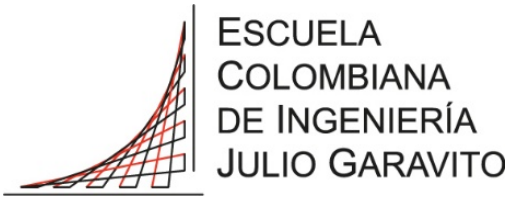
Procesamiento del Lenguaje Natural para el Apoyo en el Diagnóstico de Tuberculosis

Estudiante
Andrés Felipe Romero Gómez

Trabajo Dirigido

Tutores
Álvaro David Orjuela Cañón (D.Sc)
Andrés Leonardo Jutinico Alarcón



UNIVERSIDAD DEL ROSARIO
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
PROGRAMA DE INGENIERÍA BIOMÉDICA
BOGOTÁ D.C
2021

Agradecimientos

Agradezco a mi madre, Martha Viviana Gómez, por siempre apoyarme y ayudarme a dar lo mejor de mi. También al director de esta tesis, Álvaro David Orjuela Cañón, por la dedicación y el apoyo que me ha brindado en este trabajo, y sus enseñanzas. A mis compañeros de universidad con quienes crecí como persona, y compartí diversos aprendizajes y adversidades a lo largo de la carrera. Por último, a los miembros de Semillero Semill-IAS, por hacer posible un espacio donde se fomente el aprendizaje de temas como el de este trabajo, que no se suelen ver dentro de la carrera.

Resumen

La tuberculosis (TB) es una enfermedad infecciosa causada por la *Mycobacterium Tuberculosis*, que puede afectar a cualquier órgano del cuerpo, siendo la TB pulmonar la forma más frecuente de la enfermedad y la que más muertes causa. Según la Organización Mundial de la Salud (OMS), la TB se encuentra entre las 10 principales causas de muerte a nivel mundial, y en el caso de Colombia la TB es una enfermedad de interés en cuanto a la salud pública, por el alto número de casos que son reportados en el territorio, respecto a otras enfermedades transmisibles.

Uno de los principales problemas para manejo de la TB está en los métodos de diagnóstico, para los cuales se necesita de personal e infraestructura que no siempre están disponibles en lugares con sistemas de salud deficientes. Según el protocolo nacional para la detección de la TB, el diagnóstico de la TB pulmonar se debe hacer mediante una confirmación microbiológica, para lo cual se tienen tres tipos de pruebas, las baciloscopias, las pruebas moleculares y los cultivos. Todas las pruebas tienen un coste asociado y su disponibilidad es limitada, por lo que la generación de herramientas que den apoyo en el diagnóstico de la TB, pueden ayudar a tener un mejor control de la enfermedad.

La inteligencia artificial (IA) es un área de la informática que busca dotar a las máquinas de comportamientos inteligentes, con el fin de que realicen una tarea especifica. Una de las aplicaciones de la IA son los sistemas de apoyo a la toma de decisiones del inglés *Decision Support System* (DSS), estos sistemas aplicados en salud, buscan generar modelos que se basan en grandes volúmenes de datos y conocimientos clínicos previos, para ayudar al médico en la toma de mejores decisiones respecto a los pacientes.

Con el fin de generar herramientas que ayuden en el manejo de la TB, en el presente trabajo se utilizan técnicas de IA para el desarrollo un DSS que de apoyo en el diagnóstico de la TB, usando la información contenida en las historias clínicas electrónicas (HCE). Las HCE son fuentes de información ampliamente usadas por los médicos, en las cuales se registra el estado de salud de los pacientes, por lo que se espera que con la información contenida en ellas, se pueda generar una herramienta computacional que ayude a los profesionales de la salud en el manejo de la TB.

Para el desarrollo del trabajo se construyó una base de datos a partir de 151 HCE de pacientes sospechosos de TB pulmonar, en la base de datos se encuentran los reportes clínicos de los pacientes en fechas previas a la realización de las pruebas diagnósticas, de manera que en los reportes no se encuentra información sobre el diagnóstico final de TB. Para la creación de la herramienta diagnóstica, se tomaron los reportes clínicos y se les aplicó un preprocesamiento para limpiar el texto, luego, se extrajeron características usando 2 métodos TF-IDF (del inglés, *term-frequency - inverse document frequency*) y Word2Vec; posteriormente, se usaron modelos de aprendizaje automático para hacer la predicción de la TB. La exploración de modelos se realizó mediante validación cruzada, encontrando que los mejores resultados se obtienen haciendo una reducción de la dimensionalidad de las características obtenidas con TF-IDF, y usando del algoritmo de árboles aleatorios para la clasificación. Las métricas de desempeño obtenidas sobre los conjuntos de prueba con este modelo son: 0.721, 0.802, 0.462, y 0.723, en exactitud, sensibilidad, especificidad, y *F1-score* respectivamente.

Este trabajo se desarrolló dentro del proyecto“Generación de modelos alternativos basados en inteligencia computacional para tamización y diagnóstico de tuberculosis pulmonar” (minciencias, Universidad del Rosario, Universidad Antonio Nariño, Subred Integrada de Servicios

de Salud Centro-Oriente–Hospital Santa Clara), el cual es un proyecto conformado por un equipo conjunto de médicos e ingenieros, y tiene por objetivo generar herramientas computacionales, que puedan ser empleadas en lugares con infraestructura precaria para el diagnóstico de la TB pulmonar. Dentro del proyecto se están desarrollando modelos computacionales usando variables clínicas, epidemiológicas y sociodemográficas, se espera en un futuro integrar este trabajo con otras estrategias generadas dentro del proyecto, para la construcción de un sistema más robusto, que pueda apoyar al médico en el diagnóstico de la TB pulmonar.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

INTRODUCCIÓN

La tuberculosis (TB) es una enfermedad infecciosa causada por la micobacteria *Mycobacterium Tuberculosis*, esta enfermedad se encuentra entre las 10 principales causas de muerte a nivel mundial y hasta el 2020 era la primera causada por agente infeccioso único, superada ahora por el COVID-19 [1][2]. Se estima que un cuarto de la población mundial esta infectada con TB, pero la mayoría de estas personas no llegan a desarrollar la enfermedad gracias a la acción del sistema inmune, en los casos en los que si se desarrolla la enfermedad, esta puede llegar a ser mortal si no es tratada adecuadamente, para lo cual es necesario un diagnóstico oportuno [1][3][4].

El diagnóstico de la TB se basa en los síntomas del paciente, hallazgos imagenológicos y pruebas diagnosticas especificas para la enfermedad, estas pruebas diagnósticas son necesarias para dar comienzo al tratamiento, sin embargo, en el caso de no poder realizarse o dependiendo del estado del paciente, el médico puede iniciar el tratamiento a su criterio [5]. En el caso de la TB pulmonar, según el protocolo nacional para la detección de la TB, el diagnóstico debe hacerse mediante una confirmación microbiológica, de la presencia de la micobacteria en el esputo de los pacientes, para lo cual existen tres tipos de pruebas, las baciloscopias, las pruebas moleculares y los cultivos [5]. Dependiendo del tipo de prueba microbiológica se tendrá mayor o menor sensibilidad, siendo las de mayor sensibilidad las que requieren personal e infraestructura más especializada y costosa.

1.1 Motivación

En Colombia la TB es considerada un tema de salud pública, ya que es una de las enfermedades transmisibles que más casos registra en el Sistema de Vigilancia (Sivigila) colombiano, por ejemplo, para el 2019 se reportó una tasa de incidencia de 27,3 casos por cada 100 mil habitantes, de los cuales el 83,3 % fueron de TB pulmonar [6]. Esta tasa de incidencia es alta en comparación con países más desarrollados como los europeos, en los cuales la enfermedad se considera controlada, y las tasas de incidencia que pueden llegar a ser menores a 10 casos por cada 100 mil habitantes [2]. Uno de los problemas en Colombia, al igual que en otros países subdesarrollados, que impide un mejor manejo de la enfermedad, es que existe una falta de personal especializado y de la infraestructura necesaria para hacer un diagnóstico adecuado y oportuno de enfermedades como la TB [7][8].

De los tres tipos de pruebas que se pueden usar para el diagnóstico de la TB pulmonar, la baciloscopia es la prueba más sencilla y menos costosa, sin embargo, su sensibilidad es baja

y depende mucho de la calidad de la muestra. Las pruebas moleculares tienen una mayor sensibilidad, pero necesitan de equipos especializados para su realización y es más costosa respecto a la baciloscopia. Por último los cultivos, son los métodos de mayor peso, con una alta sensibilidad y especificidad, pero también requieren de una infraestructura especializada y costosa, así como personal capacitado para el manejo de las muestras y su resultado puede tardar semanas [5][9]. En todos los métodos de diagnóstico se requiere de personal y de equipos que no siempre están disponibles, además, los tiempos y los costos asociados a cada una de las pruebas, constituyen una brecha en la accesibilidad a las pruebas, por lo cual es necesaria la creación de nuevas tecnologías de bajo costo y rápidas que puedan ser usadas para apoyar a los profesionales de la salud en el diagnóstico de la enfermedad [7].

En los últimos años, las técnicas de inteligencia artificial (IA) y aprendizaje automático del inglés *Machine Learning* (ML), se han utilizado en medicina para desarrollar sistemas de apoyo a la toma de decisiones del inglés *Decision Support System* (DSS) [10][11], estos son sistemas informáticos en los que los profesionales en salud se pueden apoyar para tomar mejores decisiones, la ventaja de estos sistemas está en su capacidad de procesar grandes volúmenes de datos, y poseer conocimientos aprendidos o dados por un experto respecto a una tarea especifica [12]. Además, los DSS pueden ser utilizados por el personal sanitario de primera línea, lo que podría disminuir su carga de trabajo y presentan las ventajas de tener un bajo coste de implementación y ser sencillos de usar, lo que los hace muy útiles en situaciones donde no se tengan recursos suficientes para uso de las herramientas convencionales [13].

Para el caso de la TB se han empleado técnicas de IA para desarrollar modelos computacionales que permitan generar un diagnóstico o dar un mejor manejo de los pacientes sospechosos de TB. Por ejemplo, en [14] y [15] se evidencia como las redes neuronales artificiales del inglés *artificial neural networks* (ANN), pueden ser empleadas para el diagnóstico de la TB, usando variables clínicas de los pacientes. Además, en [15] y [16] se utilizan variantes de las ANN destinadas al agrupamiento, para determinar tres grupos de riesgo (alto, medio y bajo riesgo) de la población respecto a la TB, mostrando buenos resultados.

El procesamiento del lenguaje natural del inglés *Natural Lenguage Processing* (NLP), es una rama de la IA que permite un análisis de texto que no necesariamente está escrito en un lenguaje estructurado. En el ámbito clínico el NLP ha sido usado para construir sistemas de IA que ayuden en tareas de como la búsqueda de información relevante [17], determinar elegibilidad y hacer seguimiento de pacientes [18][19], o en el diagnóstico de enfermedades [20], generalmente usando la información contenida en los reportes clínicos de los pacientes. Para la construcción de estos sistemas de NLP se pueden usar modelos de IA basados en reglas dadas por un experto, modelos que infieren estas reglas de los datos, o una combinación de ambos; sin embargo, se ha visto que los modelos que aprenden las reglas a partir de los datos son los que mejores resultados han mostrado y los que mayores avances han tenido en el área recientemente [21].

Entendiendo la importancia de la generación de modelos alternativos que ayuden en el diagnóstico de la TB pulmonar, en países como Colombia donde no siempre están disponibles los métodos convencionales de diagnóstico, en este trabajo se plantea el desarrollo de modelos de IA que usen la información contenida en las historias clínicas electrónicas (HCE) de los pacientes, para predecir la TB pulmonar. Para ello se construyó una base de datos que contiene reportes clínicos de 151 pacientes sospechosos de TB, y se usaron técnicas de ML y NLP, con las cuales es posible extraer la información relevante del texto y usar esta información para realizar un diagnóstico del paciente.

1.2 Proyecto Relacionado

Este trabajo se enmarca dentro del proyecto "Generación de modelos alternativos basados en inteligencia computacional para tamización y diagnóstico de tuberculosis pulmonar" (minciencias, Universidad del Rosario, Universidad Antonio Nariño, Subred Integrada de Servicios de Salud Centro-Oriente–Hospital Santa Clara) en el cual hay una coordinación entre universidades y el hospital Santa Clara, para la conformación de un equipo de médicos e ingenieros, que tienen por objetivo generar herramientas computacionales, que puedan ser empleadas en lugares con infraestructura precaria para el diagnóstico de la TB pulmonar.

Las herramientas computacionales que se buscan desarrollar dentro del proyecto, se basan en modelos matemáticos que hacen uso de variables clínicas, epidemiológicas y sociodemográficas, para tamización y diagnóstico de la TB pulmonar. El trabajo que se desarrolló, pretende usar la información contenida en las HCE de los pacientes sospechosos de TB pulmonar, para generar una de estas herramientas computacionales basada únicamente en el análisis de texto, y se espera en un futuro integrar este trabajo con otras estrategias generadas dentro del proyecto, para la construcción de un sistema más robusto, que pueda apoyar al médico en la toma de decisión acerca del tratamiento que debe recibir un sujeto con sospecha de TB pulmonar.

1.3 Objetivos del Proyecto

1.3.1. Objetivo General

Implementar un sistema de NLP para el apoyo al diagnóstico de tuberculosis pulmonar a través de información extraída de historias clínicas.

1.3.2. Objetivos Específicos

- Acondicionar los datos en forma de texto de las historias clínicas disponibles para el proyecto empleando preprocesamiento.
- Determinar qué métodos de representación y extracción de características a partir de texto proporcionan mejores tasas de detección de la tuberculosis.
- Evaluar el empleo de tres técnicas de aprendizaje automático para la detección de tuberculosis a partir de texto, con métricas como sensibilidad, especificidad y área bajo la curva ROC.

1.4 Organización del documento

En este primer capítulo se mostró brevemente que la TB es una problemática vigente tanto en Colombia como a nivel mundial, y que uno de los principales problemas en el manejo de la enfermedad, es la brecha existente en la accesibilidad a las pruebas diagnósticas de la TB, a causa de la escasez de personal y de la infraestructura necesaria para realizar las pruebas, así como los costes asociados a estas. Dentro del proyecto *Generación de modelos alternativos basados en inteligencia computacional para tamización y diagnóstico de tuberculosis pulmonar*, se tiene por objetivo generar herramientas computacionales que puedan ser usadas en el

diagnóstico de TB pulmonar. En particular, en este trabajo se crea una de estas herramientas, la cual emplea técnicas de NLP y ML para generar un modelo que usa la información contenida en las HCE para dar apoyo en el diagnóstico de la TB.

En el segundo capítulo se presenta el marco teórico, el cual contiene la información más relevante acerca de la TB, y se profundiza más en la problemática local y mundial de la enfermedad. Además, se explica brevemente como son usadas las historias cínicas en Colombia y la normativa vigente de las HCE, ya que las estas serán usadas como fuente de información para la construcción de la base de datos. Por último, se dan las bases teóricas de la IA sobre las cuales se desarrollo en trabajo, enfocándose en el NLP para el tratamiento del texto, en el ML para la construcción de un modelo que aprenda a predecir TB.

En el tercer capítulo se presenta la metodología implementada, en este se describe cómo se realizó la extracción de texto de las historias clínicas, el preprocesamiento aplicado al texto, las técnicas de representación de texto usadas para extraer la información más relevante, y los modelos de ML que se pretenden usar para desarrollar el DSS para el apoyo al diagnóstico de la TB.

En el cuarto capítulo se muestran los resultados de todos los experimentos realizados, y la comparación del uso de las diferentes representaciones de texto así como el desempeño de los diferentes modelos. Posteriormente, en el quinto capítulo se discuten dichos resultados, explicando la posible causa de ellos y las ventajas y desventajas de emplear uno u otro modelo. Por último, se encuentran las conclusiones, recomendaciones y trabajos futuros.

Capítulo 2

MARCO TEÓRICO

Adicional a lo mencionado en el capítulo anterior, la TB es una enfermedad de relevancia mundial por ser una de las enfermedades que más muertes causa en todo el mundo, llegando en el 2019 a causar la muerte de 1,4 millones de personas en el mundo [1]. En este sentido la OMS y otras entidades se han propuesto diferentes metas para reducir las cifras de la TB, haciendo énfasis en sectores de la población que son más vulnerables ante la enfermedad y que no tienen un fácil acceso a los sistemas de salud [7][22]. Entre las problemáticas existentes que impiden un mejor control de la TB, están los costes y la disponibilidad de las diferentes pruebas diagnósticas, por lo que necesaria la creación de tecnologías que permitan hacer un diagnóstico oportuno, de forma que la enfermedad pueda ser tratada de la mejor manera posible.

Entre las tecnologías disponibles en la actualidad que permiten el desarrollo de estas herramientas, están los DSS que han sido utilizados en la salud desde los años 80's y que en la actualidad con los avances en IA han ido tomando mayor relevancia [10][11]. Los DSS hacen uso de la gran cantidad de datos que se acumulan de los diferentes ámbitos de la salud, para la generación de modelos computacionales que pueden ser usados en tareas como la optimización de procesos en el área, reducir la carga de trabajo, ayudar en el diagnóstico de enfermedades, y en general servir para prestar una mejorar atención a los pacientes. Un ejemplo de fuentes de información que pueden ser usadas para el desarrollo de DSS, son las HCE en donde se registran las condiciones de salud de los pacientes, y son ampliamente usadas en la actualidad por parte del personal de salud para hacerle un seguimiento constante a los pacientes. En este trabajo se pretende generar modelos de IA, que usen la información de las HCE para dar apoyo en el diagnóstico de la TB pulmonar, para ello se debe primero entender la enfermedad y su contexto mundial y local, así como las herramientas de IA que serán usadas para dar solución al problema.

2.1 Tuberculosis

En primer lugar, la TB es causada por una micobacteria que entra al organismo cuando una persona inhala los bacilos expulsados por otra persona enferma de TB, luego de que esta tose, estornuda o escupe [1]. Basta de unos pocos bacilos para la persona que los inhale pueda quedar infectada, y en el caso de que la persona se infecte de TB, pueden suceder dos cosas, que la persona enferme o que el sistema inmune combata la enfermedad e impida que se propague por el organismo; lo más común es el segundo caso donde el sistema inmune es capaz

de combatir la infección, en cuyo caso la persona no corre ningún riesgo, pero si el sistema inmune no puede combatir la TB, esta puede ser mortal si no es tratada oportunamente [1][3][4]. Se estima que una cuarta parte de la población mundial tiene TB latente, lo que significa que no presenta síntomas de la enfermedad ya que su sistema inmune fue capaz de contener la infección; por el contrario, las personas que si desarrollan la enfermedad se dice que tienen TB activa, en ellas lo normal es que la enfermedad afecte a los pulmones, pero la TB también pude afectar otros órganos del cuerpo, lo que puede complicar tanto el diagnóstico como el tratamiento de la misma [1][2][23].

Como la TB puede afectar cualquier órgano, dependiendo del órgano afectado se presentan diferentes síntomas, y los métodos de diagnóstico pueden variar [23]. En este trabajo se hace énfasis en la TB pulmonar, la cual se manifiesta mediante síntomas como tos intensa de más de tres semanas, dolor en el pecho y tos con sangre o esputo [24], y su diagnóstico se realiza con la confirmación microbiológica de la micobacteria en el esputo [5]. Cuando la TB se presenta fuera del pulmón se le denomina TB extrapulmonar y los síntomas varían respecto a la TB pulmonar, estos síntomas pueden llegar a confundirse con otras los de otras enfermedades, es por ello que su diagnóstico es más complicado.

La probabilidad de desarrollar la TB depende del sistema inmune de la persona, por lo cual entre los principales factores de riesgo están enfermedades como el VIH (Virus de la Inmunodeficiencia Humana), la diabetes, enfermedades de riñón, abuso de sustancias, cáncer y condiciones en las que haya un mal funcionamiento del sistema inmune [25]. Además, en los casos de TB latente, la micobacteria se encuentra solo controlada, pero puede permanecer latente dentro del organismo de una persona infectada durante toda la vida, debido a esto si por algún motivo el sistema inmune de una persona con TB latente se ve debilitado, esta tendrá una alta probabilidad de desarrollar la enfermedad, porque el sistema inmune ya no será capaz de seguir conteniéndola [1][25].

A su vez, existen factores sociales y demográficos que corresponden con altos niveles de incidencia de la enfermedad, esto se evidencia en que cerca del 90 % de las personas que enferman de TB cada año están concentradas en solo 30 países, estos países suelen tener una gran cantidad de población y sistemas de salud de baja calidad [2]. Por la amplia brecha existente entre países respecto al control de la TB, los esfuerzos de las organizaciones mundiales como la OMS, se centran más que todo en el control de la enfermedad en países con altas tasas de incidencia, y aunque el número de personas que se infectan cada año ha ido disminuyendo gracias a estas medidas, todavía se esta muy lejos de alcanzar las metas planteadas por la OMS para estos años [2]. En el reporte anual de la OMS del 2020 acerca de la TB [2], entre otras cosas, se muestra el progreso en los objetivos planteados respecto al 2020, en la figura 2.1 se muestran tres de los objetivos en cuanto a la tasa de incidencia, número de muertes por TB y la cantidad de personas que tienen que asumir los costos de los tratamientos; se puede observar que ninguno de los objetivos se ha cumplido, lo que muestra que aún queda un largo camino que recorrer en la eliminación de la TB a nivel mundial [2].

Por otro lado, la TB es una enfermedad prevenible y curable, cerca del 85 % de las personas que se enferman de TB pueden ser tratadas con drogas durante 6 meses, en países como Colombia el sistema de salud cubre este tratamiento [5], pero hay lugares del mundo en donde las personas tienen que afrontar los costes del tratamiento, que como que se ve en la figura 2.1, casi el 50 % de las personas se enfrentan a estos costos. Además, existen casos de personas con TB resistentes a los fármacos convencionales, para estos casos se debe recurrir a tratamientos más agresivos que a su vez son más costosos, como puede ser la quimioterapia antituberculosa

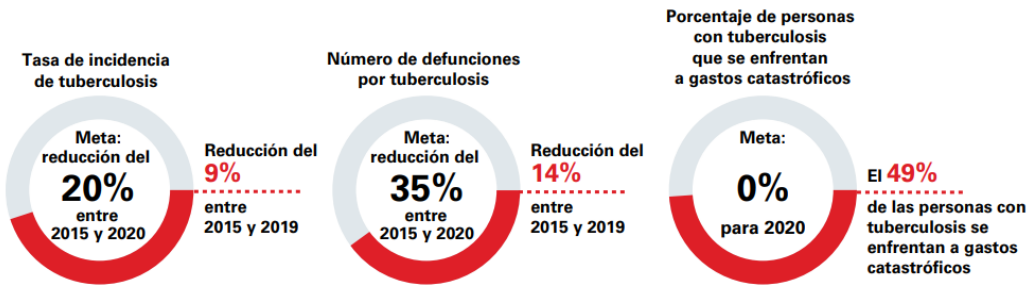


Figura 2.1: Metas de reducción en la incidencia de la TB, muertes por TB y costos catastróficos [2].

[23][1]. En cuanto a la TB latente, esta también puede ser tratada, lo que es muy importante en los casos de personas con un factor de riesgo alto, ya que como se mencionó antes la probabilidad de que desarrollen la enfermedad es alta, y es necesario que estas personas sean tratadas con anterioridad si se quiere disminuir la tasa de incidencia de la enfermedad [26].

2.1.1. Tuberculosis en Colombia

En Colombia en el 2019 se notificaron al sistema de vigilancia (Sivigila) 14684 casos de TB en todas sus formas, con una incidencia de 27,3 casos por cada 100 mil habitantes, siendo el departamento del Amazonas el más afectado, además, se evidencia que la TB afecta más a los hombres que a las mujeres, y se concentra en población laboralmente activa en un rango de edad entre los 29 y 59 años [5][6][23]. Esta incidencia es alta se mide respecto a países más desarrollados como europeos y de América del norte, pero no es un caso critico como los 30 países que concentran casi toda la población con TB, por lo que Colombia va por buen camino en la eliminación de la TB pero aún hacen falta esfuerzos para que la enfermedad deje de ser un problema para el sistema de salud colombiano [2].

El ministerio de salud y protección social colombiano (minsalud) recientemente en la resolución 227 de 2020 [5] presentó los nuevos lineamientos técnicos y operativos del Programa Nacional de Prevención y Control de la Tuberculosis (PNPCT). En esta resolución se presentan entre diferentes cuestiones, los protocolos para el diagnóstico y tratamiento de la TB, así como las definiciones establecidas para referirse a la enfermedad, estas directrices son importantes para establecer una metodología específica para el control de la enfermedad, y comunicarse mejor fuera y dentro del ámbito clínico. De acuerdo con dichos protocolos nacionales, para considerar a una persona sintomática respiratoria esta debe presentar tos y expectoración por más de 15 días, sin embargo, dado que la enfermedad puede presentarse de diferentes formas en el organismo, un profesional de medicina puede considerar una presunción de la enfermedad aún sin expectoración. En el caso de grupos de riesgo, como personas con inmunosupresión o vulnerables por su condición social podrá tomarse un tiempo menor a 15 días para la presunción de la enfermedad, ya que su condición hace necesario un pronto inicio del tratamiento.

A partir de las políticas para el diagnóstico de la TB en Colombia, se pueden dar dos casos de TB diagnosticada: 1) *Caso de tuberculosis bacteriológicamente confirmado*, donde el profesional en salud se basa en resultados de pruebas de laboratorios como la baciloscopia,

cultivo o prueba molecular. 2) *Caso de tuberculosis clínicamente confirmado*, es aquel caso diagnosticado por un profesional de la medicina que decide dar tratamiento a la tuberculosis con pruebas bacteriológicamente negativas, o sin tener su resultado aún [5]. El segundo caso mediante el cual se puede dar inicio al tratamiento sin tener los resultados microbiológicos, es necesario ya que un inicio tardío del tratamiento está asociado a mayor mortalidad y morbilidad, porque se incrementa la carga bacilar en el individuo [3][4], también es por esto que, la implementación de tecnologías de apoyo al diagnóstico de la enfermedad son importantes, para que los profesionales en salud puedan dar un diagnóstico más temprano y certero de la enfermedad.

2.1.2. Métodos de Diagnóstico

Los síntomas son el primer indicativo de una posible enfermedad causada por TB. Estos síntomas dependen del área del cuerpo afectada, y ya que la forma de la enfermedad más común es la TB pulmonar, los principales síntomas son: tos intensa de más de tres semanas, dolor en el pecho y tos con sangre o esputo. Otros posibles síntomas son fatiga, perdida de peso, falta de apetito, escalofríos, sudores nocturnos y fiebre. Ninguno de estos síntomas se presentan en personas con TB latente [24].

Entre las herramientas usadas para el diagnóstico de la TB están la prueba de derivado proteico purificado de tuberculina (PPD, por sus siglas en inglés) y la prueba de interferón gamma (IGRA). La primera es una prueba en la cual se inyecta PPD bajo la piel del antebrazo, y después de 48 a 72 horas, un profesional médico, evalúa la respuesta ante la PPD; un bulto elevado rojo y duro significa que hay una posible infección por TB, sin embargo, el profesional será quien verdaderamente determine el significado de los resultados. En cuanto a la prueba IGRA que se realiza en sangre, también se evalúa la reacción del sistema inmunitario de la persona ante las bacterias de TB. Ambas pruebas pueden realizarse en casi cualquier persona, pero ninguna de las dos determina si la persona tiene TB latente o activa, es por ello que para el diagnóstico de la TB activa se debe hacer una de las tres pruebas mencionadas con anterioridad de confirmación microbiológica [27].

Dentro de la resolución 227 de 2020 [5] donde se establecen las directrices para el diagnóstico de la TB pulmonar, enuncia que la sospecha de la TB pulmonar se basa en manifestaciones clínicas como las mencionadas previamente y en estudios radiológicos, y que para la confirmación de la enfermedad se debe dar una demostración de la presencia de la micobacteria con un estudio microbiológico. En el extremo caso en el que no se haya logrado hacer una confirmación bacteriológica, es aceptable tomar los aspectos clínicos e imagenológicos para el diagnóstico. Para el estudio microbiológico, se realiza un cultivo liquido y/o pruebas de biología molecular, para la TB pulmonar este examen se hace a partir del esputo (mucosidad producida por los pulmones).

- **Baciloscopia:**
  La baciloscopia seriada de esputo es una técnica con la cual se determina la cantidad de bacilos en el esputo. Para que una muestra sea considerada positiva, debe tener entre 5 mil y 10 mil bacilos por m^3. La sensibilidad de la detección depende en gran medida de la calidad de la muestra [28], Con una sensibilidad entre el 40 % y el 60 % [29]. En los casos en los que la persona no produzca esputo, este deberá ser inducido mediante una nebulización. Además, para el tratamiento de la muestra hay un protocolo establecido el cual debe cumplirse para mantener la integridad de la muestra[5].

- **Detección Molecular:**
  La pruebas de detección molecular son pruebas basadas en la detección de componentes del complejo *M. Tuberculosis*, mediante la reacción en cadena de la polimerasa (PCR), en tiempo real. Esta prueba tiene la ventaja de detectar a su vez mutaciones genéticas que pueden dar resistencia a algunos medicamentos, por ejemplo, para la detección de la resistencia a la rifampicina (uno de los medicamentos del tratamiento) se tiene una sensibilidad del 95 % y una especificidad del 98 %. Dentro de este grupo de pruebas de detección molecular, se encuentran los LiPA (del inglés *line probe assay*), con la cual también se puede determinar la resistencia a fármacos del bacilo que infecto al paciente [5]. En general estas pruebas tienen una sensibilidad y especificidad alrededor del 85 % y el 97 % respectivamente y en conjunto con una baciloscopia confirmada ascienden a 90 % y 99 % [9].

  El método estándar para el diagnóstico de TB es el método Xpert MTB/RIF, en el cual se emplea la plataforma dada por GeneXpert con la que se automatizan los tres procesos necesarios para la PCR en un solo cartucho: preparación de las muestras, amplificación del ADN y detección de la tuberculosis[30].

- **Cultivos:**
  Los últimos métodos son los cultivos, los cuales tienen una sensibilidad entre el 90 % y el 96 %, y los resultados de las pruebas pueden tardar de dos a tres semanas. Los cultivos deben ser realizados en laboratorios con personal capacitado, y equipos especializados capaces de mantener la calidad del proceso [28]. A todo cultivo, además, según la resolución [5], se le deberá realizar una prueba molecular para determinar la resistencia a los fármacos.

Como se mostró, cada una de las pruebas presenta sus ventajas y desventajas, y según la disponibilidad puede que a un paciente sospechosos de TB se le hagan una o más pruebas. En cuanto a la baciloscopia, es una prueba sencilla pero que no presenta tanta confiabilidad como las otras dos, tiene la ventaja de que es menos costosa, y sus resultados se deben entregar al día siguiente [5]. Para las pruebas moleculares, se requiere de un equipo especializado para realizarlas, pero este no está siempre disponible en los centros médicos, sin embargo, presenta mejores valores de sensibilidad y especificidad que la baciloscopia, dan información sobre la resistencia a fármacos, y puede demorar menos de dos horas en procesar una muestra [5][9]. Por último los cultivos, son los métodos de mayor peso, con una alta sensibilidad y especificidad, pero sus resultados pueden tardar semanas, y requieren de personal y equipo muy especializado para que estos se lleven a cabo adecuadamente.

2.1.3. Tratamiento

Para el tratamiento de la TB pulmonar el médico puede iniciar el tratamiento aún si los resultados microbiológicos son negativos o desconocidos, a partir soporte clínico y los hallazgos radiológicos. Este tratamiento se puede hacer, en la mayoría de los casos, usando varios medicamentos por un periodo de 6 a 9 meses, a estos fármacos se los denomina de primera línea y son la Isoniazida, Rifampina, Etambutol y Pirazinamida, los cuatro funcionan en conjunto y se debe medir las dosis de acuerdo a las características de cada paciente [5][31]. Cuando se tiene TB resistente a uno o varios de los medicamentos de primera línea se debe

recurrir a otros tipos de tratamientos y se debe estar supervisando constantemente al paciente; además, se debe remitir al paciente con un experto para que sea este quien determine el tratamiento que será empleado. Estos casos de resistencia a los fármacos son de especial cuidado, por lo cual para Colombia en [5], se establece que el esquema de tratamiento debe definirse con un plazo máximo de 15 días.

2.1.4. COVID-19 y TB

La presión en el sistema de salud causada por la pandemia por COVID-19, según la OMS puede causar un retroceso en las medidas que se habían tomado para el tratamiento y la prevención de la TB, especialmente en países con altas tasas de TB [2]. Esto sucede porque, aunque las medidas de aislamiento en el corto plazo pueden disminuir el número de personas infectadas, este efecto puede verse contrarrestado por un empeoramiento en la calidad de vida de las personas, a causa de un aumento de la pobreza, lo que aumenta la brecha ya existente en la accesibilidad a los sistemas de salud. Además, la pandemia ha provocado que se destinen menos recursos al control de la TB, como consecuencia de la necesidad de desviar esos recursos hacia el manejo del COVID-19, por ejemplo, las máquinas GeneXpert que se utilizan para detectar molecular TB, pueden y son usadas ahora para realizar el diagnóstico de COVID-19 [2].

2.2 Historias Clínicas

Las historias clínicas de acuerdo con la Ley 23 de 1981, en el artículo 34 se definen como "El registro obligatorio de las condiciones de salud del paciente. Es un documento privado sometido a reserva que únicamente puede ser conocido por terceros previa autorización del paciente o en los casos previstos por la Ley" [32], la historia clínica es un documento privado ya que contiene detalles íntimos del paciente, y no puede ser usada fuera del marco asistencial sin la autorización previa de su titular, por la sensibilidad de los datos [32]. Con el surgimiento de nuevas tecnologías, en la actualidad las instituciones suelen hacer uso de las HCE, las cuales son un registro integral y cronológico de las condiciones de salud del paciente, que se encuentran rodeadas de sistemas de información y de aplicaciones que facilitan su uso [33]. Las HCE reemplazan los métodos pasados para la manipulación de las historias clínicas de los pacientes, y tienen el objetivo de ayudar y agilizar el acceso y ejercicio de los servicios de salud, así como el manejo de la información de los pacientes.

Otro de los beneficios del uso de las HCE, es la interoperabilidad de las historias clínicas, que hace referencia a la capacidad de varios sistemas para intercambiar información, de manera que la información es compartida y accesible desde cualquier punto de la red [33]. En Colombia las HCE y su interoperabilidad son algo nuevo que se reglamento en el año 2020 con la ley 2015 de enero de 2020, en ella se estableció un plazo máximo de 5 años para que se implementen los términos y condiciones para la interoperabilidad que serán dictados por el Ministerio de Salud y Protección Social [33]. Esto es importante ya que facilitará el intercambio de la información de los pacientes entre instituciones, y de esta manera poder prestar un mejor servicio de salud. Además, la interoperabilidad de las HCE también permitirá que en un futuro la creación de modelos computacionales, como el presentado en este trabajo, sean más fáciles de desarrollar, ya que tener unificada la forma de guardar la información de las HCE, permite que técnicas de NLP ser más sencillas de aplicar, al no tener que preocuparse por los diferentes formatos

de cada una de las instituciones; además, la capacidad de unificar bases de datos con la información de pacientes será más sencillo, lo que permitiría tener más datos con los que crear mejores modelos.

2.3 Procesamiento del Lenguaje Natural

El NLP es una área de la IA en la cual se busca dotar a los computadores de la capacidad de comprender y manipular el lenguaje natural [34][35]. Por lenguaje natural, se entiende el lenguaje que usamos los humanos para comunicarnos, el cual no es sencillo de describir y ni ponerlo en términos que un computador pueda entender, debido a diferentes aspectos como los tiempos verbales, el uso de palabras con significados que dependen del contexto, las expresiones, entre otros. Para el uso del NLP los computadores reciben el lenguaje ya sea en forma de texto o en forma de discurso hablado, en el contexto de este trabajo se trata de un lenguaje especializado en el área de medicina en formato de texto, proveniente de las historias clínicas.

Entre los usos del NLP se encuentran aplicaciones como hacer resúmenes automáticos, traducir texto, predecir la siguiente palabra de un texto, buscar información dentro del texto, hacer inferencias del contenido general del texto, y otras en las cuales se deba de alguna manera procesar el texto con el fin de resolver una tarea [35]. Para ello los modelos computacionales deben aprender uno o varios de los diferentes niveles del lenguaje humano, como la sintaxis, el léxico, el contenido semántico y morfológico, y unos más complicados como el contexto y el tipo de discurso [34]. Históricamente han habido tres aproximaciones para dotar a los computadores de estos niveles de entendimiento, la primera es basada en reglas dadas por un experto, en donde una persona propone algoritmos o reglas que el computador usa para procesar el texto, esta estrategia es difícil de implementar por la complejidad del lenguaje humano por lo que su uso ha ido disminuyendo. La segunda, más reciente y que ha mostrado mejores resultados,es la creación de modelos que se basan en el texto para inferir las reglas mediante algún mecanismo de aprendizaje, estas reglas inferidas dependerán del problema que se quiera solucionar, y en su desarrollo por lo general primero hay una representación numérica del contenido del texto y luego se emplean modelos de ML para inferir esas reglas. Por último están los modelos mixtos, que combinan estas dos aproximaciones usando el conocimiento dado por un experto y lo que los modelos puedan aprender por su cuenta del texto [21].

En las HCE, el NLP puede ser usado para buscar información, resumir las historias clínicas, o como es el caso de este trabajo, proponer un DSS para el diagnóstico de alguna enfermedad. Dependiendo de la aplicación se procede de diferentes maneras, en este caso como se quiere hacer un modelo que de apoyo en el diagnóstico de la TB usando la información de las HCE, primero se hace un preprocesamiento del texto, donde se limpia y se elimina el ruido dentro del texto, con el fin de eliminar componentes que no aportan información relevante; posteriormente, se busca una representación de texto en forma numérica, para que los modelos de ML puedan manejar los datos y estos modelos puedan identificar y aprender los patrones en el texto.

2.3.1. Preprocesamiento para NLP

El preprocesamiento es una etapa necesaria en los sistemas de NLP, ya que permite eliminar el ruido y componentes como símbolos que pueden hacer que los modelos no se desempeñen

de la mejor forma posible [36]. Las técnicas de preprocesamiento dependen de la aplicación, así como del contenido de la base de datos, sin embargo, hay procedimientos que se aplican en la mayoría de ocasiones, como convertir todo el texto a minúsculas y eliminar o transformar símbolos fuera del abecedario [36].

A su vez, existen técnicas más complejas de preprocesamiento, como estandarizar palabras con significados similares (normalización), extraer o identificar las raíces de las palabras (*Stemming* y *Lemmatization*), o eliminar palabras sin significado (*stopwords*) [36]. En tareas como la de predecir la siguiente palabra de un texto, este tipo de preprocesamiento no es usado ya que reduce el número de palabras y, por ejemplo con *Stemming*, las transforma. Sin embargo, en casos de clasificación de textos y con bases de datos muy reducidas, donde no es necesario aprender elementos como la sintaxis de las oraciones, estas técnicas se emplean para mejorar el desempeño de los modelos.

2.3.2. Representación de Texto

Para representar el texto se hace un proceso de tokenización, el cual consiste en separar por palabras, sub-palabras o caracteres el texto. Con esta separación se puede asignar un número a cada token y usarlo para extraer características y entrenar modelos de ML, o bien para usar ML directamente. De esta representación salen los n-gramas que son subsecuencias de elementos, en este caso los tokens, de una secuencia dada, que seria el texto. Por ejemplo, para la frase: *esto es una oración*, los tokens separando por palabras serian: *esto, es, una, oración*; en este caso se están representando con unigramas. Los bigramas que se pueden tomar de la misma oración serian: *esto es, es una, una oración.* De la misma manera se pueden obtener n-gramas donde n es la longitud de la subsecuencias.

Las características que se extraen luego de la tokenización varían respecto al tipo de problema que se pretende solucionar, por ejemplo en [37] se usan características que están relacionadas con entidades (conceptos médicos), como puede ser distancias entre palabras y representaciones en otro espacio (*embeddings*) de las entidades, para encontrar asociaciones entre drogas y conceptos médicos. Para la clasificación de textos se suelen usar características que permitan extraer información de contenido tanto general como especifico del texto.

Entre los modelos más comunes para hacer la representación de documentos, se encuentra el modelo de bolsa de palabras (*bag of words*). En este modelo se tiene un conjunto de palabras o secuencias de palabras (*bag of n-gramas*), y se hace un conteo de cuántas veces aparece dicha palabra o secuencia en el documento. El conjunto de palabras que componen la bolsa de palabras se obtienen de un conjunto de documentos. Luego de hacer el conteo de la aparición de cada palabra, los documentos quedarán representados por un vector de la dimensión de la bolsa de palabras, que contiene el conteo de dichas palabras. El problema de esta representación, es que el simple conteo de palabras no siempre es una buena representación del contenido del texto, ya que no se tienen en cuenta cuestiones como relaciones entre palabras, o el orden de las palabras en la construcción de las frases, además, la dimensión del vector depende de la cantidad de palabras la cual suele ser muy alta, y la cantidad de ceros en el vector normalmente es elevada.

Para solucionar los problemas del modelo de bolsa de palabras, se usan métodos de extracción de características como TF-IDF (del inglés, *term-frequency - inverse document frequency*) con el que se busca una estadística numérica que refleje la importancia de una palabra en un documento de una colección. También se encuentran los *embeddings* que son modelos en los

que cada palabra esta representada por un vector de una dimensión fija, con los *embeddings* las palabras o secuencias de palabras con significados similares o que estén asociadas de alguna manera, tendrán una representación vectorial similar.

2.3.2.1 TF-IDF

En grandes conjuntos de documentos, aparecen palabras que no tienen mucha información relevante, estas palabras se suelen quitar con etapas de preprocesamiento, sin embrago, seguirán existiendo palabras que para una tarea de clasificación específica no aporten información relevante. Si se usan directamente todas las frecuencias de conteo como con el modelo de bolsa de palabras, los clasificadores no encontrarán palabras interesantes que permitan clasificar los documentos [38]. Con el modelo de TF-IDF se construye un vocabulario a partir de las palabras que aparecen en el conjunto de documentos, y para cada documento se obtiene un vector que contiene un valor por cada palabra, este valor es el resultado de la multiplicación entre el conteo de la aparición de cada palabra en el documento (TF), con el inverso del número de documentos en los que aparece dicha palabra (IDF).

Para el cálculo de IDF, se hacen ciertas modificaciones para evitar divisiones por cero en el inverso y se aplica un logaritmo a la función inversa. La ecuación para calcular los IDF usada es [38]:

$$IDF(t) = \log \frac{1+n}{1+df(t)} + 1 \tag{2.1}$$

Donde n es el número total de documentos y $df(t)$ es el número de documentos que contienen un término t.

Luego de calcular los TF-IDF para cada documento, se obtienen un vector que los representa, a este vector se le puede aplicar una normalización para que los valores sean más comparables entre documentos.

2.3.2.2 Embeddings

Los *embeddings* son un método de representación de texto donde a cada palabra se le asigna un vector. La idea es que palabras con un significado similar o que estén asociadas tengan una representación vectorial similar. Estos vectores suelen ser del orden de los decenas o cientos, a diferencia de modelos de representación como el de bolsa de palabras donde los vectores pueden dimensiones de miles, cientos de miles o más, y no hay una asociación entre palabras [39].

Existen diferentes formas de obtener los *embeddings*, se pueden obtener a partir de una capa de un modelo de ML entrenándolo para una tarea específica de forma supervisada, o también se pueden obtener de forma no supervisada a partir de estadísticas de los documentos [39]. En el método supervisado existe el problema de que el proceso de aprendizaje es lento y requiere de un conjunto de entrenamiento grande. Por el contrario, los métodos no supervisados, trabajan de forma más eficiente y no requieren de grandes conjuntos de datos en el caso de la tarea de clasificación [39].

Uno de los métodos más populares para hacer *embeddings* es el de Word2Vec [40], en el que se tienen dos algoritmos diferentes para de obtener esos *embeddings*, en la figura 2.2 se muestra una representación de esos dos algoritmos. El primero COBW (del inglés *Continuous*

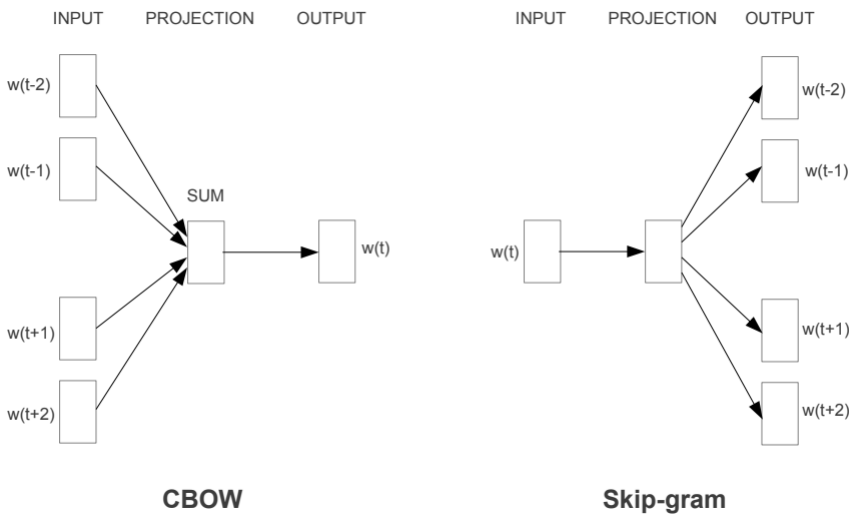


Figura 2.2: Representación de los algoritmos COBW y *spkip gram* para la construcción de los *embeddings* [40].

Bag-of-Words) genera el *embeddings* a partir de la tarea de predecir la palabra en medio de una ventana usando palabras que están a su alrededor. El segundo *Skip Gram*, genera el *embedding* a partir de predecir las palabras alrededor de una palabra específica [40]. Cuando los modelos ya se entrenaron para sus respectivas tareas, se usan los pesos de la red para obtener los *embeddings*.

2.3.3. Técnicas de Aprendizaje Automático

El ML es una rama de la IA en la que se busca dotar a las máquinas de una capacidad de aprendizaje. Los algoritmos de ML buscan aprender de los datos y que las máquinas no dependan de reglas humanas para desarrollar sus tareas. La ventaja de estos algoritmos es su velocidad y el desempeño que en algunos casos supera al humano. En general existen tres tipos de ML [41]:

- **Aprendizaje Supervisado** Con el aprendizaje supervisado se busca a partir de datos predecir un valor como la pertenencia a una clase (clasificación) o un valor continuo (regresión).
- **Aprendizaje no Supervisado** En el aprendizaje no supervisado no se tiene un valor correcto respecto al problema, sino que se busca explorar los datos para hacer un agrupamiento, o una reducción de la dimensionalidad de los datos. Esto sirve para destacar información relevante en los datos, visualizar grandes cantidades de datos, agrupar conjuntos de datos, entre otros.
- **Aprendizaje Reforzado** En el aprendizaje reforzado la meta es desarrollar sistemas que vayan mejorando su desempeño en una tarea a partir de su interacción con el ambiente.

De los tres tipos de aprendizaje automático son de especial interés en el NLP el aprendizaje supervisado y el no supervisado, dentro del aprendizaje supervisado existen diferentes algoritmos para realizar las tareas de clasificación o regresión, que sirven por ejemplo en la tarea de predecir una palabra como los dos algoritmos de Word2Vec. Los tres algoritmos de ML que se usaran en este trabajo son ANN, máquinas de soporte vectorial del inglés *Support Vector Machine* (SVM), y bosques aleatorios del inglés *Random Forest* (RF), ya que son modelos comunes y ampliamente usados dentro del ML por su capacidad de generalizar, todos ellos se usan para la tarea de clasificación. Los algoritmos no supervisados se usan más que todo para realizar una representación de los datos y reducir la dimensionalidad de las características.

2.3.3.1 Redes Neuronales Artificiales

Las ANN son modelos computacionales que están inspirados en cómo se pensaba que funcionaba el cerebro. Las ANN clásicas se basan en un conjunto de neuronas que están conectadas entre si por capas, cada neurona recibe ya sea entradas del modelo o salidas de otras neuronas, luego, realizan una suma ponderada de cada una de las entradas y se aplican una función de activación que añade no linealidades al sistema; estas no linealidades son una de las fortalezas de estos modelos ya que permiten hallar patrones en espacios de características que no son necesariamente lineales [41]. En la figura 2.3 se muestra una representación gráfica de las neuronas que componen a la ANN, las entradas de la neurona se representan con x_m, y los pesos para la suma ponderada son w_m, estos pesos se van actualizando en el proceso de entrenamiento de la red. En la figura 2.4 se muestra una ANN clásica con solo una capa intermedia.

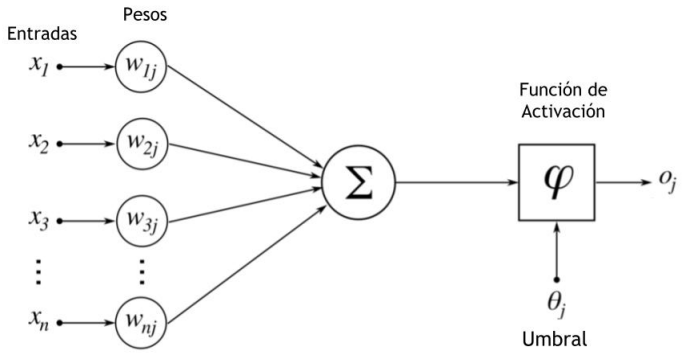


Figura 2.3: Esquema de una neurona. Imagen modificada de [42].

El proceso de aprendizaje de las ANN es un proceso iterativo, donde en primer lugar, se pasan las entradas a través de la red y se calcula el error en la salida respecto a un valor esperado, ya con el error, mediante un algoritmo de retropropagación, se van actualizando los pesos de cada una de las neuronas. A las características del modelo, como el número de iteraciones y número de neuronas, se les denomina hiperparámetros. Los hiperparámetros se deben explorar con el fin de buscar los que mejor desempeño tengan en la tarea [41].

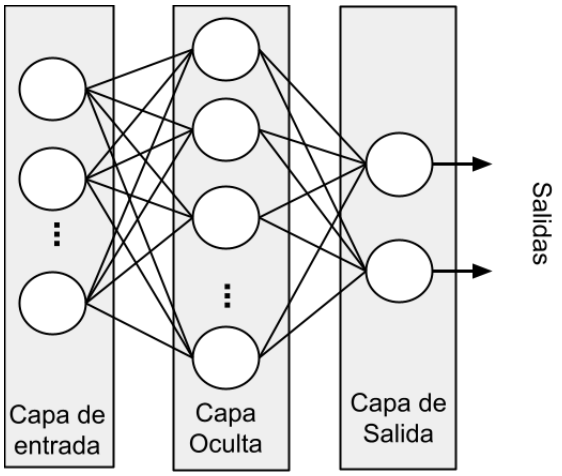


Figura 2.4: Esquema de una red neuronal de una sola capa oculta.

2.3.3.2 Máquinas de Soporte Vectorial

En las SVM para la clasificación, el objetivo es maximizar el margen, el cual se define como la distancia entre el hiperplano de separación y las muestras de entrenamiento más cercanas a este hiperplano, a estas muestras más cercanas se les denomina vectores de soporte [41]. En la figura 2.5 se muestra un ejemplo del uso de una SVM con un conjunto de datos linealmente separables, en la imagen se puede apreciar como se pueden plantear diferentes planos de separación de los datos (H1,H2 y H3), sin embargo, solo el plano H3 fue obtenido mediante una SVM que separa de forma óptima los datos.

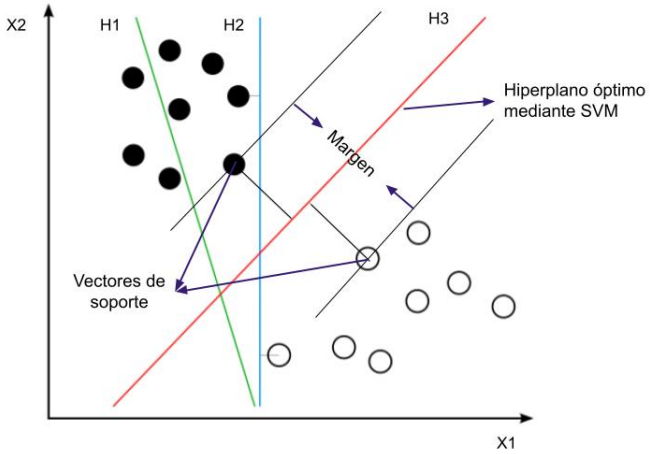


Figura 2.5: Ejemplo del funcionamiento de una SVM para un conjunto de datos linealmente separable. Imagen adaptada de [43].

Por lo general, no se tienen variables linealmente separables, por lo que es necesario relajar el modelo y permitir cierta tolerancia a clasificaciones erróneas. El parámetro mediante el cual se controla la penalización se suele denotar con la variable C. Además, dentro de las SVM se suele hacer un mapeado del espacio de las características a un espacio de mayor dimensión

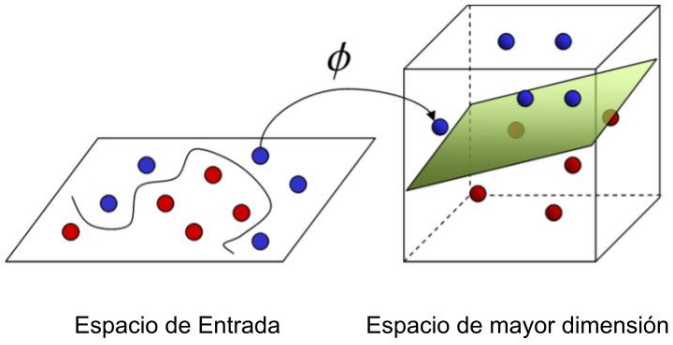


Figura 2.6: Representación del mapeo a otra dimensión para hacer la clasificación con SVM. Imagen adaptada de [44].

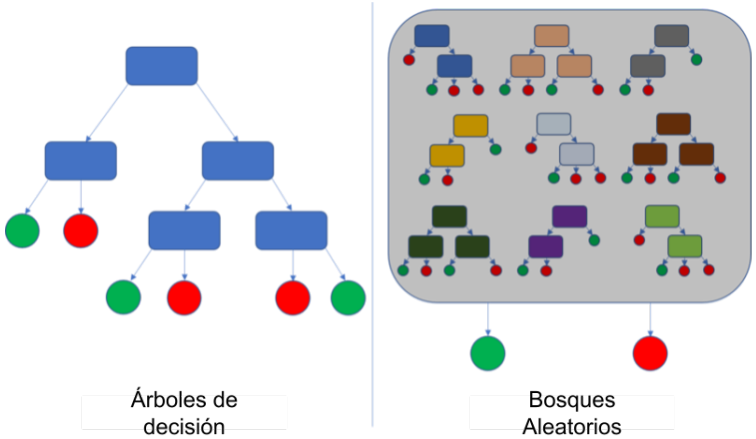


Figura 2.7: Esquema del modelo de árbol de decisión y RF [45]

buscando que en esa dimensión se pueda encontrar un mejor hiperplano para la separación de clases [41]. En la figura 2.6 se muestra el mapeo del espacio de características en un espacio de mayor dimensión, donde es más fácil encontrar el hiperplano óptimo de separación.

2.3.3.3 Bosques Aleatorios

El modelo de RF junta un conjunto de arboles de decisión que se entrenan por separado. Estos árboles de decisión se entrenan con el objetivo de maximizar la ganancia de información. Sin embargo, uno solo de estos arboles no suele ser capaz de resolver un problema de clasificación por si solo. Lo que logra RF es juntar un conjunto de clasificadores débiles para la tarea de clasificación. En la figura 2.7 se muestra un esquema de un árbol de decisión y un esquema de RF donde se aprecia que este está compuesto de varios arboles de decisión. Un paso importante para que el modelo funcione es presentar tanto muestras como características diferentes a cada árbol, con el fin de que cada uno se especialice en una tarea específica y aprendan patrones distintos entre ellos, luego la clasificación se hace mediante una votación de los arboles de decisión [41].

Capítulo 3

METODOLOGÍA

En este capítulo se presenta la metodología implementada en el trabajo, en primer lugar, se construyó una base de datos usando las HCE de pacientes sospechosos de TB pulmonar, en esta base de datos se encuentran las anotaciones de los médicos realizadas de forma previa a la realización de las pruebas microbiológicas y al inicio del tratamiento de los pacientes, estas anotaciones se guardaron en formato *txt* un archivo por cada paciente. Posteriormente, con la base de datos ya construida, se realizó un preprocesamiento del texto, con el que se busca limpiar y transformar palabras para mejorar la calidad de los datos, luego, se extrajeron características que representan en forma numérica el contenido del texto, y finalmente, se entrenan los modelos de ML con el objetivo de diagnosticar TB pulmonar activa en pacientes sospechosos de TB. Los parámetros e hiperparámetros de los métodos de extracción de características y modelos de ML fueron explorados mediante validación cruzada, la separación en conjuntos de entrenamiento y prueba se hizo antes de la extracción de características, ya que en los métodos de representación de texto empleados, las características extraídas de un documento dependen también de los datos presentes en los demás documentos. En la figura 3.1 se muestra un esquema general de la metodología descrita en este párrafo, donde se observa el proceso de la creación de la base de datos, luego el preprocesamiento, y la validación cruzada que divide los datos previo a la extracción de características y entrenamiento de modelos de ML.

3.1 Construcción de la Base de Datos

Para la construcción de la base de datos se cuenta con las historias clínicas de 165 pacientes que fueron sospechosos de TB en los años 2017, 2018 y 2019. Para todos los pacientes se registró dentro del hospital en un documento aparte algunos de sus datos personales, la fecha de inicio de síntomas, la fecha del inicio del tratamiento (en caso de haber sido tratado), y se guardaron los resultados de las pruebas microbiológicas realizadas a cada paciente así como la fecha de realización de cada prueba, esto se hizo dentro del proyecto con el fin de tener separada la información relevante de cada paciente. Por otro lado, el diagnóstico definitivo de cada paciente fue dado por un médico especialista que esta vinculado al proyecto, quien tuvo en cuenta que a los pacientes se les podía realizar una, dos o las tres pruebas diagnosticas (baciloscopia, cultivo o prueba molecular), dependiendo de los resultados de cada prueba el especialista dio su diagnóstico definitivo.

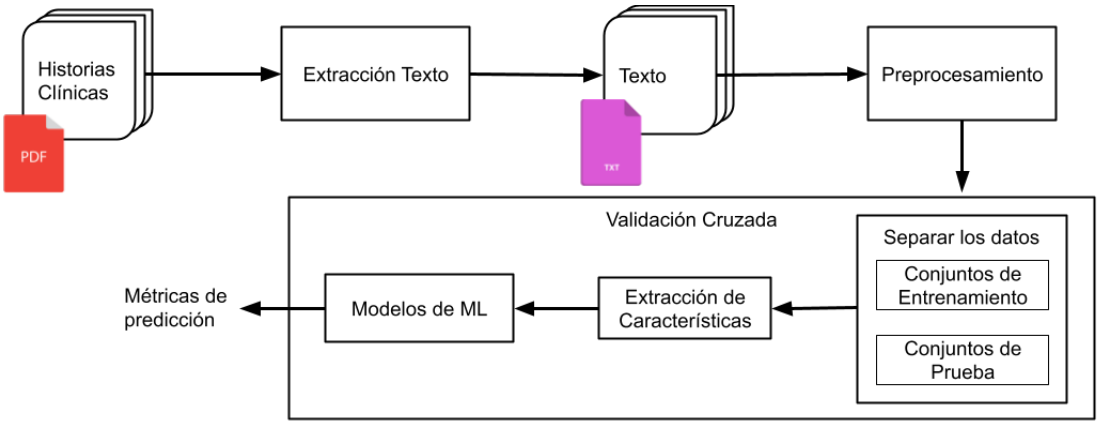


Figura 3.1: Esquema general de la metodología implementada

3.1.1. Estructura de las Historias Clínicas

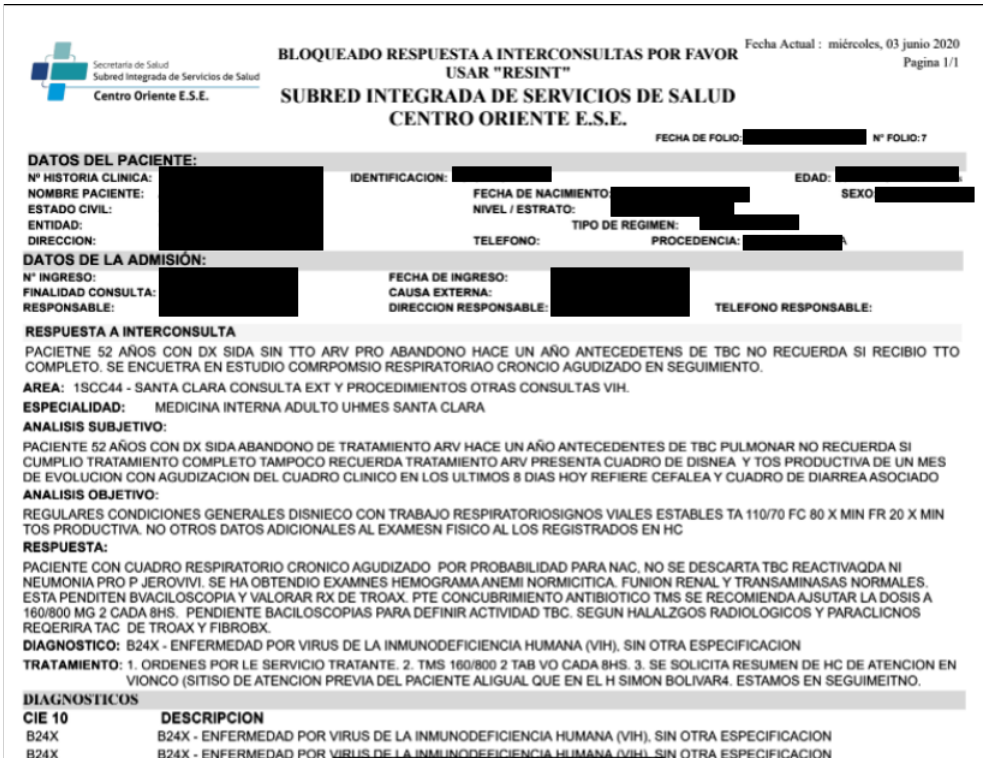
Las HCE fueron suministradas por el proyecto en formato PDF (del inglés, *Portable Document Format*), ya que era la forma más cómoda de exportar los datos que estaban dentro del sistema. Dentro de la HCE se tiene organizado cronológicamente los datos del paciente por folios, cada folio hace referencia a un instante en el que un médico hace un reporte del paciente en la HCE, y dentro de cada folio la forma en que es guardada la información depende de la razón por la cual el paciente se dirigió al centro de salud. Para este trabajo fue necesario identificar como es la estructura de los folios relacionados con el diagnóstico de la TB, para ello se tomaron aleatoriamente 20 pacientes y se buscaron folios con fechas cercanas al inicio del tratamiento que tuvieran información relacionada con la TB. De esta exploración se lograron identificar dos estructuras de folio que se usan en la mayoría de los casos. En las figuras 3.2 y 3.3 se muestran estas dos estructuras generales.

Por otro lado, la cantidad de páginas de las HCE varia demasiado entre pacientes, ya que esta depende de la cantidad de veces que el sujeto se dirigió al centro de salud por algún motivo médico. Se pueden tener PDFs de 1 sola página hasta más de 4000 páginas, por lo que es necesario buscar la manera de que solo se tomen folios que sean relevantes para el diagnóstico de la TB en la construcción de la base de datos.

3.1.2. Extracción de Texto

Para la extracción de texto de los PDFs, se tuvo en cuenta la opinión de los médicos pertenecientes al proyecto, con los que se discutió la forma en cómo se diligencian las HCE y como se realizó el diagnóstico. Se acordó usar solo la información previa al diagnóstico de TB que en teoría debe ser anterior a la fecha de inicio de tratamiento y a la fecha de la realización de la primera baciloscopia.

El algoritmo para extraer el texto, consta de los siguientes pasos. 1) Leer los PDFs y pasarlos a formato de texto. 2) Identificar el inicio y fin de página y de folio. 3) Determinar si el folio tiene una de las dos estructuras de la figuras 3.2 y 3.2. 4) Obtener la fecha de cada folio. 5) Si la fecha de folio está en un intervalo de 30 días antes de la fecha de inicio de tratamiento o de realización de la baciloscopia, se toma como candidato para guardar su texto. 6) De los



Secretaría de Salud
Subred Integrada de Servicios de Salud
Centro Oriente E.S.E.

BLOQUEADO RESPUESTA A INTERCONSULTAS POR FAVOR USAR "RESINT"

Fecha Actual : miércoles, 03 junio 2020
Pagina 1/1

SUBRED INTEGRADA DE SERVICIOS DE SALUD CENTRO ORIENTE E.S.E.

FECHA DE FOLIO: [redacted] Nº FOLIO:7

DATOS DEL PACIENTE:

Nº HISTORIA CLINICA: [redacted] IDENTIFICACION: [redacted] EDAD: [redacted]
NOMBRE PACIENTE: [redacted] FECHA DE NACIMIENTO: [redacted] SEXO [redacted]
ESTADO CIVIL: [redacted] NIVEL / ESTRATO: [redacted]
ENTIDAD: [redacted] TIPO DE REGIMEN: [redacted]
DIRECCION: [redacted] TELEFONO: [redacted] PROCEDENCIA: [redacted]

DATOS DE LA ADMISIÓN:

Nº INGRESO: [redacted] FECHA DE INGRESO: [redacted]
FINALIDAD CONSULTA: [redacted] CAUSA EXTERNA: [redacted]
RESPONSABLE: [redacted] DIRECCION RESPONSABLE: [redacted] TELEFONO RESPONSABLE: [redacted]

RESPUESTA A INTERCONSULTA

PACIETNE 52 AÑOS CON DX SIDA SIN TTO ARV PRO ABANDONO HACE UN AÑO ANTECEDETENS DE TBC NO RECUERDA SI RECIBIO TTO COMPLETO. SE ENCUETRA EN ESTUDIO COMRPOMSIO RESPIRATORIAO CRONCIO AGUDIZADO EN SEGUIMIENTO.

AREA: 1SCC44 - SANTA CLARA CONSULTA EXT Y PROCEDIMIENTOS OTRAS CONSULTAS VIH.

ESPECIALIDAD: MEDICINA INTERNA ADULTO UHMES SANTA CLARA

ANALISIS SUBJETIVO:

PACIENTE 52 AÑOS CON DX SIDA ABANDONO DE TRATAMIENTO ARV HACE UN AÑO ANTECEDENTES DE TBC PULMONAR NO RECUERDA SI CUMPLIO TRATAMIENTO COMPLETO TAMPOCO RECUERDA TRATAMIENTO ARV PRESENTA CUADRO DE DISNEA Y TOS PRODUCTIVA DE UN MES DE EVOLUCION CON AGUDIZACION DEL CUADRO CLINICO EN LOS ULTIMOS 8 DIAS HOY REFIERE CEFALEA Y CUADRO DE DIARREA ASOCIADO

ANALISIS OBJETIVO:

REGULARES CONDICIONES GENERALES DISNIECO CON TRABAJO RESPIRATORIOSIGNOS VIALES ESTABLES TA 110/70 FC 80 X MIN FR 20 X MIN TOS PRODUCTIVA. NO OTROS DATOS ADICIONALES AL EXAMESN FISICO AL LOS REGISTRADOS EN HC

RESPUESTA:

PACIENTE CON CUADRO RESPIRATORIO CRONICO AGUDIZADO POR PROBABILIDAD PARA NAC, NO SE DESCARTA TBC REACTIVAQDA NI NEUMONIA PRO P JEROVIVI. SE HA OBTENDIO EXAMNES HEMOGRAMA ANEMI NORMICITICA. FUNION RENAL Y TRANSAMINASAS NORMALES. ESTA PENDITEN BVACILOSCOPIA Y VALORAR RX DE TROAX. PTE CONCUBRIMIENTO ANTIBIOTICO TMS SE RECOMIENDA AJSUTAR LA DOSIS A 160/800 MG 2 CADA 8HS. PENDIENTE BACILOSCOPIAS PARA DEFINIR ACTIVIDAD TBC. SEGUN HALALZGOS RADIOLOGICOS Y PARACLICNOS REQERIRA TAC DE TROAX Y FIBROBX.

DIAGNOSTICO: B24X - ENFERMEDAD POR VIRUS DE LA INMUNODEFICIENCIA HUMANA (VIH), SIN OTRA ESPECIFICACION

TRATAMIENTO: 1. ORDENES POR LE SERVICIO TRATANTE. 2. TMS 160/800 2 TAB VO CADA 8HS. 3. SE SOLICITA RESUMEN DE HC DE ATENCION EN VIONCO (SITISO DE ATENCION PREVIA DEL PACIENTE ALIGUAL QUE EN EL H SIMON BOLIVAR4. ESTAMOS EN SEGUIMEITNO.

DIAGNOSTICOS

CIE 10	DESCRIPCION
B24X	B24X - ENFERMEDAD POR VIRUS DE LA INMUNODEFICIENCIA HUMANA (VIH), SIN OTRA ESPECIFICACION
B24X	B24X - ENFERMEDAD POR VIRUS DE LA INMUNODEFICIENCIA HUMANA (VIH), SIN OTRA ESPECIFICACION

Figura 3.2: Primera estructura general del folio encontrada relevante para el diagnóstico de TB

DATOS DEL PACIENTE

Nº Historia Clínica: FECHA DE FOLIO: [redacted] Nº FOLIO: 1
Nombre Paciente: [redacted] Identificación: [redacted]
Fecha Nacimiento: Estado Civil:
Dirección: Teléfono:
Procedencia: Ocupación:

DATOS DE AFILIACIÓN

Entidad: [redacted] Régimen:
Plan Beneficios: Nivel - Estrato:
Area de servicio: Centro de Atención:
Fecha HC:

DATOS DE LA ADMISIÓN:

Finalidad Consulta: No_Aplica Causa Externa: Enfermedad_General
Responsable: Direccion Responsable: Teléfono Responsable:
Centro de Atención: 1SC - UMHES SANTA CLARA Área de servicio 1SCC15 - SANTA CLARA CONSULTA EXT Y PROCEDIMIENTOS CONSULTA NEUMOLOGIA ADULTO

NOTA EVOLUCION:

FECHA: [redacted]
NOMBRE:
EDAD
SEXO
IDENTIFICACIÓN
PROCEDENCIA:
INDICACIÓN: CONSOLIDACION A ESTUDIO

Previa firma consentimiento informado y bajo anestesia tópica se realiza fibrobroncoscopia con los siguientes hallazgos:
VIA: Trasnasal derecha
FARINGE: Normal.
LARINGE: Epiglotis normal. Valléculas normales. Amígdalas linguales normales. Senos piriformes normales. Aritenoides y bandas ventriculares normales. Cuerdas vocales normales, movilidad normal a la fonación.
TRAQUEA: Presencia de secreción mucopurulenta proveniente de ambos bronquios fuentes, mucosa normal, no lesiones.
CARINA: normal, central, con secreción mucopurulenta
ARBOL BRONQUIAL DERECHO: Divisiones lobares y segmentarias normales. Presencia de secreción mucopurulenta
ARBOL BRONQUIAL IZQUIERDO: Bronquio fuente, divisiones lobares y segmentarias con secreción mucopurulenta
PROCEDIMIENTOS: Lavado broncoalveolar en el segmento anterior del lóbulo superior derecho
COMPLICACIONES: no

Figura 3.3: Segunda estructura general del folio encontrada relevante para el diagnóstico de TB. El encabezado del folio es igual que el de la figura 3.2

folios candidatos se seleccionan los 5 folios más próximos a estas fechas. 7) Extraer el texto de los folios. En la figura 3.4 se muestra un esquema de este proceso. Cada uno de los pasos se explican con más detalle a continuación.

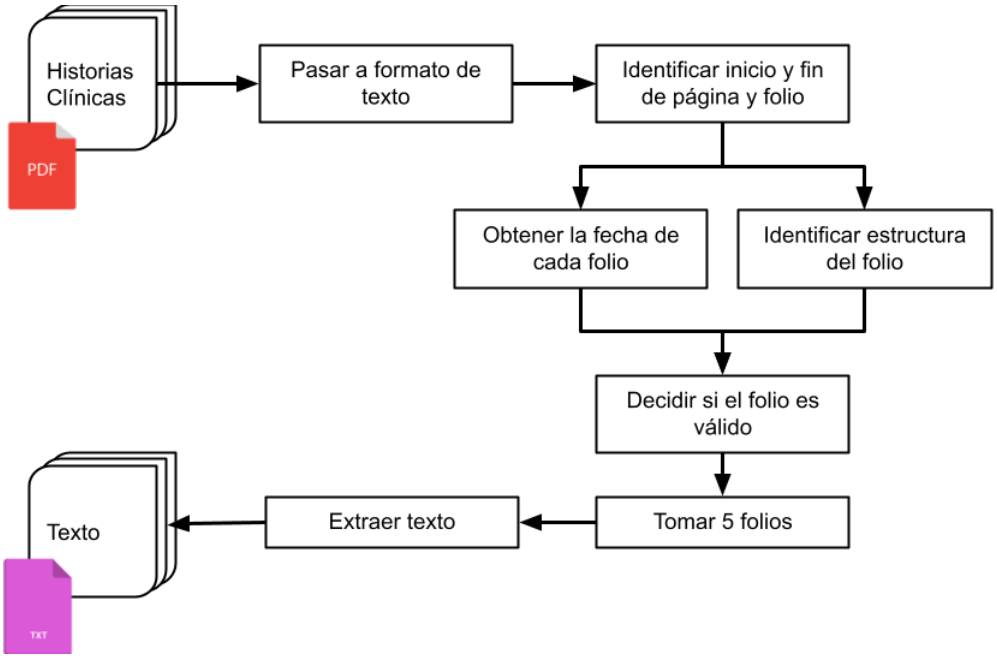


Figura 3.4: Esquema del proceso para la extracción de texto de las historias clínicas

Paso 1:

Para pasar los PDF a formato texto se uso la librería pdfminersix de Python [46], que esta especializada en la lectura de PDF's y contiene diferentes herramientas para ello. La función *extract_text* de la librería, permite leer todo el PDF y devuelve su texto en un formato especifico, en este caso UTF-8. Para leer los PDFs hay que tener en cuenta que dentro del formato los caracteres no están organizados de forma secuencial, como si lo están por ejemplo en formatos como *doc*. Lo que hace la librería es identificar cada uno de los caracteres, luego calcula la distancia entre ellos y mediante heurísticas, los agrupa de forma que se van construyendo las palabras, frases y párrafos. Además, esta función recibe parámetros de como se realiza la lectura del texto en el caso de encontrase por ejemplo en columnas.

Paso 2:

Gracias a que las HCE tiene un formato específico, fue posible identificar donde comenzaba cada página con el encabezado y el pie de página. Para identificar los folios se buscaron las palabras *fecha de folio:*, estos patrones de palabras se hallan dentro del texto usando expresiones regulares, que son secuencias de caracteres con las que se puede definir un patrón de búsqueda.

Pasos 3 y 4:

También con expresiones regulares se buscó el patrón que se usa en las fechas y el patrón para definir si el folio tiene alguna de las estructuras de las figuras 3.2 y 3.3.

Paso 5 y 6:

Dento de las HCE como ya se mencionó antes, hay información que puede no ser relevante para el estudio que se está realizando. Por esta razón se decidió solo tomar los folios que

estén cercanos a la fecha de inicio de tratamiento o de la primera prueba de baciloscopia. Estas fechas en general correspondían a lo que se hallaba en las HCE, sin embargo, habían casos que al buscar las fechas dentro de las HCE, estas no estaban o no correspondían a nada relacionado con la TB. Estos errores se deben a posibles errores en la digitación o por problemas técnicos, que se escapan del trabajo. Los errores en las fechas ocasionaron que el algoritmo no encontrara folios válidos para algunos pacientes, disminuyendo el tamaño de la base de datos. También se dieron casos en los que habían menos de cinco folios válidos para un paciente, dado que en ese periodo de 30 días simplemente no habían registros del paciente en la HCE. En la figura 3.5 se encuentra visualizada la cantidad de pacientes a los que se les encontró cierta cantidad de folios válidos, desde cero folios válidos hasta 5 folios válidos; en la figura 3.5 se puede observar que para la mayoría de pacientes fue posible obtener 5 folios válidos.

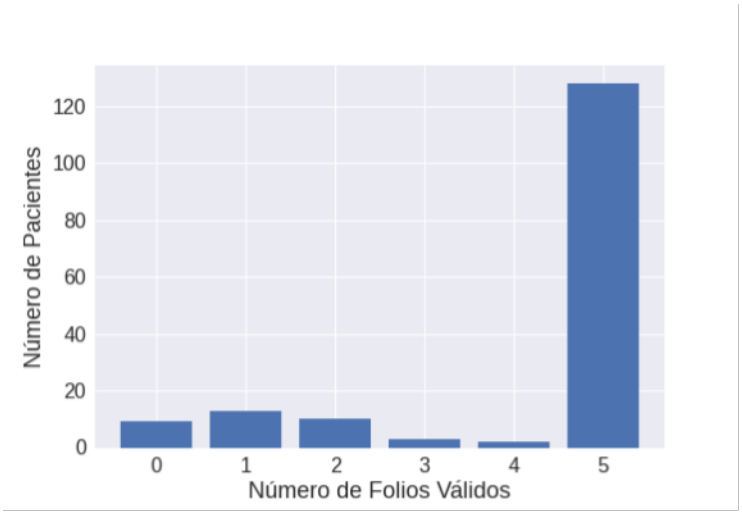


Figura 3.5: Conteo de pacientes con un número determinado de segmentos extraídos.

Paso 7:

El texto extraído consiste de las anotaciones que hace el médico sobre el paciente. En la figura 3.6 se muestra nuevamente el ejemplo de HCE de la figura 3.2, y se marca el segmento de texto donde están las anotaciones que hace el médico, a este segmento lo llamaremos segmento de interés. A su vez, en la figura 3.6, se muestra la ubicación de la fecha de folio que se mencionó en los pasos anteriores, con la cual se realizó la selección de los 5 folios más próximos a las fechas de inicio de tratamiento o de baciloscopia.

Para segmentar el segmento de interés, se ubicaron en el folio las palabras: *respuesta a interconsulta, respuesta interconsulta, nota evolución y nota de evolución*, las cuales indican el inicio del segmento. Para marcar el fin del segmento de interés se ubicaron segmentos de texto que coincidieran alguno de los patrones: *DIAGNOSTICOS Nombre Código* o *DIAGNOSTICOS CIE 10*. Esto como ya se mencionó se hizo teniendo en cuenta las sugerencias de los médicos del proyecto sobre que información tomar.

Por último, como se muestra en la figura 3.4, luego de extraer el texto, este fue almacenado en archivos con formato *txt*, un archivo por paciente. Estos archivos en pocos casos contienen información como el nombre y la edad de la persona. Esto sucedió porque aunque

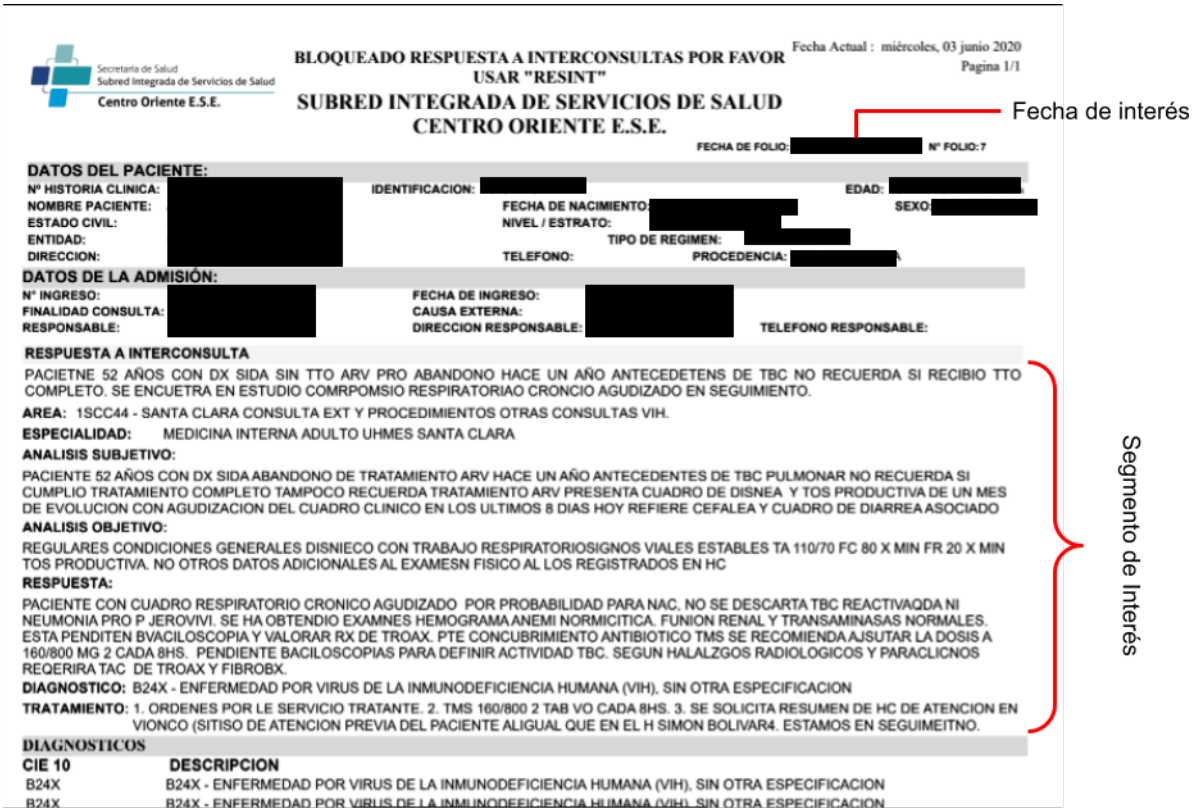


Figura 3.6: Ejemplo de la ubicación de la fecha y segmento de interés dentro de una HCE, para la extracción de texto.

la información del paciente está arriba de los segmentos de interés, en algunos casos como el ejemplo de la figura 3.3, la información aparece dentro del segmento de interés.

3.2 Preprocesamiento

Ya con los archivos en formato *txt*, los cuales conforman la base de datos, se procede a hacer un preprocesamiento de los datos, el cual consiste en limpiar el texto y remover las palabras sin significados (*stopwords*).

3.2.1. Limpieza de Texto

En la limpieza del texto se eliminaron caracteres que no fueran letras del abecedario español, incluyendo tildes, signos de puntuación, números y otros caracteres extraños, también se quitaron los dobles espacios y los saltos de línea, esto con el fin de quitar del texto elementos que no sirven y pueden confundir a los modelos que se aplicaron.

En este punto, se decide guardar en un archivo en formato *csv* los segmentos extraídos a cada paciente, junto con el diagnóstico dado por el especialista, y un texto en el cual se condensan todos los segmentos, además, dentro del archivo se descartaron los casos donde no se puedo tomar ningún segmento, quedando en total 151 pacientes, 116 con TB pulmonar

confirmada y 35 con TB pulmonar descartada, lo que evidencia un desbalance en la cantidad de pacientes en cada grupo, hecho que se tuvo en cuenta para las siguientes partes de la metodología. En la figura 3.7 se ve la cabecera del archivo *csv*, en donde cada uno de los 5 segmentos está en la columna px donde x es el número de segmento, en la columna *target* esta el diagnóstico, y en la columna *full text* se encuentran todos los segmentos del paciente juntos. Los textos de la columna *full text* son los que serán usados para el diagnóstico de la TB, sin importar si un paciente tiene menos de 5 segmentos extraídos.

ID	p1	p2	p3	p4	p5	target	year	Full text
[redacted]	evolucion adicional pte con dx sida cun cuadr...	area scu santa clara urgencias consulta urgen...	area scu santa clara urgencias observacion ad...	area scu santa clara urgencias observacion ad...	evolucion de medicina interna medicina genera...	TB CONFIRMADA	2017	evolucion adicional pte con dx sida cun cuadr...
[redacted]	evolucion medicina interna paciente masculino...	respuesta inerconsulta neumologia paciente ha...	paciente de años a quien el dia de hoy bk de...	paciente reportado a sdis para la caracteirza...	ingreso piso medicina interna sc umhes santa ...	TB CONFIRMADA	2017	evolucion medicina interna paciente masculino...
[redacted]	ss rx de torax portatil pos paso de cavafix p...	analisis subjetivo neurologia analisis objeti...	rutina para el dia de mañana principal dx in...	se abre folio para responder interconsulta es...	nota procedimiento previa asepsia y antisepsi...	TB DESCARTADA	2017	ss rx de torax portatil pos paso de cavafix p...
[redacted]	paciente de años de edad nativo y residente ...	evolucion medicina hopsitalaria medicina inte...	paciente habitante de calle con cuadro respir...	area scu santa clara urgencias observacion ad...	evolucion medicina hopsitalaria medicina inte...	TB CONFIRMADA	2017	paciente de años de edad nativo y residente ...
[redacted]	nota tarde medicina inetrna paciente de años...	toxicologia paciente de años con idx de sind...	se abre folio para formulacion no_aplica fern...	evolucion medicina interna medicina hospitala...	evolucion medicina interna medicina hospitala...	TB CONFIRMADA	2017	nota tarde medicina inetrna paciente de años...

Figura 3.7: Cabecera del archivo *csv* creado luego de la limpieza de texto.

3.2.2. Stopwords

Las *stopwords* son palabras que carecen de significado dentro de un lenguaje, como los artículos, pronombres y preposiciones. Remover estas palabras es importante porque en el contexto de este trabajo, estas palabras sin significado no aportan información en la tarea de clasificación. Para la remoción de las *stopwords* se uso la libraría *Natural Language Toolkit* (NLTK) [47], que es una librería de Python desarrollada especialmente para el NLP, en la que las *stopwords* para el lenguaje español ya se encuentran guardadas dentro de la librería. Por otro lado, las *stopwords* no son las únicas palabras que pueden no tener significado o relevancia y ser ruido para el diagnóstico de la TB, dentro de las historias clínicas se encontraron elementos como abreviaciones y errores de ortografía que influirán de la misma manera en el desempeño de los modelos, sin embargo, con un posprocesamiento de las características, se espera reducir el efecto de este ruido.

3.3 Validación Cruzada

La validación cruzada se usa para probar la capacidad de los modelos para generalizar, y encontrar los hiperparámetros que mejor se ajustan a la solución de un problema, también busca que los resultados de los modelos sean independientes de la partición entre datos de prueba y de entrenamiento. En este caso se usó una validación cruzada con tres subconjuntos de datos, donde cada uno contiene datos para entrenamiento y datos de prueba, se tomaron solo tres subconjuntos porque la base de datos no es muy grande, y se necesita que haya una buena cantidad de datos en los conjuntos de prueba para evaluar correctamente los

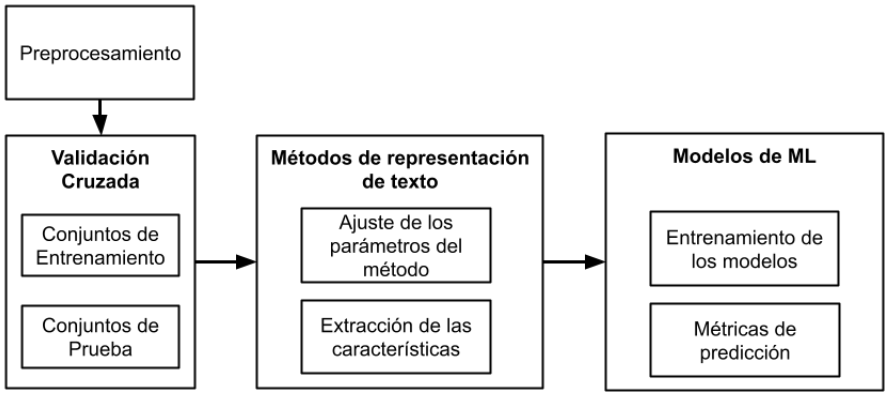


Figura 3.8: Esquema de la validación cruzada usada para la exploración de modelos.

modelos; además, en cada conjunto de prueba y entrenamiento se conserva la proporción de casos confirmados y descartados de TB, y todos los modelos probados usan los mismos subconjuntos de forma que se pueda hacer una comparación adecuada entre los modelos.

En la figura 3.8 se muestra como fue implementada la validación cruzada, en esta figura se observa que los datos son separados previo a la extracción de características, esto se hace de así ya que en los métodos de representación de texto usados, las características que representan a cada documento no depende únicamente del mismo, sino de la información contenida en los demás documentos, por ejemplo, en la construcción del vocabulario de términos o en el calculo de IDF(t) (ver ecuación 2.1). Además, al separar los datos antes de la extracción de características se simula de mejor manera lo que pasaría si se intenta aplicar el sistema a un documento nuevo, al cual se le extraerían las características basándose en lo encontrado de antemano en la base de datos, y luego si se usarían los modelos de ML entrenados.

3.4 Técnicas de Extracción de Características

La extracción de características se hizo a través de los modelos de TF-IDF y Word2Vec. En el método TF-IDF se usó la formula 2.1 para el cálculo de IDF, y se realizó una normalización l2 a los vectores de cada documento. Además, con el método TF-IDF se exploraron diferentes formas de construir el vocabulario a partir de diferentes combinaciones de n-gramas: 1) unigramas, 2) unigramas y bigramas, 3) bigramas, 4) unigramas, bigramas y trigramas, 5) trigramas. Del vocabulario construido con las diferentes combinaciones de n-gramas, se tomó solo un conjunto de los términos que más frecuencia tuvieran en el conjunto de documentos, ya que si un término no aparece mucho entre los documentos es muy posible que no sea relevante, el número de términos seleccionados se fue variando encontrando que había que tomar entre 800 y 2000 términos para obtener buenos resultados. Cuando se tienen los vectores de características generados con TF-IDF estos presentan una alta dimensionalidad, por lo que se aplicó una reducción de la dimensionalidad con un número definido de componentes (también a explorar), luego, el vector resultante de esta reducción es el que se uso para entrenar los modelos.

Para el método de *embeddings* con Word2Vec, se probaron los dos modelos (COBW y *Skip*

Gram) para la construcción del *embedding*, se decidió usar estos modelos porque su tiempo de ejecución y resultados fueron mejores en comparación a si, por ejemplo, se tomara un *embedding* generado a partir de una capa en una red que predijera TB pulmonar. En estos modelos de *embeddings* se varió el tamaño de la ventana, que define la cantidad de palabras al rededor de las cuales se aplica el algoritmo, y se varió el número mínimo de veces que aparece una palabra para ser parte del vocabulario. Una vez ajustados los pesos de la red que generan los *embeddings* con el respectivo conjunto de entrenamiento, se sacan los vectores que representan las palabras de los documentos, dando como resultado una matriz del largo del documento que contiene los vectores de cada palabra, como los modelos de ML necesitan tener una entrada de dimensión fija, se extrajo por separado la media y el valor máximo como capa de agrupamiento (*pooling*) a lo largo del documento, dando como resultado un vector de dimensión igual al tamaño del *embedding* que representa cada documento.

Para el método de TF-IDF se uso la librería scikit-learn [38] y para el modelo de Word2Vec se uso Gensim [48]. A su vez, en el entrenamiento de los modelos se tomó en cuenta si se hacia o no una normalización previa de las características, escalándolas entre -1 y 1.

3.5 Técnicas de Aprendizaje Automático

Se emplearon los tres modelos descritos en el marco teórico, ANN, SVM y RF, por ser algoritmos clásicos que usan las características en la tarea de clasificación y permiten hallar superficies de separación para datos que no son lineales [41]. En cuanto a la composición de los modelos, las ANN tienen una sola capa oculta, y usa el 10 % de los datos de entrenamiento para la condición de parada. En los experimentos se exploraron el número de neuronas de la capa oculta, la función de activación de las neuronas en la capa oculta y la tasa de aprendizaje para el optimizador Adam. En las SVM en los experimentos se varió principalmente el parámetro de regularización C y el tipo de kernel empleado. Por último, para los modelos de RF se exploraron los hiperparámetros de el número de árboles aleatorios y el número mínimo de muestras necesarias para partir un nodo.

3.6 Métricas de Desempeño

Los experimentos fueron diseñados para explorar los parámetros en el proceso de extracción de características, y los hiperparámetros de los modelos usados. Para realizar una comparación entre los modelos, se tomó el promedio de cuatro métricas sobre los conjuntos de prueba de la validación cruzada. Las cuatro métrica empleadas fueron: 1) exactitud, 2) sensibilidad, 3) especificidad, y 4) *F1-score*, porque se emplean habitualmente en salud, por ejemplo para evaluar que tan buena es una prueba, como en el caso de las pruebas microbiológicas para detectar TB [5][28], y en el caso de *F1-score* aunque no es muy mencionada se usa como métrica que puede tener más cuenta el desbalance entre clases [38].

Más específicamente, la exactitud se define como la cantidad de muestras bien clasificadas sobre el total de muestras. Esta es una métrica estándar en el ML, sin embargo, para conjuntos con clases desbalanceadas no es la mejor métrica para evaluar el modelo. La sensibilidad es la capacidad del modelo para detectar verdaderos positivos, y la especificidad es la capacidad del modelo para detectar los verdaderos negativos. El *F1-score* es la media harmónica entre la sensibilidad y la precisión, en este caso se usa un *F1-score* ponderado respecto a las muestras de cada clase, para tener en cuenta el desbalance de clases que se menciono en la creación de

la base de datos, esto es posible debido a que se puede calcular la sensibilidad y la precisión de cada clase de forma independiente mediante las siguientes ecuaciones:

$$sensibilidad = \frac{verdaderos\,positivos}{verdaderos\,positivos + falsos\,negativos} \tag{3.1}$$

$$precisión = \frac{verdaderos\,positivos}{verdaderos\,positivos + falsos\,positivos} \tag{3.2}$$

A partir de la sensibilidad y la precisión para cada clase se calcula el *F1-score* mediante la ecuación 3.3, y se hace una media ponderada respecto a la cantidad de muestras de cada clase.

$$F1 - score = 2 \cdot \frac{precisión \cdot sensibilidad}{precisión + sensibilidad} \tag{3.3}$$

Capítulo 4

RESULTADOS

Uno de los primeros resultados del desarrollo del presente trabajo, tiene que ver con la obtención de una base de datos de texto de las HCE de pacientes con sospecha clínica de tener TB pulmonar. La base de datos se compone de un conjunto de archivos de texto que contienen los reportes de los médicos en un periodo de 30 días antes de la fecha de la primera baciloscopia o de la fecha del inicio del tratamiento, esto por sugerencia del grupo que conforma el proyecto en el que se enmarca el trabajo. En cuanto a la cantidad de pacientes en la base de datos con diagnóstico confirmado de TB pulmonar activa y pacientes con TB descartada, se tienen 116 y 35 pacientes respectivamente. Además, la cantidad de reportes tomados en esos 30 días y puede ir de uno a cinco reportes por paciente, se tomaron cinco reportes como el número máximo de reportes pensando en que no hubiera también un desbalance en el número de reportes de cada paciente.

El preprocesamiento fue aplicado a todos los reportes de todos los pacientes, y el texto limpiado fue guardado en un archivo *csv*, para que fuera más sencillo usar la base de datos. La extracción de características se realizó sobre los textos de la columna *full text* del archivo *csv* creado (ver figura 3.7), para extraer las características se tienen dos métodos, TF-IDF y *embeddings* con Word2Vec, los cuales permiten hacer una representación numérica del contenido del texto; luego, con cada método de extracción de características se implementaron tres modelos de ML: ANN, SVM y RF, modelos que han sido usados por su capacidad de generalización y fácil implementación [15][18][41]. Como resultados se muestran solo las mejores combinaciones de parámetros e hiperparámetros de la exploración realizada, esto en base a las métricas de exactitud, sensibilidad, especificidad y *F1-score*.

En la tabla 4.1 se muestran las mejores combinaciones de parámetros del método TF-IDF para extracción de características, con los hiperparámetros de los tres modelos implementados de ML. Dentro de la tabla 4.1 el número de términos hace referencia al tamaño del vocabulario para TF-IDF, el # Componentes significa el tamaño del vector para hacer la reducción de la dimensionalidad, y con normalizar se refiere a si se aplica una normalización a las características, escalándolas a valores entre -1 y 1. En las métricas de desempeño se encuentra el promedio de cada una de las métricas en los conjuntos de prueba de la validación cruzada

Además de los resultados de la tabla 4.1, se realizó una exploración de que términos tomar en base a un umbral máximo y otro mínimo del valor IDF del vocabulario completo. Luego de la exploración de estos umbrales solo hubo mejoría en el modelo de SVM con métricas de 0.695, 0.810, 0.310 y 0.693, en exactitud, sensibilidad, especificidad y *F1-score* respectivamente. Los hiperparámetros de la SVM para estos resultados son $C = 1$ y $kernel = rbf$, y se tomaron

Tabla 4.1: Mejores modelos usando TF-IDF para extraer características.

			Métricas de Desempeño			
Modelo de ML	Hiperparámetros	Parámetros de TF-IDF	Exactitud	Sensibilidad	Especificidad	F1-score
SVM	C=1.5 kernel= rbf	unigramas y bigramas 1000 términos # componentens=100 Sin normalizar	0.675	0.81	0.227	0.666
ANN	Función de activación = 'tanh', Neuronas en la capa oculta = 17, tasa de aprendizaje = 0.001	uni, bi y tri-gramas 1000 términos # Componentens=150 Con normalización	0.768	0.932	0.227	0.719
RF	Número min para partir nodo = 4 Número de arboles = 4	unigramas y bigramas 1000 términos # Componentens=150 Sin normalizar	**0.721**	**0.802**	**0.462**	**0.723**

unigramas y bigramas para el vocabulario del método TF-IDF.

Por otro lado, en la tabla 4.2 se muestran las mejores combinaciones de parámetros para extraer el *embedding* mediante Word2Vec, con los hiperparámetros de los tres modelos implementados de ML. En las métricas de desempeño se encuentra el promedio de cada una de las métricas en los conjuntos de prueba de la validación cruzada. Dentro de la tabla 4.2 solo hay resultados usando el modelo COBW con una ventana de 6 palabras para obtener el *embedding*, y aplicando una normalización de las características luego de la capa de agrupamiento, ya que esta forma de obtener los *embeddings*, fue la que mejor resultados obtuvo en todos los modelos. Además, en los parámetros del *embedding*, el tamaño se refiere al tamaño del vector del *embedding*, y la agrupación (o *pooling* en ingles) se refiere a si se toma el máximo o el promedio de la matriz de *embeddings* de cada documento.

Tabla 4.2: Mejores modelos usando embeddings para extraer las características.

			Métricas de Desempeño			
Modelo de ML	Hiperparámetros	Parámetros del embedding	Exactitud	Sensibilidad	Especificidad	F1-score
SVM	C = 100 kernel= Polinomial	Tamaño = 120 Agrupación = Promedio	0.714	0.811	0.399	0.717
ANN	Función de activación = 'tanh', Neuronas en la capa oculta = 5, tasa de aprendizaje = 0.001	Tamaño = 100 Agrupación = Máximo	0.688	0.818	0.257	0.680
RF	Número min para partir nodo = 4 Número de arboles = 9	Tamaño = 120 Agrupación = Máximo	0.716	0.829	0.343	0.695

Por otro lado, como se mencionó en el marco teórico, una de las ventajas de usar *embeddings* es la capacidad de interpretabilidad de esta forma de representación de los datos. Como cada palabra queda representada por un vector, mediante la distancia entre los puntos de los vectores se puede determinar que términos estás más relacionados entre ellos, a su vez, luego de obtener los términos más cercanos respecto a un término especifico, se puede hacer una visualización de que tan cercanos están los términos, aplicando una reducción de la dimensionalidad de los *embeddings* a un espacio bi o tri-dimensional, con el que es posible identificar esa cercanía entre conceptos. En las figuras 4.1 y 4.2 se visualiza el resultado de reducir a dos componentes o dimensiones, los *embeddings* que representan a los 10 términos más cercanos a las palabras tuberculosis y esputo respectivamente, para la reducción de la dimensionalidad se empleó la técnica de análisis de componentes principales del ingles *Principal Component Analysis* (PCA). Estas dos palabras fueron escogidas porque se usan mucho cuando se ha-

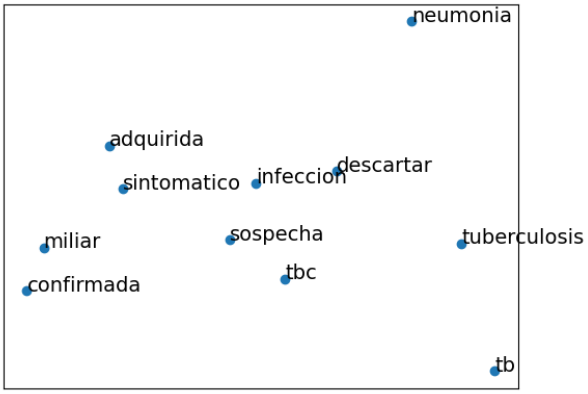


Figura 4.1: Espacio del *embedding* para palabras cercanas a tuberculosis reducido a 2 componentes usando PCA.

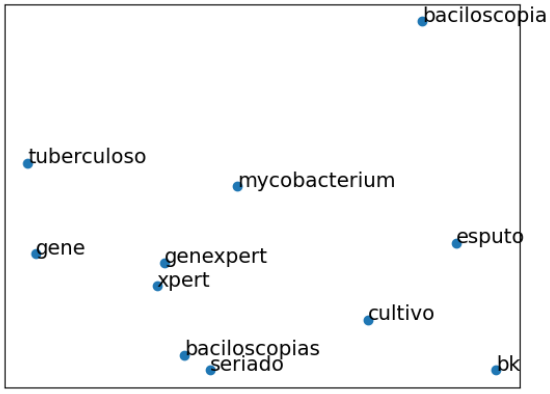


Figura 4.2: Espacio del *embedding* para palabras cercanas a esputo reducido a 2 componentes usando PCA.

bla de TB, y es posible observar como aparecen entre los términos cercanos, palabras muy relacionadas entre ellas y de las que se ha hablado en el desarrollo del trabajo.

En resumen, se realizó la creación de una base de datos mediante la extracción de texto de las HCE de pacientes sospechosos de TB pulmonar, y se implementaron modelos de NLP que tienen como objetivo hacer una predicción de TB pulmonar a partir del procesamiento del texto, y cuyos resultados están en las tablas 4.1 y 4.2. La implementación e interpretación de los resultados se hará en el siguiente capítulo, donde además profundizará en las características de los modelos implementados que mejores resultados obtuvieron.

Capítulo 5

DISCUSIÓN

La primera cuestión a discutir es la creación de la base de datos, para lo cual fue necesario el diálogo con los profesionales de la salud pertenecientes al proyecto, con el fin de conocer su punto de vista respecto a la información que la base de datos debía contener, ya que fueron ellos quienes realizaron el diagnóstico de TB pulmonar, teniendo en cuenta tanto las características clínicas de los pacientes como los resultados de las pruebas microbiológicas que se les realizaron. Debido a que el proyecto en el que se enmarca este trabajo espera generar herramientas computacionales que ayuden en el diagnóstico de TB en pacientes que ya son sospechosos de TB, la base de datos debía contener registros que no tuvieran información de la confirmación de las pruebas microbiológicas, por eso la construcción se hizo a partir de los folios anteriores a las fechas de inicio de tratamiento y de la realización de las pruebas, además, el texto extraído contiene los reportes médicos con el fin de que estos contengan la información que el médico considera relevante del paciente en esos días previos al diagnóstico de la TB.

En cuanto a los modelos implementados, para decidir cuales modelos son los que tienen mejor desempeño del amplio conjunto de modelos explorados, fue necesario decidir que métricas tenían mayor relevancia, en este caso al tener una base de datos tan desbalanceada a favor de los casos positivos de TB, se analizaron en conjunto las cuatro métricas. Las métricas de sensibilidad y especificidad fueron escogidas ya que son usadas cuando se habla del desempeño de las pruebas diagnósticas, como en las pruebas microbiológicas que son empleadas en el diagnóstico de la TB pulmonar, lo que permite tener un contraste de los resultados. Por otro lado, el *F1-score* aunque no es muy usado para referirse a las pruebas diagnósticas, tiene la ventaja de tomar de consideración el desbalance de clases, sin embargo, para este caso el desbalance entre las clases era muy marcado y para los casos en los que los modelos clasificaban casi todo como TB se obtenían valores cercanos al 65 %, valor que no es representativo de lo que pasa realmente con el modelo. Teniendo esto en cuenta, se le dio más importancia al resultado de la especificidad a la hora de seleccionar los mejores modelos, dado que la clase con menor número de muestras era la de pacientes con TB descartada, y además, es una métrica que suele ser usada para dar el desempeño de las pruebas diagnósticas; sin embargo, un incremento en la especificidad normalmente venía asociado con un decremento en las otras métricas, así que si las demás métricas disminuían su valor considerablemente al aumentar la especificidad, estos modelos también eran descartados.

En las tablas 4.1 y 4.2, se muestra que ninguno de los modelos considerados como los mejores, obtuvieron una especificidad por encima del 50 %, esto puede deberse a que en la

construcción de la base de datos se tienen pacientes que ya son sospechosos de TB, lo que hace que las HCE de pacientes con TB descartada, tengan características similares a pacientes con TB confirmada. Además, según los criterios para seleccionar los mejores modelos, aunque la especificidad fuera muy alta, por ejemplo de más del 60 %, las demás métricas bajaban a valores cercanos al 30 %, por lo que estos modelos se descartaban. En los resultados aunque estos no muestren una especificidad muy elevada, es necesario tener en cuenta el desbalance de las clases que juega un papel importante dentro de los métodos de ML, sumado a esto, solo se está usando la información de los reportes dentro de las HCE y no ninguna otra fuente de información como imágenes radiográficas o variables clínicas, esto se espera mejorar con la conjunción con otros modelos que si usen datos clínicos, epidemiológicos y/o sociodemográficos como los que se están explorando dentro del proyecto, y en un futuro crear una herramienta más robusta para ayudar en el diagnóstico de pacientes sospechosos de TB pulmonar.

En las tablas de los resultados también se visualiza que los mejores parámetros de los métodos de extracción de características son muy similares entre los modelos de ML, por ejemplo, en TF-IDF el mejor tamaño de vocabulario se mantuvo constante en 1000 términos, y se evidencia la necesidad de usar bigramas y trigramas para obtener buenos desempeños (ver tabla 4.1). Por su lado, en los *embeddings* como se mencionó en la sección de resultados, los mejores resultados se obtuvieron usando COBW con una ventana de 6 palabras, pero ninguno de estos superó lo obtenido con TF-IDF. Una de las posibles mejoras que se pueden implementar es usar en los *embeddings* bigramas y trigramas, ya que como se vio con TF-IDF en la tabla 4.1, ninguno de los mejores resultados usan solo unigramas como si es el caso de los *embeddings*. Para finalizar, respecto a las tablas4.1 y 4.2, se puede observar que los mejores resultados se obtuvieron usando el método TF-IDF para la extracción de características, y el modelo de RF para la clasificación.

Por otro lado, los resultados del uso de *embeddings* aunque no hayan sido mejores que el uso de TF-IDF, estos tienen la ventaja de su interpretabilidad a partir de ver las relaciones entre los vectores de las palabras como se muestra en las figuras 4.1 y 4.2. En estas figuras por ejemplo, es interesante observar como en ambas aparecen términos muy relacionados con tuberculosis y esputo que han sido mencionados durante el trabajo, para el caso de la figura 4.1 entre los términos más cercanos a tuberculosis, están abreviaciones que los médicos usan para referirse a la enfermedad como tbc y tb; en el caso del término esputo de la figura 4.2, se ve una alta relación con los métodos diagnósticos de la enfermedad, como es de esperarse ya que es con el esputo que se realizan las pruebas diagnósticas, además, en la figura 4.2 también se ve como los términos que hacen referencia a las pruebas, como baciloscopias, cultivo, y genexpert (de la prueba molecular) son muy cercanos entre ellos.

Para finalizar, se evidencia la posibilidad de usar la información contenida en el texto, para la creación de herramientas que puedan ayudar a los profesionales de la salud en el cuidado de los pacientes. Estas herramientas no solo deben estar orientadas al diagnóstico y ser una caja negra que prediga un valor, también pueden usarse para buscar información relevante o simplemente procesar el texto y encontrar relaciones entre conceptos como se mostró en las figuras 4.1 y 4.2; además, como se interpreten sus resultados depende del profesional que las use y se debe tener en cuenta que es necesaria más información del paciente si se quiere construir un DSS que sea capaz de dar una ayuda clara y oportuna a los profesionales de la salud.

Capítulo 6

CONCLUSIONES

En este trabajo se presenta la creación una base de datos a partir de las HCE de pacientes con sospecha de TB pulmonar, y la implementación de un sistema de NLP con el que se hace un preprocesamiento de los textos, luego, una extracción de características, y se usan en tres modelos de ML que buscan predecir TB pulmonar, todo esto con el fin de dar apoyo en el diagnostico de la enfermedad. Al final, el modelo que mejor desempeño mostró en los conjuntos de prueba de la base de datos creada, obtuvo 0.721, 0.802, 0.462, y 0.723 en las métricas exactitud, sensibilidad, especificidad y *F1-score* respectivamente, dicho modelo usa una reducción de la dimensionalidad de las características obtenidas con TF-IDF, e implementa el algoritmo de RF, para hacer la clasificación de los pacientes.

La base de datos creada tiene importancia dada la poca cantidad bases de datos del mismo estilo, en la búsqueda realizada de bases de datos similares, no se encontró ninguna base de datos de texto en español cuyo énfasis fuera en pacientes con TB, por lo que esta es una de las primeras bases de datos en el área que se usan pacientes con sospecha de TB. El contenido de la base de datos consiste en reportes clínicos que los médicos realizaron previo al diagnóstico de TB, los reportes fueron guardados sin ningún tipo de preprocesamiento en archivos de texto. Luego de tener la base de datos construida, se procedió implementar modelos de IA que buscan predecir TB pulmonar con confirmación microbiológica.

La metodología empleada para el manejo de los reportes clínicos consta de un preprocesamiento de los textos, una extracción de características y modelos de ML que predicen TB, estos modelos fueron pensados con de manera que tomaran la información relevante del texto y aprendieran los patrones que identifiquen cada clase. En el preprocesamiento se limpia el texto y se eliminan palabras que no aportan información (*stopwords*), esto para que los datos que entren a los modelos de ML no estén contaminados con información irrelevante para la tarea. Luego, en la extracción de características se usaron dos métodos, TF-IDF y *embeddings*, el primero tiene en cuenta la aparición de las palabras en todos los documentos con el fin de destacar palabras que sean más relevantes en términos generales dentro de las HCE; el segundo método busca que palabras que estén asociadas entre ellas sean representadas de una forma similar, y así ayudar al modelo a encontrar patrones que las asocien. Por último, los modelos de ML que se usaron fueron ANN, SVM y RF, por su capacidad de generalizar y de encontrar superficies de decisión para clases que no son linealmente separables, además su implementación es sencilla y no tienen de un alto coste computacional [41].

Tanto los parámetros de los métodos de extracción de características, como los hiperparámetros de los modelos de ML, fueron explorados usando validación cruzada y en todos ellos se

usaron los mismos conjuntos de datos para que los resultados fueran comparables entre ellos. Para evaluar el desempeño de los modelos se usaron cuatro métricas que son la exactitud, sensibilidad, especificidad y *F1-score* ponderada respecto al número de muestras en cada clase, y la selección de los mejores modelos se realizó a partir del análisis de esas cuatro métricas, dándole mayor prioridad a la especificidad porque esta métrica está asociada a que tan bueno es el modelo en la predicción de pacientes con TB pulmonar descartada, que es la clase con menor cantidad de muestras, sin embargo, para evitar una disminución de las demás métricas se descartaron también modelos en los cuales estas métricas disminuyeran notablemente su valor.

Capítulo 7

RECOMENDACIONES Y TRABAJOS FUTUROS

En cuanto a trabajos futuros se debe seguir profundizando en el uso de la base de datos, por ejemplo, el preprocesamiento de los textos se puede mejorar con el fin de eliminar más ruido y enfatizar conceptos que permitan un mejor desempeño de los modelos, también se pueden transformar palabras que sean cercanas entre ellas y convertirlas a un solo concepto, por ejemplo, en la figura 4.1 se observa el uso de abreviaciones de tuberculosis (tb y tbc) que pueden ser convertidas a una sola palabra (por ejemplo tuberculosis), y de esta manera ayudar al modelo a que se concentre en identificar otros patrones. En la extracción de características se podrían usar *embeddings* pre-entrenados o emplear otras formas de obtenerlos, estos *embeddings* pre-entrenados pueden ser desarrollados especialmente para textos médicos como en [49][50], y tienen la ventaja de que su contenido ya ira dirigido a la solución de problemas que tengan que ver con textos médicos.

Además, recientemente han habido grandes avances en los modelos de aprendizaje profundo para el NLP, que hacen uso de redes neuronales recurrentes y transformers [51][52], estos modelos son más complejos ya que usan una mayor cantidad de parámetros y grandes bases de datos para ser entrenados lo que mejora su desempeño. El tamaño de la base de datos creada no es muy grande y esta dirigida a TB, lo que impide en principio el uso de estos modelos de aprendizaje profundo; sin embargo, se podría encontrar la manera de hacer una transferencia de aprendizaje y usar también modelos pre-entrenados en esta base de datos, esperando obtener mejores resultados. Todas estas mejoras son fáciles de proponer luego de tener la base de datos construida, por eso también es importante su desarrollo dentro del proyecto, y mostrar que es otra fuente de información valida para la creación de las herramientas computacionales que ayuden en el diagnóstico y manejo de pacientes con TB. Además, el uso de otras variables clínicas, epidemiologías y demás, pueden integrase en un modelo computacional que pueda ser usado por los médicos en el diagnóstico de la TB.

Bibliografía

[1] (14 de oct. de 2020). «Tuberculosis,» who.int, (visitado 04-03-2021).

[2] «Global Tuberculosis Report,» World Health Organization, Geneva, 2020.

[3] D. M. C, G. R. H, C. V. S y S. M. C, «Factores asociados al diagnóstico tardío de pacientes con tuberculosis pulmonar en Lima Este, Perú,» *Revista Peruana de Medicina Experimental y Salud Pública*, vol. 21, n.º 1, págs. 18-22, 2004, Publisher: Instituto Nacional de Salud, ISSN: 1726-4642, 1726-4634. (visitado 10-05-2021).

[4] D. Palmero, G. C. de Casado, J. Castagnino, R. M. Musella, O. Aidar, M. Ambroggi, A. J. José, M. C. Brian, E. Canedo, M. Cufré, V. Curras, P. G. Montaner, S. Krugliansky, N. Leidi y E. Moraña, «Guías de diagnóstico, tratamiento y prevención de la tuberculosis, HOSPITAL MUÑIZ - INSTITUTO VACCAREZZA,» pág. 43, 2010.

[5] MINSALUD, *Resolución Número 0000227 de 2020*, 20 de feb. de 2020.

[6] «Boletin Epidemiológico Semanal, Semana 12,» Instituto Nacional de Salud, Colombia, 2020.

[7] MINSALUD, «Plan Estratégico: Hacia el fin de la Tuberculosis, Colombia 2016-2025,» Colombia, 2016.

[8] A. Sánchez, «Por cada 1.000 habitantes en Colombia, hay alrededor de 1,5 médicos generales,» 25 de mar. de 2020.

[9] A. A. Flores-Ibarra, M. D. Ochoa-Vázquez y G. A. S. Tec, «Estrategias diagnósticas aplicadas en la Clínica de Tuberculosis del Hospital General Centro Médico Nacional la Raza,» *Rev Med Inst Mex Seguro Soc.*, pág. 6, 2015.

[10] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak y K. I. Kroeker, «An overview of clinical decision support systems: Benefits, risks, and strategies for success,» *npj Digital Medicine*, vol. 3, n.º 1, pág. 17, dic. de 2020, ISSN: 2398-6352. DOI: 10.1038/s41746-020-0221-y. (visitado 07-05-2021).

[11] (15 de oct. de 2016). «Cardiovascular disease: Clinical decision-support systems (CDSS),» The Guide to Community Preventive Services (The Community Guide), (visitado 07-05-2021).

[12] E. H. Shortliffe y M. J. Sepúlveda, «Clinical Decision Support in the Era of Artificial Intelligence,» *JAMA*, vol. 320, n.º 21, págs. 2199-2200, 4 de dic. de 2018, ISSN: 0098-7484. DOI: 10.1001/jama.2018.17163. (visitado 15-04-2021).

[13] Amisha, P. Malik, M. Pathania y V. K. Rathaur, «Overview of artificial intelligence in medicine,» *Journal of family medicine and primary care*, vol. 8, n.º 7, págs. 2328-2331, jul. de 2019, Publisher: Wolters Kluwer - Medknow, ISSN: 2249-4863. DOI: 10.4103/jfmpc.jfmpc_440_19. PMID: 31463251.

[14] P. Dande y P. Samant, «Acquaintance to Artificial Neural Networks and use of artificial intelligence as a diagnostic tool for tuberculosis: A review,» *Tuberculosis*, vol. 108, págs. 1-9, 1 de ene. de 2018, ISSN: 1472-9792. DOI: 10.1016/j.tube.2017.09.006.

[15] A. D. Orjuela-Cañón, J. E. Camargo Mendoza, C. E. Awad García y E. P. Vergara Vela, «Tuberculosis diagnosis support analysis for precarious health information systems,» *Computer Methods and Programs in Biomedicine*, vol. 157, págs. 11-17, abr. de 2018, ISSN: 1872-7565. DOI: 10.1016/j.cmpb.2018.01.009.

[16] A. D. Orjuela-Cañón y J. de Seixas. (). «Fuzzy-ART neural networks for triage in pleural tuberculosis | IEEE Conference Publication | IEEE Xplore,» (visitado 15-04-2021).

[17] S.-F. Sung, K. Chen, D. P. Wu, L.-C. Hung, Y.-H. Su e Y.-H. Hu, «Applying natural language processing techniques to develop a task-specific EMR interface for timely stroke thrombolysis: A feasibility study,» *International Journal of Medical Informatics*, vol. 112, págs. 149-157, abr. de 2018, ISSN: 1872-8243. DOI: 10.1016/j.ijmedinf.2018.02.005.

[18] T. D. Imler, J. Morea, C. Kahi y T. F. Imperiale, «Natural language processing accurately categorizes findings from colonoscopy and pathology reports,» *Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association*, vol. 11, n.º 6, págs. 689-694, jun. de 2013, ISSN: 1542-7714. DOI: 10.1016/j.cgh.2012.11.035.

[19] Y. Wang, J. Luo, S. Hao, H. Xu, A. Y. Shin, B. Jin, R. Liu, X. Deng, L. Wang, L. Zheng, Y. Zhao, C. Zhu, Z. Hu, C. Fu, Y. Hao, Y. Zhao, Y. Jiang, D. Dai, D. S. Culver, S. T. Alfreds, R. Todd, F. Stearns, K. G. Sylvester, E. Widen y X. B. Ling, «NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records,» *International Journal of Medical Informatics*, vol. 84, n.º 12, págs. 1039-1047, dic. de 2015, ISSN: 1872-8243. DOI: 10.1016/j.ijmedinf.2015.06.007.

[20] T. Cai, A. A. Giannopoulos, S. Yu, T. Kelil, B. Ripley, K. K. Kumamaru, F. J. Rybicki y D. Mitsouras, «Natural Language Processing Technologies in Radiology Research and Clinical Applications,» *Radiographics: A Review Publication of the Radiological Society of North America, Inc*, vol. 36, n.º 1, págs. 176-191, feb. de 2016, ISSN: 1527-1323. DOI: 10.1148/rg.2016150080.

[21] S. Doan, M. Conway, T. M. Phuong y L. Ohno-Machado, «Natural language processing in biomedicine: a unified system architecture overview,» *Methods in Molecular Biology (Clifton, N.J.)*, vol. 1168, págs. 275-294, 2014, ISSN: 1940-6029. DOI: 10.1007/978-1-4939-0847-9_16.

[22] W. H. Organization. (2015). «WHO End TB Strategy: Global strategy and targets for tuberculosis prevention, care and control after 2015,» WHO. Publisher: World Health Organization, (visitado 09-05-2021).

[23] «Protocolo de vigilancia en salud pública: Tuberculosis,» Instituto Nacional de Salud, Colombia, 2020.

[24] (17 de abr. de 2016). «Signs and Symptoms,» cdc.gov, (visitado 04-03-2021).

[25] D. of Tuberculosis Elimination. (1 de jun. de 2016). «TB Risk Factors,» cdc.gov, (visitado 04-03-2021).

[26] (1 de jun. de 2016). «Deciding When to Treat Latent TB Infection,» cdc.gov, (visitado 04-03-2021).

[27] CDC, *Purebas de Detección de tuberculosis*, ago. de 2013. dirección: https://www.cdc.gov/tb/esp/pdf/Pruebas-de-detecci%5C%C3%5C%B3n-de-tuberculosis.pdf.

[28] M. L. Pérez del Molino, V. Tuñez Bastida, M. R. García Ramos y F. L. Lado Lado, «Diagnóstico microbiológico de la tuberculosis,» *Medicina Integral*, vol. 39, n.º 5, págs. 207-215, 1 de mar. de 2002, Publisher: Elsevier, ISSN: 0210-9433. (visitado 22-04-2021).

[29] «Guía para la vigilancia por laboratorio de tuberculosis,» Instituto Nacional de Salud, Colombia, 2020.

[30] PAHO, *Preguntas frecuentes sobre el método Xpert MTB/RIF*, 2011. dirección: https://www.paho.org/hq/dmdocuments/2011/Preguntas_frecuentes_Xper_MTB-RIF_final.pdf.

[31] (15 de abr. de 2020). «Tratamiento para la enfermedad de la TB | Tratamiento | TB | CDC,» (visitado 21-04-2021).

[32] «interoperabilidad de Datos de la Historia Clínica en Colombia: Términos y siglas,» Ministerio de Salud y Protección Social, Colombia, 2019.

[33] «Ley Número 2015 de 2020,» República de Colombia, 31 de ene. de 2020.

[34] G. G. Chowdhury, «Natural language processing,» *Annual Review of Information Science and Technology*, vol. 37, n.º 1, págs. 51-89, 2003, ISSN: 1550-8382. DOI: https://doi.org/10.1002/aris.1440370103. (visitado 28-04-2021).

[35] E. D. Liddy, «Natural language processing,» *In Encyclopedia of Library and Information Science*, pág. 15, 2001.

[36] K. Ganesan. (23 de feb. de 2019). «Text preprocessing for machine learning & NLP,» Kavita Ganesan, Ph.D, (visitado 28-04-2021).

[37] I. Alimova y E. Tutubalina, «Multiple features for clinical relation extraction: A machine learning approach,» *Journal of Biomedical Informatics*, vol. 103, pág. 103 382, 2020, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2020.103382.

[38] 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 y E. Duchesnay, «Scikit-learn: Machine Learning in Python,» *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[39] J. Brownlee. (10 de oct. de 2017). «What are word embeddings for text?» Machine Learning Mastery, (visitado 30-04-2021).

[40] T. Mikolov, K. Chen, G. Corrado y J. Dean, «Efficient Estimation of Word Representations in Vector Space,» *arXiv:1301.3781 [cs]*, 6 de sep. de 2013. arXiv: 1301.3781. (visitado 30-04-2021).

[41] S. Raschka, *Python Machine Learning: Unlock deeper insights into Machine Leaning with this vital guide to cutting-edge predictive analytics.* 23 de sep. de 2015, 456 págs.

[42] Chrislb, *English: Diagram of an artificial neuron.* 14 de jul. de 2005. (visitado 24-05-2021).

[43] U. User:Cyc based on PNG version by, *English: Graphic showing how a support vector machine would choose a separating hyperplane for two classes of points in 2D. H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximum margin.* 26 de nov. de 2012. (visitado 24-05-2021).

[44] Shehzadex, *Türkçe: dogrusal olarak ayrilamayan veriyi bir baska boyutta dogrusal ayrilabilir duruma getirme islemi*, 1 de nov. de 2016. (visitado 29-04-2021).

[45] Jeremybeauchamp, *English: A visual comparison between the complexity of decision trees and random forests.* 13 de dic. de 2020. (visitado 29-04-2021).

[46] *Pdfminer.six*, 2020. dirección: https://pypi.org/project/pdfminer.six/.

[47] E. Loper y S. Bird, «NLTK: The Natural Language Toolkit,» *CoRR*, vol. cs.CL/0205028, 2002.

[48] R. Řehůřek y P. Sojka, «Software Framework for Topic Modelling with Large Corpora,» English, en *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: ELRA, mayo de 2010, págs. 45-50.

[49] F. Soares, M. Villegas, A. Gonzalez-Agirre, M. Krallinger y J. Armengol-Estapé, «Medical Word Embeddings for Spanish: Development and Evaluation,» en *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, Minneapolis, Minnesota, USA: Association for Computational Linguistics, jun. de 2019, págs. 124-133. DOI: 10.18653/v1/W19-1916.

[50] A. Gutiérrez-Fandiño, J. Armengol-Estapé, C. P. Carrino, O. De Gibert, A. Gonzalez-Agirre y M. Villegas, «Spanish Biomedical and Clinical Language Embeddings,» *arXiv:2102.12843 [cs]*, 25 de feb. de 2021. arXiv: 2102.12843. (visitado 13-05-2021).

[51] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever y D. Amodei, «Language Models are Few-Shot Learners,» *arXiv:2005.14165 [cs]*, 22 de jul. de 2020. arXiv: 2005.14165. (visitado 04-05-2021).

[52] P. López Úbeda, M. C. Díaz-Galiano, L. A. Urena Lopez, M. Martin, T. Martín-Noguerol y A. Luna, «Transfer learning applied to text classification in Spanish radiological reports,» en *Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020)*, Marseille, France: European Language Resources Association, mayo de 2020, págs. 29-32, ISBN: 979-10-95546-65-8. (visitado 13-05-2021).