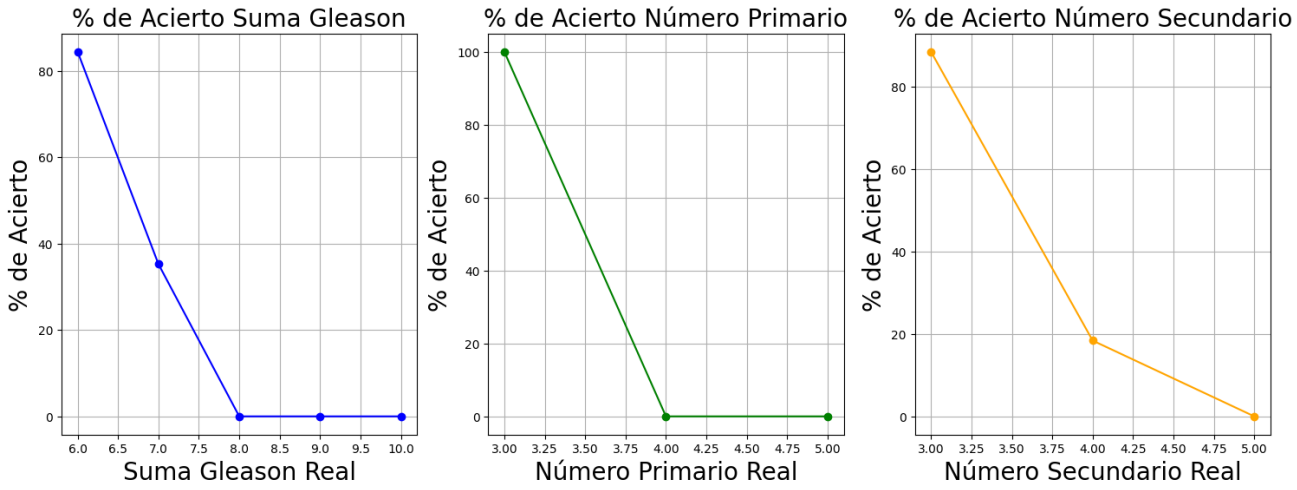

FIGURA 19: DIAGRAMA DE LÍNEAS QUE MUESTRAN EL % DE ACIERTO PARA: Nº PRIMARIO, Nº SECUNDARIO Y SUMA GLEASON

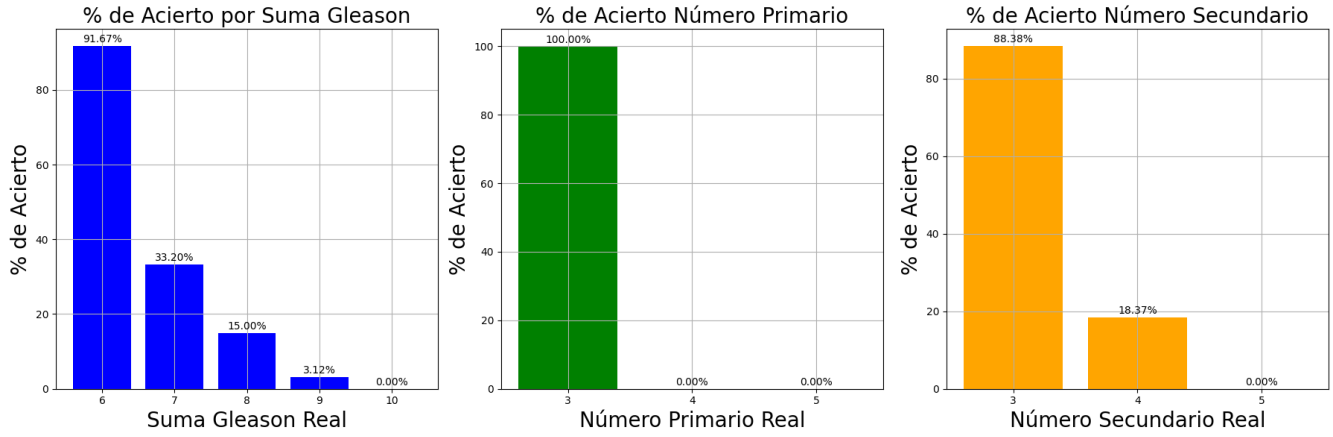

FIGURA 20: DIAGRAMA DE BARRAS QUE MUESTRAN EL % DE ACIERTO PARA: Nº PRIMARIO, Nº SECUNDARIO Y SUMA GLEASON

Si examinamos los resultados obtenidos en el análisis, podemos observar que el clasificador logra una tasa de acierto del 50,0%. Una observación detallada de los diagramas de barras revela patrones interesantes. En particular, para el número primario, siempre acierta cuando su valor es 3, mientras que para otros valores no logra la misma precisión. En el caso del número secundario, existe una alta probabilidad de acierto para el número 3, una probabilidad menor para el número 4 y, en general, una menor precisión para otros valores. Este comportamiento se debe a que la mayoría de los pacientes suelen presentar un Gleason de tipo 3+3 o 3+4, lo que concuerda con una mayor certeza en la predicción para los Gleason 6 y 7.

Es importante notar que la infrecuencia de los casos restantes disminuye la disponibilidad de datos para estos valores, resultando en una capacidad limitada del algoritmo para hacer predicciones precisas. Un aumento en el volumen de datos probablemente incrementaría la tasa de acierto.

No obstante, basándonos en los datos recopilados, no podemos afirmar con certeza la existencia de una relación significativa entre las variables radiológicas y patológicas. Para ello, se requeriría un estudio más riguroso y el uso de clasificadores más adecuados para este tipo de datos, como las redes neuronales, que podrían mejorar la capacidad de entrenamiento y predicción.

Actualmente no se ha encontrado ningún estudio que afirme que existe una relación directa entre el valor Pi-Rads y el valor de Gleason, pero si existen estudios que relacionan los resultados obtenidos por biopsias con los obtenidos en imagen médica. Uno de estos estudios ha sido publicado por las Actas Urológicas Españolas, una revista internacional dedicada a las enfermedades urológicas y al trasplante renal, en el volumen 40, Issue 5, June 2016, Pages 295-302

En ella detallan que, aunque las imágenes de resonancia magnética pueden llegar a detectar en algunos casos el cáncer de próstata, actualmente sigue siendo más preciso la biopsia y no siempre se obtienen los mismos resultados en ambas pruebas.

CAPÍTULO 5. CONCLUSIONES Y LÍNEAS FUTURAS

En conclusión, el procesamiento de lenguaje natural mediante expresiones regulares ha demostrado ser una etapa fundamental en la investigación.

A medida que hemos avanzado en el proceso de análisis, ha sido normal observar una reducción en el volumen de datos. Esto tiene sentido, ya que en un principio no todos los informes proporcionados contenían la información relevante que estábamos buscando. Además, hemos aplicado una selección cuidadosa, eligiendo solo el dato más relevante por paciente y fecha, lo que ha resultado en la eliminación de una gran cantidad de datos. Al unificar las bases de datos, también se ha requerido descartar pacientes con información incompleta, lo que ha llevado a una base de datos final con un volumen mucho más reducido en comparación con las bases iniciales.

La aplicación del clasificador KNN ha arrojado resultados interesantes en la predicción del valor de Gleason a partir de Pi-Rads. Si bien los porcentajes de acierto no son uniformes para todas las variables, se ha observado una alta precisión en ciertos casos, como el Gleason 6. Esta tendencia se explica por la predominancia de ciertos tipos de cáncer de próstata y su representación en los datos.

No obstante, estos resultados deben interpretarse con cautela. Aunque el clasificador muestra cierta capacidad predictiva, no se puede concluir con certeza la existencia de una relación causal entre las variables radiológicas y patológicas. La naturaleza limitada de los datos y la complejidad de la relación entre estas variables requieren un enfoque más profundo y sofisticado, posiblemente utilizando métodos más avanzados como las redes neuronales.

LÍNEAS FUTURAS

En cuanto a las perspectivas de investigación futura, se plantea la optimización de los sistemas de extracción de un conjunto más amplio de variables clínicas a partir de informes médicos, tales como el volumen tumoral, la localización precisa de la lesión, o la presencia de metástasis, utilizando expresiones regulares. Hasta ahora, este estudio se ha enfocado en la extracción de dos variables de interés fundamentales: Pi-Rads y Gleason.

A pesar de nuestros esfuerzos por incluir más variables, hemos enfrentado limitaciones en la capacidad de detección en la mayoría de los casos, lo que ha resultado en una reducción significativa en la cantidad de datos disponibles para el análisis.

Como paso adelante, se considera la implementación de clasificadores más avanzados, como las redes neuronales, con el propósito de mejorar el modelo de predicción y aumentar la precisión en las predicciones de diagnóstico.

Además, se busca extender la aplicación de estas técnicas a otros tipos de cáncer, como el cáncer de mama, para evaluar su efectividad en diferentes contextos clínicos y contribuir a un diagnóstico más temprano y preciso.

Estas líneas de investigación representan un esfuerzo continuo por perfeccionar y expandir las capacidades de análisis de datos en el ámbito médico, con el objetivo de brindar una atención de salud más efectiva y personalizada.

BIBLIOGRAFÍA

[1] (*Las cifras del cáncer en España*, 2023)

[(American Society of Clinical Oncology (ASCO), 2022)

[3](Asociación Española contra el cáncer, n.d.-a)

[4] (Asociación Española Contra el Cáncer, n.d.)

[5] (American Cancer Society, 2019)

[6] (Asociación Española contra el cáncer, n.d.-b)

[7] (American Cancer Society, 2023a)

[8] (*BOE-A-2002-22188-consolidado*, n.d.)

[9] (Obermeyer et al., 2017)

[10] (*medicine & technology*, n.d.)

[11] (Friedl, 2009)

[12] (American Cancer Society, 2023b)

[13] (Sociedad Americana Contra El Cáncer, 2018)

[14] "Joan.domenech91. (2011). Método de retención [Archivo de imagen].

[15] (Sandra Sánchez García (SERAM, 2018)

[16] (Servicio de Anatomía Patológica. Hospital Universitario Vall'dHebrón. Barcelona, 2007)

Anexo 1. PRESUPUESTO

	Horas	Precio/hora	TOTAL
Recursos humanos (costes directos)	450	20 €	**9.000 €**

Recursos materiales (costes directos)	Precio de coste	Uso (meses)	Amortización	TOTAL
Ordenador portátil: HP Pavilion x360 2-in-1 14-	899,00	9	5	134,85
Python	- €	9	-	-
Paquete Microsoft Office	69,00	9	1	23,00
TOTAL				**157,85 €**

Costes generales (costes indirectos)	15%	sobre CD	**1.483,86 €**
Beneficio industrial	6%	sobre CD+CI	**682,58 €**

Subtotal		**11.324,29 €**
IVA Aplicable	**21%**	**2.378,10 €**

Presupuesto TOTAL	**13.702,39 €**

Anexo 2. Ejemplo visual de la base de datos

BASE DE DATOS OBTENIDA RADIOLOGÍA

ID	fecha	Palabra	Valor Rads
1739258	2015-09-24 00:00:00	rads	1
8743856	2016-02-04 00:00:00	rads	1
1380715	2016-03-31 00:00:00	rads	1
7298175	2020-05-14 00:00:00	rads	1
1307510	2020-10-06 00:00:00	rads	1
1266620	2020-12-18 00:00:00	rads	1
48500661	2015-03-26 00:00:00	rads	2
5759082	2015-09-03 00:00:00	rads	2
5732532	2015-12-17 00:00:00	rads	2
1013232	2016-01-13 00:00:00	rads	2
7770527	2016-04-14 00:00:00	rads	2
6310569	2019-08-02 00:00:00	rads	2
7280309	2019-11-22 00:00:00	rads	2
8113992	2020-04-07 00:00:00	rads	2
2442022	2021-11-22 00:00:00	rads	2
2051174	2015-03-26 00:00:00	rads	3
542307	2015-12-15 00:00:00	rads	3
9004458	2016-03-10 00:00:00	rads	3
7530561	2016-04-29 00:00:00	rads	3
2228901	2016-05-04 00:00:00	rads	3
5027428	2016-05-13 00:00:00	rads	3
8123951	2016-06-22 00:00:00	rads	3
4264585	2016-07-07 00:00:00	rads	3
6928643	2016-07-07 00:00:00	rads	3
125366	2016-12-29 00:00:00	rads	3
863600	2017-01-26 00:00:00	rads	3
4501303	2017-05-25 00:00:00	rads	3
4171607	2017-06-05 00:00:00	rads	3
1643729	2017-08-24 00:00:00	rads	3

BASE DE DATOS ANATOMÍA PATOLÓGICA

ID	Fecha	Número Primario	Número Secundario	Suma Gleason
24972	2022-01-18 00:00:00	3	4	7
53950	2016-02-08 00:00:00	3	4	7
89874	2019-01-12 00:00:00	3	4	7
113471	2016-06-24 00:00:00	3	4	7
136158	2019-03-04 00:00:00	3	4	7
197260	2016-08-30 00:00:00	3	3	6
202327	2016-05-16 00:00:00	3	4	7
206176	2021-02-17 00:00:00	3	4	7
260268	2021-03-01 00:00:00	3	4	7
276593	2016-11-10 00:00:00	3	5	8
302695	2018-07-18 00:00:00	3	4	7
325524	2017-02-02 00:00:00	4	5	9
362807	2015-02-16 00:00:00	3	3	6
368817	2020-03-23 00:00:00	3	3	6
460632	2016-06-06 00:00:00	3	3	6
542307	2016-03-03 00:00:00	3	3	6
668715	2018-01-24 00:00:00	4	5	9
687326	2017-04-06 00:00:00	4	3	7
726712	2019-08-05 00:00:00	3	4	7
728367	2015-09-08 00:00:00	3	4	7
753046	2018-07-24 00:00:00	3	4	7
753367	2021-10-22 00:00:00	3	3	6
788752	2018-10-04 00:00:00	3	3	6
800036	2016-11-03 00:00:00	3	4	7
805898	2016-10-19 00:00:00	4	4	8
817111	2017-04-03 00:00:00	3	4	7
856944	2016-02-22 00:00:00	3	4	7
864255	2021-02-15 00:00:00	3	3	6
872740	2020-12-28 00:00:00	4	5	9

BASE DE DATOS CONJUNTA

ID	Fecha	Número Primario	Número Secundario	Suma Gleason	fecha2	Palabra	Valor Rads
53950	2016-02-08 00:00:00	3	4	7	2016-10-27 00:00:00	rads	5
136158	2019-03-04 00:00:00	3	4	7	2017-08-08 00:00:00	rads	4
202327	2016-05-16 00:00:00	3	4	7	2015-12-23 00:00:00	rads	5
325524	2017-02-02 00:00:00	4	5	9	2017-04-06 00:00:00	rads	5
368817	2020-03-23 00:00:00	3	3	6	2021-09-21 00:00:00	rads	4
542307	2016-03-03 00:00:00	3	3	6	2015-12-15 00:00:00	rads	3
668715	2018-01-24 00:00:00	4	5	9	2017-08-08 00:00:00	rads	5
687326	2017-04-06 00:00:00	4	3	7	2016-09-27 00:00:00	rads	4
728367	2015-09-08 00:00:00	3	4	7	2015-08-24 00:00:00	rads	4
753367	2021-10-22 00:00:00	3	3	6	2021-07-23 00:00:00	rads	4
856944	2016-02-22 00:00:00	3	4	7	2016-04-19 00:00:00	rads	4
864255	2021-02-15 00:00:00	3	3	6	2021-03-15 00:00:00	rads	3
1013232	2015-08-26 00:00:00	3	3	6	2016-01-13 00:00:00	rads	2
1014526	2015-07-14 00:00:00	4	5	9	2015-09-03 00:00:00	rads	5
1017466	2016-09-14 00:00:00	3	3	6	2017-02-23 00:00:00	rads	4
1023918	2020-10-27 00:00:00	3	4	7	2021-01-27 00:00:00	rads	4
1026014	2017-04-19 00:00:00	4	4	8	2017-09-06 00:00:00	rads	4
1030159	2016-02-16 00:00:00	4	3	7	2016-10-28 00:00:00	rads	4
1043429	2016-10-05 00:00:00	3	3	6	2017-02-02 00:00:00	rads	4
1043931	2022-01-31 00:00:00	3	3	6	2021-10-31 00:00:00	rads	4
1046832	2018-06-20 00:00:00	3	3	6	2021-09-27 00:00:00	rads	4
1051909	2018-08-08 00:00:00	3	4	7	2018-07-17 00:00:00	rads	5
1082214	2021-12-10 00:00:00	4	3	7	2021-10-10 00:00:00	rads	4