Sistema de predicción de diagnósticos médicos en base a las notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático.

Cruz Caiza, Macarena Lizbeth y Quishpe Rocha, Luis Lenin

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Software.

Trabajo de integración curricular, previo a la obtención del título de Ingeniero de Software

Ing. Uyaguari Uyaguari, Alvaro Danilo Msc.

22 de febrero del 2023

Latacunga

CERTIFICADO DE ANÁLISIS
magister

Tesina_MIC_Quishpe_Cruz_14-02-2023

6% Similitudes

< 1% Texto entre comillas
0% similitudes entre comillas
< 1% Idioma no reconocido

Nombre del documento: Tesina_MIC_Quishpe_Cruz_14-02-2023.docx
ID del documento: 2a9548fef78c680f8f9aa70268f9db3769f6046c
Tamaño del documento original: 1,35 Mo

Depositante: JOSÉ LUIS CARRILLO
Fecha de depósito: 17/2/2023
Tipo de carga: interface
fecha de fin de análisis: 17/2/2023

Número de palabras: 14.133
Número de caracteres: 92.689

Ubicación de las similitudes en el documento:

Fuentes principales detectadas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	Tesina_Titulación_Pallo_Salazar_08-02-2023.docx | Tesina_Titulación_Pallo_S... #6dcccc El documento proviene de mi biblioteca de referencias	1%		Palabras idénticas : 1% (165 palabras)
2	www.apd.es | ¿Cuáles son los tipos de algoritmos del machine learning? | APD https://www.apd.es/algoritmos-del-machine-learning/ 1 fuente similar	1%		Palabras idénticas : 1% (161 palabras)
3	repositorio.utc.edu.ec | Diseño de un algoritmo utilizando Machine Learning para la... http://repositorio.utc.edu.ec/bitstream/27000/8014/3/MUTC-001017.pdf.txt	< 1%		Palabras idénticas : < 1% (128 palabras)
4	Tesis_Castillo_Chuquitarco-Antiplagio-Compilation_15-02-2023_.docx | Tesi... #2a158a El documento proviene de mi biblioteca de referencias	< 1%		Palabras idénticas : < 1% (112 palabras)
5	www.elsevier.es | Modelos predictivos en salud basados en aprendizaje de maquin... https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-articulo-modelos-predictivos-... 1 fuente similar	< 1%		Palabras idénticas : < 1% (103 palabras)

Fuentes con similitudes fortuitas

Nº	Descripciones	Similitudes	Ubicaciones	Datos adicionales
1	hdl.handle.net | Procesamiento del Lenguaje Natural. N. 67 (2021) http://hdl.handle.net/10045/117498	< 1%		Palabras idénticas : < 1% (38 palabras)
2	Documento de otro usuario #ab3ae3 El documento proviene de otro grupo	< 1%		Palabras idénticas : < 1% (36 palabras)
3	www.scielo.org.co http://www.scielo.org.co/pdf/rci/n102/0121-5612-rci-102-41.pdf	< 1%		Palabras idénticas : < 1% (35 palabras)
4	Documento de otro usuario #29679f El documento proviene de otro grupo	< 1%		Palabras idénticas : < 1% (12 palabras)
5	es.wikipedia.org | Reconocimiento de entidades nombradas - Wikipedia, la enciclop... https://es.wikipedia.org/wiki/Reconocimiento_de_entidades_nombradas	< 1%		Palabras idénticas : < 1% (20 palabras)

Fuentes mencionadas (sin similitudes detectadas)

Estas fuentes han sido citadas en el documento sin encontrar similitudes.

1 https://repositorio.upeu.edu.pe/handle/20.500.12840/2511
2 http://repobib.ubiobio.cl/jspui/handle/123456789/1772
3 http://e-spacio.uned.es/fez/view/tesisuned:ED-Pg-SisInt-Hfabregat
4 https://doi.org/10.1109/UKRCON.2017.8100379
5 https://riunet.upv.es/handle/10251/133840

Ing. Uyaguari Uyaguari, Álvaro Danilo Msc
C.C.: 0103411112

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Software

Certificación

Certifico que el trabajo de integración curricular: **"Sistema de predicción de diagnósticos médicos en base a las notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático."** fue realizado por los señores **Cruz Caiza Macarena Lizbeth** y **Quispe Rocha Luis Lenin,** el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizada en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Latacunga, 22 de febrero de 2023

Firma:

Ing. Uyaguari Uyaguari, Álvaro Danilo, Msc

C. C.: 0103411112

ESPE
UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Software

Responsabilidad de Autoría

Nosotros, **Cruz Caiza Macarena Lizbeth** con cédula de ciudadanía N° 1754422150 y **Quispe Rocha Luis Lenin** con cédula de ciudadanía N° 0504609090, declaramos que el contenido, ideas y criterios del trabajo de integración curricular**: Título: "Sistema de predicción de diagnósticos médicos en base a las notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático."** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Latacunga, 22 de febrero de 2023

Firma

Cruz Caiza, Macarena Lizbeth
C.C.: 1754422150

Firma

Quispe Rocha, Luis Lenin
C.C.: 0504609090

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Software

Autorización de Publicación

Nosotros, **Cruz Caiza Macarena Lizbeth** con cédula de ciudadanía N° 1754422150 y **Quispe Rocha Luis Lenin** con cédula de ciudadanía N° 0504609090, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de integración curricular: **Título: Sistema de predicción de diagnósticos médicos en base a las notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático."** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Latacunga, 22 de febrero de 2023

Firma

Cruz Caiza, Macarena Lizbeth
C.C.: 1754422150

Firma

Quispe Rocha, Luis Lenin
C.C.: 0504609090

Dedicatoria

El resultado de este trabajo está dedicado principalmente a Dios, a mis padres por todo su amor, trabajo y sacrificio en todos estos años, gracias por su apoyo incondicional y porque siempre estuvieron ahí en los buenos y malos momentos. Gracias por enseñarme a enfrentar cada adversidad. También me gustaría dedicar este trabajo a mi mami Nancy gracias por enseñarme los valores de compromiso y constancia, todo esto lo hace con amor y sin esperar nada a cambio. De igual forma me gustaría dedicarle este trabajo a mi hermano, gracias por siempre estar presente a lo largo de esta etapa de mi vida.

Macarena Lizbeth Cruz Caiza

Agradecimiento

Primero quiero dar gracias a Dios por darme la fuerza y sabiduría necesaria para terminar mi carrera, es un orgullo para mi mencionar a todas las personas que de una u otra forma me han ayudado a lograr esta meta tan anhelada de graduarme, gracias a mis padres Carlos y Fanny a mi hermano Rafael por siempre estar presentes en cada momento ya que son mi principal apoyo, gracias a mi segunda mamá, Nancy por creer en mí, ayudarme y guiarme en este proceso, de igual manera quiero agradecer a mis tíos Jorge y Elvia por cada consejo y por cada una de sus palabras que me guiaron durante este largo camino, gracias a mi abuelita Hermelinda por sus bendiciones, paciencia y amor recibido, de la misma manera quiero agradecer a la familia Quinteros Cruz en especial a mi tía Angelita por abrirme las puertas de su casa y apoyarme en este largo proceso académico, igualmente quiero agradecer de manera especial a mi compañero y amigo Lenin ya que juntos llevamos este trabajo adelante, asimismo quiero agradecer a mi mejor amiga Vane gracias por su apoyo incondicional, por cada palabra de aliento y por siempre estar para mí, finalmente quiero agradecer a mi tutor de tesis Msc. Álvaro Uyaguari gracias por su apoyo incondicional, por su paciencia y enseñanzas brindadas, gracias a todos ustedes hoy veo este nuevo triunfo realizado, gracias por apoyarme y creer en mí.

Macarena Lizbeth Cruz Caiza

Dedicatoria

Le dedico el resultado de este proyecto a Dios y mi familia. Principalmente a Dios que me ha dado la sabiduría y las fuerzas necesarias para lograr esta meta, por permitir que mi mente sea capaz de lograr este gran sueño, agradezco a mis padres y mis hermanos que estuvieron en los momentos difíciles y me apoyaron en todo lo que fue posible, me ayudaron brindarme los recursos necesarios, este proyecto de titulación es resultado del gran sacrificio que mi familia realizó.

Dedico este esfuerzo a toda mi familia en la Iglesia, por sus oraciones y la confianza que ha depositado en mí, por creer grandemente en mis habilidades y darme una palabra de aliento cuando lo he necesitado y mostrarme una sonrisa cuando tengo el ánimo bajo.

Dedicar este trabajo a mi propia persona ya que he recorrido un largo camino que no ha sido fácil, pero se ha logrado concluir una etapa de mi vida y abrir otra nueva como profesional, quiero seguir sueños y retos.

Luis Lenin Quispe Rocha

Agradecimiento

Agradezco a Dios y a mi familia, ellos han sido los que me soportaron en momentos de frustración, tristeza y me apoyaron dándome palabras de ánimos. Mis padres me han enseñado el valor del esfuerzo, trabajo y sacrificio, a mis hermanos y principalmente a mi hermano mayor Franklin que son su apoyo incondicional me da fuerzas para seguir esforzándome.

Agradezco de una manera muy especial a mi amiga Macarena que con ella hemos pasado varios años de la carrera juntos y me ha mostrado una amistad sincera, por estar ahí cuando he tenido momentos difíciles, se convirtió en parte de mi familia y la estimo mucho por ser mi compañera de locuras, risas, llantos y el proyecto de titulación, espero que todo en la vida le vaya bien y siga cumpliendo sueños.

También, quiero agradecer todo este esfuerzo a mi grupo de amigos “los reales” que conocí en la universidad que son personas que día a día compartir grandes experiencias y llegaron a ser parte de mi vida universitaria, mis amigos cercanos Diego, Kevin, Carlos, Maycol que a pesar de no pasar mucho tiempo junto a ellos se encuentran ahí para apoyarme.

Doy un agradecimiento muy especial a mis docentes de la universidad y a mi tutor de tesis Ing. Álvaro Uyaguari por sus enseñanzas y la paciencia para lograr este mérito, que siga compartiendo su conocimiento con paciencia y profesionalismo.

Luis Lenin Quispe Rocha

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Resumen

En el presente trabajo se desarrolló un sistema de predicción de diagnósticos médicos en base a notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático, con el objetivo de precisar la toma de decisiones mediante técnicas de análisis de datos, donde se utilizó una metodología de desarrollo de software ágil Scrum. El alcance del proyecto es llegar a construir un modelo y un sistema de predicción de diagnósticos que se ajuste a los pocos recursos lingüísticos médicos existentes en idioma español, pero que logre un desempeño competitivo o superior a los publicados en el estado del arte, por lo cual se aplicó buenas prácticas en el ciclo de desarrollo y se implementó el sistema con frameworks y arquitecturas actuales. Para el proyecto se usó un algoritmo que convierte el texto de las notas médicas en un archivo xml que es el primer paso para la construcción del sistema, los datos utilizados y convertidos al formato correspondiente se encuentran dentro de la página de huggingface como un conjunto de datos públicos. Para el entrenamiento del modelo se utilizó una librería de Python llamada sklearn, para colocar las etiquetas a las entidades médicas. También se utilizó paperspace el cual es un entorno que nos otorga gpu y cpu el cual fue utilizado para la implementación del dataset.

Palabras clave: Aprendizaje automático, análisis de datos, predicción de diagnósticos médicos.

Abstract

In the present work, a medical diagnosis prediction system was developed based on patients' clinical notes, applying machine learning techniques and models, with the objective of refining decision making through data analysis techniques, where an agile software development methodology Scrum was used. The scope of the project is to build a model and a diagnostic prediction system that fits the few existing medical linguistic resources in Spanish, but that achieves a competitive or superior performance to those published in the state of the art, so good practices were applied in the development cycle and the system was implemented with current frameworks and architectures. For the project an algorithm was used that converts the text of the medical notes into an xml file which is the first step for the construction of the system, the data used and converted to the corresponding format are found within the huggingface page as a public dataset. For the training of the model a Python library called sklearn was used to place the labels to the medical entities. We also used paperspace which is an environment that provides us with gpu and cpu which was used for the implementation of the dataset.

Key words: Machine learning, data analysis, medical diagnosis prediction.

Capítulo I

Introducción

Propósito y contextualización del tema

En la última década la inteligencia artificial se ha convertido en un campo líder de las tareas de procesamiento y generación de información a través de la aparición del aprendizaje automático o Machine Learning, que ha llevado su aplicabilidad a distintos campos como la robótica, procesamiento de voz, visión artificial, procesamiento natural del lenguaje, pretendiendo en cierta medida dotar a sistemas computacionales de la capacidad de aprender (Sancho Escrivá et al., 2020).

Los cambios tecnológicos han permitido que el lenguaje natural (LN), en el que hablan los seres humanos cotidianamente se pueda incorporar en los computadores para la realización de diversas tareas que ayudan al ser humano a una mejor redacción y comprensión de un texto extenso, traducción y otras. Sin embargo, lo que se estudia en el presente documento es la potencial aplicación de la arquitectura Transformer en el campo del procesamiento de lenguaje natural (Sancho Escrivá et al., 2020).

El aprendizaje automático (machine learning), comúnmente abreviado como ML, es un tipo de inteligencia artificial (IA) y se define como el campo que otorga a las computadoras la habilidad de aprender sin haber sido eficientemente programadas, (Morales & Jorge, 2017). En lugar de seguir reglas estáticas codificadas en un programa, esta tecnología simula procesos que realizan los humanos al momento de realizar una tarea, (Alfaro & Ospina, 2021). Para poder conseguir esa meta es necesario entrenar un modelo con el objetivo de que aprenda ese comportamiento o procesamiento (Guardiola González, 2020).

Existen tres diferentes grupos de algoritmos de machine learning, estos son: el aprendizaje supervisado, en el cual la máquina aprende con el ejemplo. De este modo, el operador proporciona al algoritmo de aprendizaje automático un conjunto de datos conocidos que incluye las entradas y salidas

deseadas, y el algoritmo debe encontrar un método para determinar cómo llegar a esas entradas y salidas.

En el aprendizaje sin supervisión el algoritmo estudia los datos para identificar patrones. No hay una clave de respuesta o un operador humano para proporcionar instrucción. En cambio, la máquina determina las correlaciones y las relaciones mediante el análisis de los datos disponibles. De acuerdo con esto el algoritmo organiza los datos de alguna manera para describir su estructura, llevando a la necesidad de agrupar los datos en grupos de manera que se vea de forma organizada y eficiente al momento de trabajar con la información.

La gran cantidad de datos que se maneja en el ámbito médico es de vital importancia ya que los registros clínicos son electrónicos y estos mismo pueden ser usados para las técnicas de aprendizaje automático, desarrollando así un análisis donde se pueda predecir y también reconocer entidades médicas por medio de modelos de cómputo, con la ayuda de estos datos se podrá obtener información y llevarlo a la práctica clínica con mayor precisión (Pineda, 2022).

Al definir las reglas, el algoritmo de aprendizaje automático intenta explorar diferentes opciones y posibilidades, monitorizando y evaluando cada resultado para determinar cuál es el óptimo (Pineda, 2022). En consecuencia, el sistema enseña a la máquina a través del proceso de ensayo y error. Aprende de experiencias pasadas y comienza a adaptar su enfoque en respuesta a la situación para lograr el mejor resultado posible.

En la presente investigación se realiza un análisis de lecturas sobre modelos para la predicción de diagnósticos basado en notas médicas textuales, juntamente con trabajos sobre métodos y técnicas de Machine Learning para la predicción de diagnósticos.

El propósito es desarrollar un sistema y nuevos métodos para la identificación de entidades biomédicas en español utilizando técnicas de procesamiento de lenguaje natural. Estos nuevos métodos y algoritmos serán integrados a un sistema para la predicción de diagnósticos, aplicando herramientas,

métodos y buenas prácticas de ingeniería de software. Sin embargo, existen pocos recursos lingüísticos en el dominio médico para identificar dichos rasgos en idioma español, limitando así la creación de nuevos modelos de aprendizaje automático dentro del contexto médico.

Justificación del interés de la investigación.

El lenguaje natural es enriquecido por su vocabulario y construcción, así se establecen características como la flexibilidad, ambigüedades e indeterminación permitiendo la variedad en la interpretación dependiendo la situación, el procesamiento natural del lenguaje (PNL) es el campo de estudio que busca entender cómo funciona el lenguaje, su construcción, la generación de nuevo lenguaje, así como todas las tareas que tienen relación con el tratamiento del lenguaje. Entre estas tareas se tiene la generación de un nuevo texto, traducciones de un idioma a otro, preguntas y respuestas, generar resumen, chatbots entre otros.

Los modelos de predicción clínica se aplican comúnmente en la práctica médica para ayudar a los profesionales de la salud a determinar el diagnóstico o pronóstico de un paciente. Existen varios enfoques y métodos para realizar este proceso de predicción, dependiendo de la naturaleza de los datos y de los recursos disponibles en el idioma en el que se redacten las notas médicas. Nuestro alcance es llegar a construir un modelo y un sistema de predicción de diagnósticos que se ajuste a los pocos recursos lingüísticos médicos existentes en idioma español, pero que logre un desempeño competitivo o superior a los publicados en el estado del arte.

Además, con la ayuda de los modelos predictivos son un grupo de técnicas que mediante los campos del aprendizaje automático la recolección de datos y el reconocimiento de patrones, pretende dar una predicción de resultados futuros, con el objetivo de precisar la toma de decisiones mediante técnicas de análisis de datos, la gran cantidad de datos que están almacenando en registros clínicos electrónicos y notas médicas y mayor poder computacional, hacen que las técnicas de aprendizaje de

máquina tenga un rol preponderante en el desarrollo de nuevos análisis predictivos y reconocimiento de patrones no conocidos con estos modelos de cómputo (Pineda, 2022).

Objetivo general

Desarrollar un sistema de predicción de diagnósticos médicos en lenguaje español.

Objetivos específicos

- Explorar modelos y corpus para el proceso de predicción de diagnósticos médicos.
- Desarrollar un modelo de predicción de diagnósticos en base a textos médicos en lenguaje español.
- Aplicar buenas prácticas en el ciclo de desarrollo e implementar el sistema con frameworks y arquitecturas actuales.
- Redacción de la tesina.

Metodología

El presente proyecto tiene como meta desarrollar un sistema de predicción de diagnósticos médicos en base a las notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático, de esta manera, cumplir con los objetivos planteados. La metodología que seguirá este proyecto consta de 3 fases:

En la primera fase se analiza la literatura científica relacionada con el objetivo de estudio esto con el fin de formular el marco teórico de este documento, para realizar esta fase se usará varios métodos teóricos como el método histórico-lógico y el método análisis-síntesis. Se analiza las características de los modelos y también de las bibliotecas que se ajustan de manera rápida y eficiente para poder acercarse al rendimiento más avanzado en la clasificación de texto.

En la segunda fase se analiza sobre la tokenización y el entrenamiento de modelos que ayuden al manejo de la información encontrada en el corpus biomédico-clínico en español. En el modelo BERT o RoBERTa el proceso a seguir consiste en un entrenamiento mediante el uso del lenguaje enmascarado a

nivel de subpalabras, además se revisa sobre la biblioteca huggingface Pytorch para ajustar de manera rápida y eficiente un modelo para acercarse a un rendimiento óptimo en la clasificación de textos médicos.

BERT (Representaciones de codificador bidireccional de transformadores), es el modelo que brinda una mejor comprensión y guía práctica para usar modelos de aprendizaje de transferencia en NLP. BERT es una arquitectura basada en redes Transformer para entrenar modelos de lenguaje el cual ayuda a la compresión de consultas adaptándose mucho mejor a los intereses del usuario.

En la última fase lo que se pretende es evaluar el sistema desarrollado y su eficiencia, para lo cual se empezará los métodos experimental y empírico, esto para poder verificar los resultados alcanzados una vez que el sistema de predicción se ha completado.

Capítulo II

Marco teórico

En este capítulo se describen los conceptos, características y elementos que pertenecen a los métodos y técnicas de predicción, así como algoritmos secuenciales que usa el machine learning ya que este campo es especializado en el reconocimiento de patrones complejos en un conjunto de datos. El aprendizaje automático tiene la característica de que es capaz de crear algoritmos que permiten a las computadoras aprender a realizar tareas a partir de datos, esto es lo que posibilita que el programa aprenda y mejore la ejecución en la tarea que ha sido asignada.

Las disciplinas de inteligencia artificial ayudan a llevar a cabo la clasificación de texto, junto con un grupo de algoritmos con mayor nivel de abstracción y menor supervisión de parte de personas y aun así poder cumplir con las tareas inteligentes (Pérez Ortiz de Landaluce, 2021).

Minería de datos

La minería de datos se puede definir inicialmente como el proceso de examinar grandes cantidades de información para descubrir nuevas e importantes relaciones, patrones y tendencias. La disponibilidad de grandes cantidades de datos y el uso generalizado de herramientas informáticas han transformado el análisis de información, dando lugar a la minería de datos o Data Mining.

Hoy en día, la minería de datos se utiliza en varios campos de la ciencia. Aplicaciones financieras y bancarias, análisis de mercado y comercio, seguros y salud privada, educación, procesos industriales, medicina, biología y biotecnología, telecomunicaciones y muchas otras áreas (Riquelme Santos et al., 2006).

Las tareas en la fase de minería de datos pueden ser descriptivas, es decir, (modelos de descubrimiento, o describir relaciones en los datos), predictivas (es decir, basado en lo previamente disponible). En otras palabras, es un campo interdisciplinario cuyo objetivo general es predecir resultados y descubrir relaciones en los datos. Para esto se utilizan herramientas automáticas, las cuales

emplean algoritmos sofisticados para encontrar principalmente patrones ocultos, asociaciones, anomalías, y/o estructuras de la gran cantidad de datos almacenados en la data warehouses u otros repositorios de información (Pérez López & Santin González, 2007).

Minería de texto.

La minería de textos, o text mining, es la aplicación de la lingüística computacional y el procesamiento de textos para facilitar la identificación y extracción de nuevos conocimientos a partir de colecciones de documentos textuales.

Existe una clara relación entre la minería de textos y la minería de datos la cual es la recuperación de datos y lingüística computacional (Valero Moreno, 2017).

La minería de texto es un paso adicional a la minería de datos. Este último está destinado a colaborar y comprender el contenido de la base de datos. Para la minería de datos, la materia prima son los datos a la que los usuarios le dan sentido, convirtiéndola en información que los expertos utilizarán para convertirla en conocimiento. La minería de datos tiene muchas aplicaciones, ya que se puede utilizar en casi todos los aspectos de la actividad humana (Valero Moreno, 2017).

Procesamiento de lenguaje natural.

Una de las labores fundamentales de la inteligencia artificial (IA) es la manipulación del lenguaje natural mediante herramientas computacionales, y en este sentido los lenguajes de programación juegan un papel importante ya que proporcionan el vínculo necesario entre el lenguaje natural y su manipulación por parte de una máquina.

PLN consiste en utilizar el lenguaje natural para comunicarse con una computadora, la computadora tiene que entender las oraciones provistas, estos lenguajes naturales facilitan el desarrollo de programas que realizan tareas relacionadas con el lenguaje y el desarrollo de modelos que nos ayudan a comprender los mecanismos humanos asociados al lenguaje (Vásquez et al., 2009).

Procesamiento de lenguaje natural aplicado en el campo médico.

Los principales síntomas del paciente, diagnósticos, medicamentos, tratamiento y los resultados de este se recopilan en documentos clínicos. La información contenida en estos registros se puede clasificar en estructurada y no estructurada. En el primer caso, tratamos con datos que pueden ser categóricos o numéricos. Por otro lado, se dice que los datos no están estructurados cuando no se pueden tabular en un formato de datos, como radiografías o datos médicos que carecen de un lenguaje verificable (Villena & Dunstan, 2019). En la práctica clínica, el texto libre no estructurado forma una parte importante de los datos de los pacientes. Estos incluyen, por ejemplo, historias clínicas de pacientes, protocolos de exámenes o notas diarias, y es muy importante poder extraer información de estas fuentes y utilizarlas para mejorar la atención clínica y la investigación. Hoy en día, los registros electrónicos de pacientes contienen cada vez más datos, y la información no estructurada a menudo se deja fuera de los proyectos de TI.

El propósito de codificar texto clínico es estructurar la información de tal manera que pueda ser fácilmente utilizada para tareas de gestión, clasificación de enfermedades, estadísticas o toma de decisiones. Ejemplos de codificación son la Clasificación Internacional de Enfermedades (CIE) y la Nomenclatura Sistematizada de Términos Médicos Clínicos (SNOMED-CT). Sin embargo, este es un proceso intensivo en capacitación que requiere tiempo tanto del personal responsable del tratamiento como del que se dedica exclusivamente a la codificación (Villena & Dunstan, 2019).

Reconocimiento de entidades nombradas

El reconocimiento de entidades nombradas (NER), también conocido como minería de entidades, es una tarea de minería de información que tiene como objetivo encontrar y clasificar las entidades nombradas que se encuentran en el texto en categorías predefinidas, como expresiones de persona, organización, lugar, tiempo y cantidad.

Las entidades nombradas a menudo se dividen conceptualmente en dos problemas distintos: detección de nombres, y clasificación de los nombres según el tipo de entidad al que hacen referencia. Es por eso que muchas veces en la literatura se lo conoce como reconocimiento y clasificación de entidades nombradas (NERC por sus siglas en inglés) (Haag, 2019). Las tareas NERC son la base para aplicaciones inteligentes como la generación de texto y gráficos de conocimiento debido a su importancia en el análisis semántico. Desde modelos basados en reglas que utilizan patrones de expresiones regulares, listas o diccionarios geográficos, hasta modelos supervisados y semisupervisados, como Hidden Markov Models Evolved (HMM), Support Vector Machine (SVM) y Conditional Random Field (CRF) modelo de aprendizaje automático supervisado) (Samy, 2021).

El papel de NERC es esencial para el desarrollo de un sistema legal inteligente. Dada la gran cantidad de texto que normalmente se procesa en este campo, existe un creciente interés en el trabajo de los procesadores de texto legales en general y NERC en particular.

Este interés está impulsado por el gran potencial de la tecnología PLN y su capacidad para brindar soluciones inteligentes que benefician a los usuarios clave del sector, como abogados, jueces, redactores legales, médicos y el sector público en general. No se trata de documentos puramente jurídicos, sino de documentos administrativos con un alto contenido jurídico, como en el caso de las compras y contratos públicos (Samy, 2021).

Reconocimiento de entidades nombradas biomédicas

En el reconocimiento de entidades biomédicas entra en relación la ingeniería biomédica que es un campo interdisciplinario donde convergen disciplinas como la ingeniería eléctrica, mecánica, medicina, biología, física, entre otros. Su objetivo es aplicar tecnología de punta para desarrollar dispositivos y métodos médicos que contribuyan al bienestar humano y mejoren nuestra comprensión de los procesos biológicos que ocurren en los humanos (Sarmiento-Ramos, 2020).

La aplicación de técnicas NER en el dominio biomédico es un campo de estudio con gran relevancia, y para poder clasificar los datos como notas clínicas o registros las técnicas NER brindan enfoques supervisados y no supervisados. Si bien las fuentes de información están garantizadas gracias a la era del internet, la generalización de colecciones documentales anotadas sigue siendo un proceso que requiere mucho tiempo (Fabregat, 2021).

El problema en el reconocimiento de entidades nombradas es verse como una tarea de clasificación secuencial, en la que un sistema NER debe asignar una etiqueta a cada palabra que pertenece a una entidad, según el tipo y la posición de esa palabra dentro de la entidad. Para evaluar la complejidad de esta tarea en diferentes niveles y para cubrir diferentes tipos de entidades y contextos, se debe usar un esquema de anotación como el IOB (Inside-Outside-Begin).

Redes neuronales

Las redes neuronales se describen como el conjunto más popular de algoritmos de aprendizaje automático, y ahora aumentan con las mejoras tecnológicas para admitir la alta intensidad de las operaciones aritméticas involucradas. Esta mejora se debe en gran parte al uso de GPUs, o unidades de procesamiento gráfico, que son plataformas encargadas de optimizar el procesamiento paralelo, lo que se traduce en una mayor simplicidad, velocidad y menores costos (Repetur, 2019).

Las redes neuronales resultan eficaces cuando se entrenan con datos estructurados. Por ejemplo, una base de datos de precios de viviendas basada en propiedades, la relación entre el fenotipo metabólico y los niveles de insulina en plasma, etc. Sin embargo, tienen problemas con los datos no estructurados, como imágenes médicas, pistas de audio y archivos electrónicos de pacientes, (Arias et al., 2019). Al igual que la escritura, la compresión de voz por máquina (en cualquier idioma) es un problema difícil de resolver. Esto se debe a que pueden tener variaciones casi infinitas (acento, entonación, velocidad, etc.) al pronunciar palabras. Otro desafío para resolver son las largas secuencias de entrada generadas en tiempo real. Con imágenes, es relativamente fácil usar una red neuronal para

restar e identificar caracteres en una imagen, pero las operaciones realizadas en archivos de audio y diálogo no son tan sencillas, ya que primero separan los caracteres.

Las redes neuronales artificiales y el aprendizaje profundo son dos de las herramientas de aprendizaje automático más poderosas para crear sistemas que aprenden automáticamente de conjuntos de datos, reconocen patrones, predicen comportamientos y generalizan información. Estas dos herramientas son áreas potenciales de investigación con aplicaciones de ingeniería (Sarmiento-Ramos, 2020).

Redes recurrentes.

Las Redes Recurrentes constituyen un grupo de modelos adecuados para el procesamiento de datos estructurados secuenciales. Son eficientes porque tienen un espacio oculto de alta dimensión con dinámica no lineal que les permite recordar y procesar información pasada (Pérez Guerrero, 2020). Las redes recurrentes son modelos para resolver problemas de reconocimiento de patrones y se basan básicamente en los mismos conceptos que los perceptrones multicapa de retroalimentación (MLP).

La gran variedad de redes recurrentes hace que las topologías y sus mecanismos de aprendizaje sean versátiles. Hacer una plataforma que le permita construir redes recurrentes requiere mucha flexibilidad (Cruz et al., 2007). Existen varios simuladores (SNNS, NeuroSolution, etc.) y lenguajes de especificación (Aspirin, CONNECT, etc.) con el propósito de tratar con redes recurrentes, pero todos llegan a presentar limitaciones en la flexibilidad dificultando el desarrollo sistemático de soluciones a problemas reales (Cruz et al., 2007).

Redes Transformers.

Las redes neuronales transformativas son una nueva clase de redes neuronales secuenciales autoconscientes que han demostrado adaptarse bien al texto y actualmente contribuyen a avances significativos en el procesamiento del lenguaje natural. En estas redes, la secuencia de entrada se procesa por completo en paralelo, en contraste con las redes recurrentes, donde los elementos de la

secuencia se procesan individualmente (Vaswani et al., 2017). A diferencia de las RNN (redes neuronales recurrentes), los transformadores no requieren ordenación de datos. Por ejemplo, si la entrada al modelo es una oración en español, Transformer no necesita comenzar con la primera palabra y continuar con la siguiente, etc., para finalmente hacer una predicción. Esto permite que las redes de tipo Transformers tengan un mayor paralelismo que los RNN, lo que significa un tiempo de entrenamiento más corto (Vaswani et al., 2017).

Transformers está revolucionando el mundo de la NPL al ser la opción más popular para el procesamiento de textos. Debido a su alta eficiencia, reemplazan las redes clásicas (hasta ahora "estado del arte"). La paralelización que conlleva su implementación ha permitido el rápido desarrollo de modelos predictivos sobre conjuntos de datos tan grandes que anteriormente resultaban imprácticos procesar (Frayre & Martínez, 2022).

Métodos y técnicas de predicción de diagnósticos médicos.

El aprendizaje automático puede aplicarse para resolver problemas de muy diversos campos del conocimiento, siempre y cuando se disponga de datos necesarios para los modelos (Sampedro & Garcıa, 2012).

En la investigación científica en cuanto a salud, han surgido importantes hallazgos y descubrimientos de forma inesperada o no planificada en el contexto de un modelo de investigación para otro fin. Un ejemplo clásico es el descubrimiento de la penicilina. Esto se puede equiparar a la minería de datos en términos de "dejar que los datos hablen", es decir, procesar analíticamente datos que aporten nueva información (agrupación o clustering, clasificación, distribución, etc.) que no ha sido considerada o sugerida previamente (Pineda, 2022).

El cambio de paradigma del aprendizaje automático y la ciencia de datos, especialmente el procesamiento de grandes volúmenes de datos es precisamente una nueva mirada, diferente a la contabilidad estadística, que utiliza modelos computacionales "a partir de los datos" para extraer nueva

información relevante. El campo se enfoca en determinar la definición o las características de la pregunta que influyen en las respuestas buscadas (predicción o clasificación automática) lo proporciona el propio modelo de aprendizaje automático, no sólo el investigador, lo que minimiza el sesgo y supera a los métodos tradicionales. Estas técnicas son importantes en medicina y salud en general; muestra una importante utilidad en los sistemas de apoyo a la decisión clínica (Pineda, 2022).

Modelos y algoritmos secuenciales para Deep Learning.

Los algoritmos, en su forma más simple, son solo una secuencia de acciones, una lista de instrucciones, los algoritmos son la base de toda la informática. Si pensamos en una computadora como una pieza de hardware, un disco duro, chips de memoria, procesadores, etc. Sería imposible la tecnología moderna, sin los algoritmos. Razones generales para estudiar algoritmos:

1. Son esenciales para la informática y los sistemas inteligentes.
2. Son importantes en muchos otros dominios (biología computacional, economía, comunicaciones, ecología, física, etc.).
3. Desempeñan un papel en la innovación tecnológica.
4. Mejoran la resolución de problemas y el pensamiento analítico.

El aprendizaje automático está relacionado con el reconocimiento de distintos patrones en situaciones en las que el humano no podría detectar, ya sea por estar trabajando con una cantidad de datos muy grande, o por la dificultad de dichos patrones (Pineda, 2022). El aprendizaje automático tiene una amplia gama de aplicaciones, ya que se puede aplicar a cualquier campo del cual se quiera hacer una investigación profunda. Algunos ejemplos de aplicación podrían ser:

- Análisis del mercado.
- Reconocimiento de imágenes, voz, caracteres, etc.
- Motores de búsqueda.
- Diagnósticos médicos.

- Recuperación de la información.

Con la ayuda de varios logros innovadores, el Deep Learning ha dado un gran impulso a todo el campo del Machine Learning. Ahora, incluso programadores que no saben casi nada de esta tecnología pueden usar herramientas sencillas y eficaces para implementar programas capaces de aprender a partir de datos (Sampedro & Garcıa, 2012).

Las propiedades de los algoritmos que se debe tomar en cuenta son los siguientes:

- El tiempo secuencial: quiere decir que un algoritmo funciona en tiempo discretizado paso a paso, esto para la precisión en la sucesión de los estados computacionales al momento de un ingreso válido.
- El estado abstracto: se debe explicar el estado computacional, utilizando una distribución de primer orden y cada algoritmo es autónomo de su implementación, por esta razón los algoritmos son objetos abstractos.
- Exploración acotada: los algoritmos tienen esta propiedad por la evolución de un estado al siguiente, la cual queda totalmente explícita por una representación fija y finita.

Los algoritmos son aquellas cosas que funcionan paso a paso, y los pasos que se dan se puede representar sin tergiversación, y además tiene límite fijo con respecto a la cantidad de datos que se van a utilizar sea leer o escribir (Gonzales Balcázar, 2022).

Con el aprendizaje automático, los registros clínicos completos son entradas para el aprendizaje de algoritmos, (Pineda, 2022) . Los datos de los registros clínicos no suelen estar estructurados o tabulados, encontrándose como textos libres. Los modelos resultantes pueden ser subsecuentes para ayudar al personal de la salud a precisar los diagnósticos.

En modelos de aprendizaje automático toma el conjunto de datos y contiene 3 grupos que son el entrenamiento, validación y las pruebas. El primer grupo se va a encargar de la entrada del algoritmo con lo cual aprenderá a resolver la tarea que se le asigna; con el segundo grupo se evalúa sobre la

eficiencia con la cual se resuelve la tarea, el algoritmo tiene la oportunidad de ajustar algunos cambios y para el último grupo se asigna la pruebas y ver los resultados de cómo se comportará el algoritmo con los datos (Pineda, 2022).

La tarea de predecir no es fácil ya que existen errores y esto impide que la predicción tenga una certeza completa en el sistema, es necesario revisar los conceptos fundamentales para el estudio de las técnicas de predicción:

- Predicción: es anunciar por revelación ciencia o conjetura algo que ha de suceder (Lalaleo Achachi, 2021).
- Sistemas: es el conjunto de las cosas ordenadas y relacionadas entre sí que contribuyen a un determinado objeto (Lalaleo Achachi, 2021).

Paradigmas de diseño de algoritmos

Para el diseño de algoritmos existen tres enfoques amplios (Lalaleo Achachi, 2021).

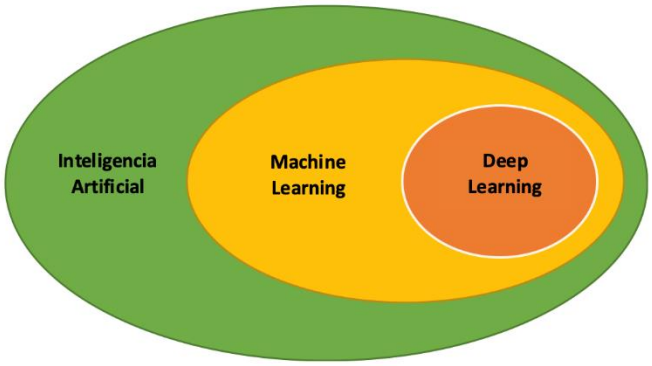
- Divide y vencerás. - menciona sobre dividir el problema en subcategorías para poder resolverlas de una manera más fácil y ágil.
- Algoritmos codiciosos. - Este enfoque siempre va a elegir el destino más cercano o el camino más corto para obtener el escenario ideal.
- Programación dinámica. - los resultados que se alcanzan se almacenan en caché y se pueden utilizar en operaciones posteriores.

Para entender el Deep Learning, hay que aclarar sus distinciones con respecto a otras disciplinas existentes que también llegan a pertenecer al área de Inteligencia Artificial (IA). El Deep Learning es como un sub-Aprendizaje del Machine Learning, pero los modelos ya incluyen como parte de su procedimiento (Pérez Ortiz de Landaluce, 2021). Este subaprendizaje no se basa en la lógica lineal, se sustenta en teorías acerca de cómo funciona el cerebro humano. Este aprendizaje tiene un potencial para aplicaciones como reconocimiento de fotografías, clasificación de imágenes y poner en práctica

creaciones de técnicas capaces de determinar emociones o eventos que aparecen en un texto (Pérez Ortiz de Landaluce, 2021). El aprendizaje profundo y su progreso es viable a la realización de predicciones acerca de comportamientos de personas y otras funciones.

Figura 1

La evolución de la IA, ML y DL



Una aplicación típica del aprendizaje automático es la clasificación de imágenes, un estudio que puede estar sujeto a diversos grados de supervisión humana según los métodos utilizados. Estos métodos favorecen los sitios con menos intervención del usuario, ya que el etiquetado manual puede llevar mucho tiempo y ser costoso, puede verse afectado por futuras actualizaciones de datos. Se dividen en tres subgrupos generales: aprendizaje supervisado, aprendizaje supervisado débilmente y aprendizaje no supervisado (Pérez Ortiz de Landaluce, 2021).

Aprendizaje supervisado

Se entiende que el proceso de ejecución de la fase de entrenamiento tiene uno o más resultados esperados para valores de entrada dados. Los datos de entrenamiento son pares de objetos, es decir, uno es la entrada y el otro es la salida esperada.

El aprendizaje supervisado es un tipo de aprendizaje que descubre la relación entre algunas variables de entrada y otras variables de salida presentando una gran cantidad de ejemplos de

entrenamiento preetiquetados o clasificados para que el algoritmo aprende a actuar sobre esas soluciones. Las personas participan en el algoritmo, lo monitorean y muestran el resultado deseado. El sistema se prueba y, si falla, los parámetros de la máquina se ajustan en la siguiente pasada para mejorar la precisión y el rendimiento. Cuanta más cantidad y calidad de datos proporcione un sistema, más exitoso debería ser. Por lo tanto, se recomienda tener una base de datos completa (Aguirre Ascona, 2019).

Aprendizaje débilmente supervisado

El aprendizaje débilmente supervisado es una rama del aprendizaje automático que se enfoca en cómo los algoritmos pueden aprender de datos débiles o mal etiquetados. Este tipo de aprendizaje es cada vez más popular porque permite entrenar modelos con recursos y tiempo limitados. También permite usar datos sin etiquetar para entrenar modelos, que pueden usarse para mejorar la precisión y la confiabilidad. El aprendizaje débilmente supervisado tiene una amplia gama de aplicaciones, como el procesamiento del lenguaje natural, el reconocimiento de imágenes, el reconocimiento de voz y más (Meseguer Esbri, 2022).

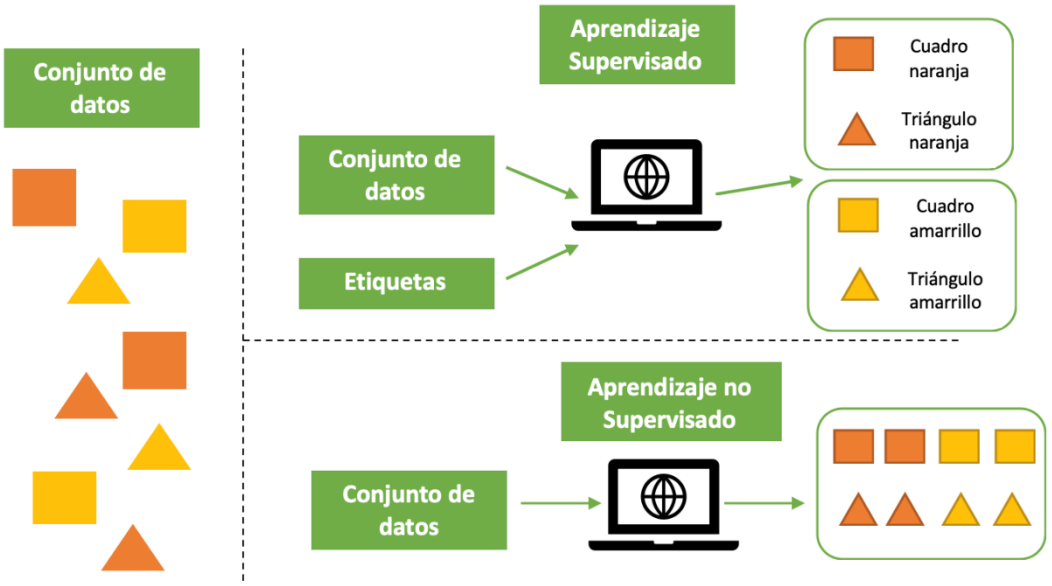
Aprendizaje no supervisado

El aprendizaje no supervisado brinda resultados sin tener que explicarle al sistema lo que desea lograr porque solo usa los datos proporcionados durante la entrada. No requieren que nadie controle las respuestas o los ejemplos etiquetados para verificar que se está realizando la clasificación correspondiente, ya que no hay intervención humana directa. Consiste en un algoritmo que busca patrones similares en los datos de entrada e intenta encontrar su estructura interna y posibles variaciones. Las bases de datos no controladas son mucho más fáciles y sencillas. A pesar de su futuro prometedor, estas redes aún no están completamente implementadas y todavía se están investigando (Arribas Jara, 2018).

El Deep Learning en lugar de basarse en la lógica lineal, el aprendizaje profundo se basa en teorías sobre cómo funciona el cerebro humano. No hay instrucciones que indiquen qué paso a seguir cuando se detecta una característica. Un programa consta de capas anidadas de nodos vinculados entre sí (Pérez Ortiz de Landaluce, 2021). La técnica de Deep Learning se define como un subaprendizaje, pero contiene modelos que incluyen como parte de su procedimiento una extracción de características.

Figura 2

Clasificación con aprendizaje supervisado y aprendizaje no supervisado



Resolver problemas de aprendizaje profundo requiere mucha potencia informática debido a la naturaleza iterativa de los algoritmos de aprendizaje profundo, su complejidad a medida que aumenta la cantidad de capas y la gran cantidad de datos necesarios para entrenar redes.

La naturaleza dinámica de los métodos de aprendizaje profundo, su capacidad para evolucionar continuamente y adaptarse a los cambios en el modelo de datos subyacente ofrece una excelente oportunidad para llevar un comportamiento más dinámico a la analítica.

Capítulo III

Desarrollo e implementación del sistema

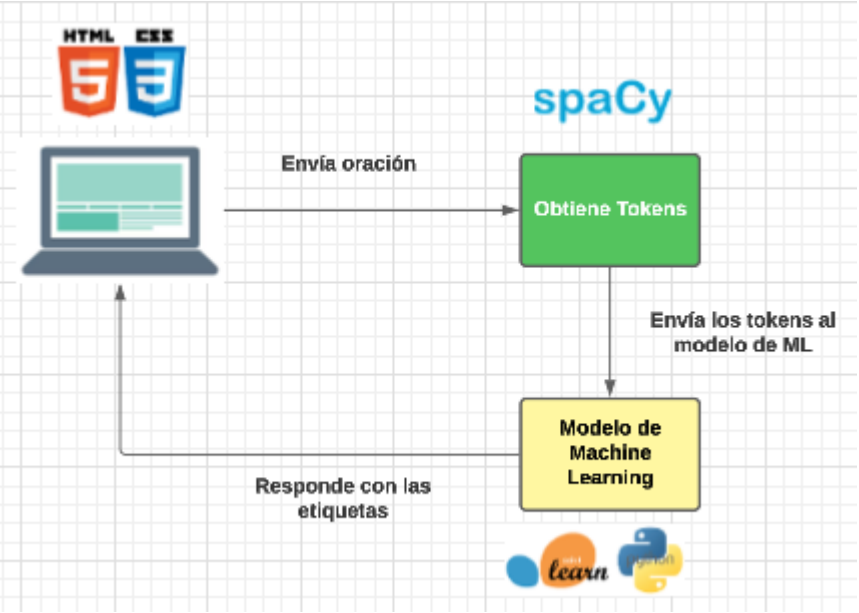
En este capítulo se describe a detalle todo el proceso que se realizó para desarrollar el sistema propuesto, un sistema de predicción de diagnósticos médicos en base a las notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático. Para el sistema se usa un corpus médico, el cual contiene un conjunto de datos que son los casos clínicos, en el cual se encuentran datos de entrenamiento, desarrollo y pruebas, todas las anotaciones de estas notas clínicas se encuentran en "CodiEsp corpus: gold estándar Spanish clinical cases coded in ICD10 (CIE10)-eHealth CLEF2020". El corpus contiene casos clínicos codificados manualmente. Todos los documentos están en idioma español y CIE10 es la terminología de codificación (es la versión en español de ICD10-CM e ICD10-PCS). Este corpus se ha muestreado aleatoriamente en tres subconjuntos: el entrenamiento, desarrollo y el conjunto para pruebas. El conjunto para el entrenamiento que este corpus ofrece contiene 500 casos clínicos, y el conjunto de desarrollo y prueba contiene 250 casos clínicos cada uno.

Las carpetas train, dev y test tienen 3 archivos separados por tabulaciones con la información de anotaciones relevantes para cada una de las 3 subpistas de CodiEsp. Una subcarpeta denominada text_file con los archivos de texto plano de los casos clínicos. Una subcarpeta llamada text_files_en con los archivos de texto sin formato traducidos automáticamente al inglés. Debido al proceso de traducción, los archivos de texto se dividen en oraciones. El corpus CodiEsp se distribuye en texto plano en codificación UTF8, donde cada caso clínico se almacena como un único archivo cuyo nombre es el identificador del caso clínico. Las anotaciones se colocan en un archivo separado por tabulaciones.

En la Figura 3 se observa el esquema de cómo será el funcionamiento del sistema web que se está desarrollando, se observará desde como empieza el envió de la oración con la interfaz web y como se procesa con la ayuda del modelo.

Figura 3

Infraestructura del sistema



En el desarrollo del sistema, se utilizó una metodología de desarrollo de software ágil, que permite implementar un producto de software, de forma organizada, ágil y en un periodo corto de tiempo, compromete el trabajo en equipo, Scrum es un proceso para construir productos, y un marco que permita gestionar el desarrollo de productos con una alta complejidad, (Kuz et al., 2018). Scrum ayuda a solventar problemas gracias a que contiene puntos positivos que son importantes en el desarrollo de proyectos software:

-Simplicidad: Los eventos organizados por Scrum están claramente identificados y se le dice a cada uno: quiénes participarán, sus objetivos, el tiempo requerido y cuál será el resultado esperado. Lo que esencialmente facilita que los miembros del equipo adopten el método (Rodríguez & Vicente, 2015).

- Inspección: Uno de los componentes destacados por Scrum es la revisión, por lo que tres de sus eventos apuntan a estos objetivos: la reunión diaria, la revisión del sprint y la retrospectiva final. A través de estos eventos, la organización puede fortalecer la metodología y descubrir en cada equipo y

cada proceso lo que necesita mejorar. Este componente es uno de los favoritos de las organizaciones porque les permite ver qué tan bien se ajusta un método a su cultura y si se están logrando los beneficios prometidos (Rodríguez & Vicente, 2015).

- Adaptación: Lo mejor del método es el deseo de cambiar las propiedades del producto. Este es uno de los componentes con mayor diferencia con respecto a los demás, ya que se pueden realizar cambios en cualquier momento, incluso durante el desarrollo del rendimiento de iteraciones o sprints diferentes, siempre que no afecte la entrega acordada. Esta personalización beneficia a la organización tanto como contribuye a la satisfacción del cliente y a los ingresos por personalización (Rodríguez & Vicente, 2015).

-Trabajo en equipo: Lo que es particularmente interesante de Scrum es cómo logra la sinergia entre las personas involucradas en el proceso en la medida en que, en cada iteración del ciclo de desarrollo, el mismo equipo se adapta a las mejoras. También significa que cada individuo es reconocido como parte integral del equipo, por lo que el efecto del cambio en las personas puede ser grande. En comparación, Scrum logra traer visibilidad al equipo porque, en las metodologías tradicionales, sus inventores no tenían una relación directa con el cliente, lo cual es importante y notable para el reconocimiento (Rodríguez & Vicente, 2015).

Metodología Scrum

En la metodología Scrum existen eventos que permiten el control de forma periódica para mostrar el avance en el desarrollo del proyecto evitando inconvenientes de tiempo, (Hema et al., 2020). En el trabajo realizado se aplica eventos y conceptos propios de esta metodología:

-Sprint: Este es un período de 1 a 4 semanas en el que el equipo realiza tareas en conjunto durante las cuales se entrega el sistema con un avance significativo del producto.

- Planificación del sprint (Sprint planning): Cada sprint tiene una planificación de acuerdo con las tareas que se llevaran a cabo y esto se comunicará en una reunión en la que participa todo el equipo de

desarrollo para determinar el objetivo del sprint en base a la integración de historias de usuario según las metas establecidas.

- Reuniones diarias (Daily Meetings): se refiere a una reunión diaria con los participantes del proyecto en un tiempo corto como de 10 a 15 minutos para informar sobre el proceso del trabajo y las tareas realizadas.

-Revisión del sprint (Sprint review): Al final del sprint, el equipo de desarrollo celebra una reunión para revisar que todas las tareas estén relacionadas con el objetivo del sprint establecido, luego se toman decisiones sobre ajustes adicionales para mejorar el sistema si es necesario.

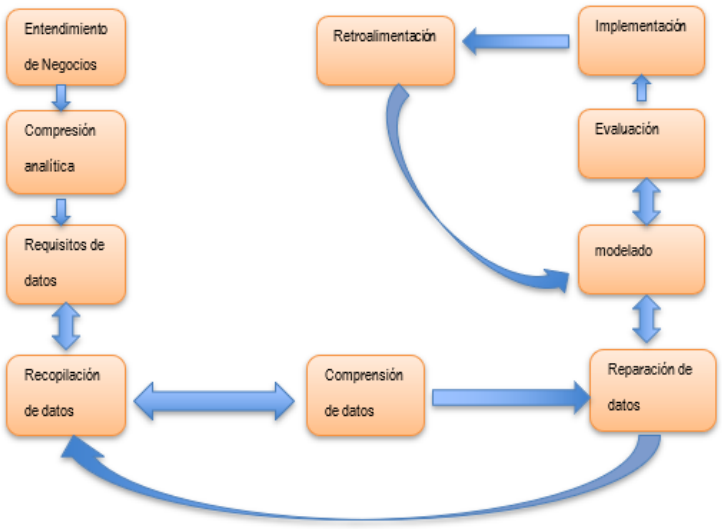
- Retrospectiva del sprint (Sprint restrospective): Luego de la reunión con el equipo, se documentan y analizan todos los obstáculos encontrados durante la ejecución del sprint y todo lo que funcionó correctamente. Así, se realizan cambios para conseguir mejoras para el siguiente sprint.

Minería de ciencia de datos

El proyecto también se dirige a conceptos como la ciencia de datos y las metodologías usadas en ese ámbito, la ciencia de datos se considera un enfoque nuevo y prometedor utilizado en la recopilación y el análisis de datos en muchas disciplinas. A pesar de la amplia difusión y aceptación, el uso de técnicas basadas en la ciencia de datos en el campo de la investigación global aún está en pañales. En ese sentido, este artículo tiene como objetivo reflexionar sobre los aportes y desafíos metodológicos que la ciencia de datos puede traer a la disciplina de los estudios globales (Lemus-Delgado & Pérez Navarro, 2020).

Figura 4

Proceso cíclico en la metodología de ciencia de datos



Comprensión empresarial: Antes de resolver cualquier problema empresarial, debe comprenderlo correctamente. Comprender el negocio crea una base concreta que conduce a la solución de una pregunta simple. Necesitamos tener claro el problema que estamos tratando de resolver.

Perspectiva analítica: El enfoque analítico a seguir debe decidirse en función del desempeño comercial anterior. Hay cuatro tipos de enfoques: enfoque descriptivo (estado actual y datos presentados), enfoque de diagnóstico, enfoque predictivo (predicción de tendencias o la probabilidad de eventos futuros) y enfoque preceptivo (cómo se debe resolver realmente el problema).

Requerimientos de datos: El método de análisis elegido arriba indica el contenido, formas y fuentes de los datos a recolectar. El proceso de requerimientos de información implica encontrar respuestas a preguntas como "qué", "dónde", "cuándo", "por qué", "cómo" y "quién".

Recopilación de datos: Los datos recopilados se pueden obtener de cualquier forma aleatoria. Por lo tanto, estos datos deben validarse según el enfoque elegido y el resultado que se desea lograr.

Comprensión de datos: Comprensión de datos responde a la pregunta "¿Los datos recopilados representan un problema que vale la pena resolver?". Las estadísticas descriptivas calculan las medidas aplicadas al material para lograr el contenido y la calidad del material. Este paso puede llevar de vuelta al paso anterior para su corrección.

Preparación de datos: Entendamos esto conectando este concepto con dos analogías. Una es lavar las verduras recién cogidas y la otra es poner en el plato sólo los productos deseados para comer en el buffet. Lavar las verduras significa eliminar la suciedad, es decir, materiales no deseados de los datos. Aquí se realiza la reducción de ruido. No vale la pena considerar tomar solo artículos comestibles en el plato para un procesamiento posterior, a menos que se necesite información más detallada. Todo este proceso implica transformación, normalización, etc.

Modelado: el modelado decide si la información preparada para el procesamiento es adecuada o necesita más refinamiento. Esta fase se centra en la construcción de modelos predictivos/descriptivos.

Evaluación: La evaluación del modelo se produce durante el desarrollo del modelo. Verifique la calidad del modelo evaluado y también si cumple con los requisitos comerciales. Pasa por una fase de medición de diagnóstico y una fase de prueba de significación estadística.

Implementación: Dado que el modelo ha sido efectivamente evaluado, está listo para su implementación en el mercado comercial. En la fase de implementación, verificamos cuánto admite el modelo en el entorno externo y se desempeña mejor en comparación con otros.

Retroalimentación: La retroalimentación es un objetivo necesario para ayudar a refinar el modelo y lograr su desempeño y efectos. Los pasos involucrados en proporcionar retroalimentación definen el proceso de revisión, registran, miden el desempeño y refinan las revisiones.

Análisis del sistema

De acuerdo con la metodología scrum y sus eventos ya mencionados, para la especificación de requisitos del sistema se usa Historias de usuario (HU) con esto se define los roles que cada miembro del

equipo tiene dentro del proyecto, el Product Owner será el encargado de optimizar y maximizar el valor del producto, así como notificar el requisito del producto. El equipo del desarrollo (Development Team) que se encarga de desarrollar el producto dentro del periodo de tiempo establecido. El (Scrum Master) debe gestionar el proceso y ayudar a eliminar impedimentos que afecten la entrega del producto. Además, se encarga de actividades como revisar el cumplimiento de reglas y principios. El Scrum Master es responsable de que el proceso se lleve adelante (Rad & Turley, 2019). Se realiza una designación de roles para el desarrollo de este proyecto, los roles tendrán consigo información sobre el rol y del miembro asignado, y una descripción acerca de la función dentro del proyecto.

Team Scrum

Para el desarrollo del sistema de predicción de diagnósticos se ha definido roles para cada miembro que se encuentra dentro del proyecto. En la tabla 1: Team Scrum se indica la información acerca de los roles asignados.

Tabla 1

Team Scrum

N°.	Rol Scrum	Integrante	Funciones
01	Product Owner	Uyaguari Uyaguari, Álvaro Danilo Msc.	Encargado de guiar el proceso de desarrollo y controlar el tiempo de entrega y evaluar el producto.
02	Scrum Master	Luis Lenin Quishpe Rocha	Responsables de revisar actividades y verificar que el tiempo requerido del sprint se cumpla, es el líder del equipo.

N°.	Rol Scrum	Integrante	Funciones
03	Development Team	Luis Lenin Quishpe Rocha Macarena Lizbeth Cruz Caiza	Son los encargados de desarrollar e implementar el sistema de predicciones de diagnósticos.

La asignación de roles permite identificar quién está a cargo de las tareas, el proyecto está conformado por dos integrantes, uno cumple con la función de Scrum Master y también tiene su parte en otras actividades correspondientes al equipo de desarrollo (Development Team). La información que se adquiere por esta asignación de roles permite documentar las historias de usuario. El Scrum Master lleva a cabo la reunión inicial con los miembros del proyecto como el Product Owner y el equipo de desarrollo.

Historias de Usuario

Las Historias de Usuario se colocan en la Tabla 2, en donde se detalla el nombre de la historia, el rol, las características y el resultado logrado en la implementación del sistema.

Tabla 2

Historias de Usuario

ID de la H. U	Nombre de la H. U	Rol de la persona	Característica	Resultado
1	H. U. 01	Como usuario dentro del equipo de desarrollo (Development Team).	Quiero crear un dataset que contenga los términos médicos de las notas clínicas (diagnósticos) ofrecidas en el corpus.	Este dataset servirá para entrenar el modelo de Machine Learning.

ID de la H. U	Nombre de la H. U	Rol de la persona	Característica	Resultado
2	H. U. 02	Como usuario dentro del equipo de desarrollo (Development Team).	Quiero crear un algoritmo que tome la información de las notas médicas (diagnósticos) para colocar etiquetas en los términos correspondientes.	Para reconocer los términos médicos usados en las notas clínicas a través de características de los archivos donde se encuentran las palabras relacionadas con diagnósticos.
3	H. U. 03	Como usuario dentro del equipo de desarrollo (Development Team).	Quiero crear un dataset que contenga los términos médicos de las notas clínicas (procedimientos) ofrecidas en el corpus.	Este dataset servirá para entrenar el modelo de Machine Learning.
4	H. U. 04	Como usuario dentro del equipo de desarrollo (Development Team).	Quiero crear un algoritmo que tome la información de las notas médicas (procedimientos)	Para reconocer los términos médicos usados en las notas

ID de la H. U	Nombre de la H. U	Rol de la persona	Característica	Resultado
			para colocar etiquetas en los términos correspondientes.	clínicas a través de características de los archivos donde se encuentran las palabras relacionadas con los procedimientos.
5	H. U. 05	Como programador dentro del equipo de desarrollo (Development Team).	Quiero usar el dataset de diagnósticos y procedimientos que se encuentran etiquetados con IOB con las entidades médicas reconocidas con la ayuda de algoritmos de Machine Learning e implementarlos con el modelo para revisar su funcionamiento con textos médicos nuevos.	Para obtener una predicción en la identificación de diagnósticos en nuevas notas de contexto médico.
6	H. U. 06	Como programador dentro del equipo de	Quiero que el sistema de predicción de diagnósticos médicos realice el	Para obtener resultados y ver cuál es la tasa de

ID de la H. U	Nombre de la H. U	Rol de la persona	Característica	Resultado
		desarrollo (Development Team).	reconocimiento de los términos introducidos usando el modelo de Machine Learning que contiene el dataset etiquetado.	predicción con el sistema implementado.

Con las historias de usuario creadas el siguiente paso es realizar el Product Backlog del proyecto que es una lista de trabajo de elementos (historias de usuarios, errores pendientes, otras tareas, etc.) que el equipo de software utiliza para coordinar las piezas del trabajo. Los desarrolladores suelen crear nuevas funciones, modificar las funciones existentes y corregir errores en función de los elementos principales del trabajo pendiente (Sedano et al., 2019), de acuerdo con la prioridad y dificultad de implementación.

La Tabla 3 indica el Product Backlog, muestra las historias de usuario que contiene el proyecto y que se desarrollarán en el transcurso del tiempo establecido, la estimación del tiempo es factor clave ya que deben cumplirse con horas, fecha de inicio, fecha de finalización y cuál es el Sprint al que corresponde.

Tabla 3

Product Backlog del Proyecto

Historia de Usuario	Nombre	Estimación de tiempo (días)	Fecha de inicio	Fecha de fin	N° de Sprint
1	H. U. 01	5	24/10/2022	28/10/2022	01
2	H. U. 02	10	31/10/2022	11/11/2022	01

3	H. U. 03	5	14/11/2022	18/11/2022	02
4	H. U. 04	15	21/11/2022	09/12/2022	02
5	H. U. 05	5	12/12/2022	16/12/2022	03
6	H. U. 06	20	19/12/2022	13/01/2022	03

Diseño del sistema

El reconocimiento de entidades nombradas es un proceso de clasificación secuencial, en el cual el sistema NER debe asignar una etiqueta a cada palabra que encuentre en una oración o un texto, esto para identificar a qué entidad pertenece, el esquema IOB ha servido como un formato común que permite usar tres tipos de etiquetas el cual intenta representar en el contexto de la tarea NER, el rol de cada palabra dentro de la oración o texto dado.

Figura 5

Oraciones etiquetadas usando el esquema IOB

Oración	**Se**	**trata**	**de**	**una**	**mujer**	**de**	**29**	**años**
Etiquetas IOB	O	O	O	O	O	O	O	O
Oración	**sometida**	**a**	**un**	**estudio**	**ecográfico**	**pélvico**		
Etiquetas IOB	O	O	O	O	B	I		

Desarrollo del sistema

En esta sección se detalla sobre cómo se llevó a cabo el desarrollo del sistema propuesto, sistema de predicción de diagnósticos médicos en base a las notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático. La entrada del sistema es un conjunto de notas clínicas obtenidas por un corpus médico, dentro de este corpus se encuentra las notas clínicas de diagnósticos y de procedimientos, estos se encuentran en texto plano, además el corpus ofrece archivos por separado de los códigos de los diagnósticos y procedimientos. La visión del proyecto se dirige hacia: i) Realizar la detección de los términos médicos sean procedimientos o diagnósticos mediante sus códigos, ii) Se extrae de las notas clínicas los términos y sus posiciones dentro del texto plano para etiquetar cada

término de acuerdo con sus características, y iii) se implementa con el modelo de Machine Learning para colocar nuevo texto médico y ver cómo identifica a cada palabra introducida.

Herramientas de software usadas para el desarrollo

La Tabla 4 muestra el nombre y la descripción de las herramientas usadas para el desarrollo del proyecto de un sistema de predicción de diagnósticos médicos.

Tabla 4

Herramientas de desarrollo

Herramienta	**Descripción**
Para la implantación de la lógica del sistema y el id de desarrollo.	Se utiliza Python con la versión 3.10.5 y el editor se usa el Visual Studio Code versión 1.74.3
Para el procesamiento de lenguaje natural y reconocimiento de entidades.	Para el procesamiento de texto se usa librerías como Spacy 3.5.0
Para clasificar el texto y colocar las etiquetas.	El formato BIO/IOB (abreviatura de interior, exterior, comienzo) es un formato de etiquetado común para etiquetar tokens en una tarea de fragmentación en lingüística computacional.
Algoritmos para la identificación de entidades médicas y etiquetado IOB.	Se usa BERT que es la representación de codificador bidireccional de transformadores el cual usa información etiquetada con formato IOB.

La metodología Scrum consiste en que después de preparar un backlog de producto con historias de usuarios y los números de sprint correspondientes, se ejecuta un plan para cada sprint para secuenciar y priorizar las tareas consideradas las más importantes desde el punto de vista del desarrollo, por lo que el concepto de sprint se debe aplicar la acumulación (Kayes et al., 2016).

Sprint 01: Creación de un dataset para diagnósticos

El Sprint uno toma en cuenta a la Historia de usuario H. U. 01 que se encuentra en la Tabla 2, donde menciona que debe crearse un dataset de los diagnósticos para luego entrenar con un modelo de Machine Learning para procesamiento de lenguaje natural.

Historia de Usuario Detalladas

La creación del dataset viene con la Historia de Usuario H. U. 01 del sistema de predicción de diagnósticos, la cual especifica la persona designada de su desarrollo y cuáles serán los elementos de aprobación para crear el conjunto de datos. La Tabla 5 indica la información que debe llevar la Historia de Usuario detallada.

Tabla 5

Historia de Usuario para la Creación del conjunto de datos

Historia de Usuario	
Número: H. U. 01	**Usuario:** Administrador
Nombre Historia: Creación de un dataset para diagnósticos	**Sprint uno**
Prioridad: 1	**Riesgo de Desarrollo:** 3
Tiempo de estimación (días): 5	**Interacción asignada:** 1
Desarrollador responsable: Macarena Lizbeth Cruz Caiza	
Descripción: Como desarrollador quiero crear un dataset que contenga los términos médicos de las notas clínicas (diagnósticos) ofrecidas en el corpus.	
Validación: ● Se realiza la revisión del archivo con códigos médicos pertenecientes a los diagnósticos que se encuentran dentro de los textos planos en el corpus.	

- Se crea un archivo .xml que será el primer paso para luego convertir al formato .csv que se requiere para el dataset.

Sprint Backlog

El Sprint Backlog especifica las tareas que se realizarán durante el desarrollo del sprint y quién es el responsable de cada ejecución, los días que desea que se ejecute el sprint, el tiempo estimado en horas y el estado actual de donde se encontró cada tarea. Se observa que el Sprint Backlog está completo. La Tabla 6: Sprint Backlog 01 presenta los detalles.

Tabla 6

Sprint Backlog 01

Sprint 1	**Fecha Inicio**	**24/10/2022**	**Fecha Fin**	**28/10/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
H. U. 01	Revisión del corpus y los códigos de los diagnósticos dentro del corpus y sus posiciones.	8	24/10/2022	24/10/2022	Macarena Lizbeth Cruz Caiza	Finalizado
H. U. 01	Creación del algoritmo para pasar de un texto normal al formato	16	25/10/2022	26/10/2022	Macarena Lizbeth Cruz Caiza	Finalizado

Sprint 1	**Fecha Inicio**	**24/10/2022**	**Fecha Fin**	**28/10/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
	necesario para trabajar dentro del sistema.					
H. U. 01	Almacenar la información que se genera con el algoritmo para guardarlo como un dataset.	16	27/10/2022	28/10/2022	Macarena Lizbeth Cruz Caiza	Finalizado

Sprint 01: Desarrollo del algoritmo que tome la información de las notas médicas (diagnósticos)

Para el desarrollo del Sprint uno, se tomó en cuenta la Historia de Usuario H. U. 02 que se encuentra en la tabla 2, se menciona que se debe desarrollar un algoritmo para tomar la información de diagnósticos que en este caso nos guiamos con el formato BIO/IOB que permite el etiquetado común para etiquetar tokens en una tarea de fragmentación en lingüística computacional como el reconocimiento de entidad nombrada.

Historia de Usuario Detalladas

El desarrollo de un algoritmo utilizando técnicas de Machine Learning para que tome la información de las notas médicas, se encuentra en la Historia de Usuario H. U. 02 del sistema de predicción de diagnósticos, la cual especifica la persona designada de su desarrollo y cuáles serán los elementos de aprobación para crear el algoritmo. La Tabla 7 indica la información que debe llevar la Historia de Usuario detallada.

Tabla 7

Historia de Usuario para el desarrollo del algoritmo del sistema de predicción de diagnósticos

Historia de Usuario	
Número: H. U. 02	**Usuario:** Administrador
Nombre Historia: Desarrollo del algoritmo que tome la información de las notas médicas (diagnósticos)	**Sprint uno**
Prioridad: 1	**Riesgo de desarrollo: 3**
Tiempo de estimación (días): 10	**Interacción asignada:** 1
Desarrollador responsable: Macarena Lizbeth Cruz Caiza	
Descripción: Quiero crear un algoritmo que tome la información de las notas médicas (diagnósticos) para colocar etiquetas en los términos correspondientes.	
Validación: ● Se desarrolla un algoritmo para tomar la información del corpus, en donde, se aplica la evaluación para colocar etiquetas en los términos de las notas clínicas. ● Se revisa que las etiquetas sean con el formato IOB que se necesita.	

Sprint Backlog

El Sprint Backlog especifica las tareas que se realizarán durante el desarrollo del sprint y quién es el responsable de cada ejecución, los días que desea que se ejecute el sprint, el tiempo estimado en horas y el estado actual de donde se encontró cada tarea. Se observa que el Sprint Backlog está completo. La Tabla 8: Sprint Backlog 02 presenta los detalles.

Tabla 8

Sprint Backlog 02

Sprint 1	**Fecha Inicio**	**31/10/2022**	**Fecha Fin**	**11/11/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
H. U. 02	Creación de un algoritmo para colocar las etiquetas correspondientes a los términos médicos que contengan código perteneciente al diagnóstico.	40	31/10/2022	04/11/2022	Macarena Lizbeth Cruz Caiza	Finalizado
H. U. 02	Creación de un algoritmo para colocar las etiquetas correspondientes a los términos que no sean médicos y no tengan relación	16	07/11/2022	08/10/2022	Macarena Lizbeth Cruz Caiza	Finalizado

Sprint 1	**Fecha Inicio**	**31/10/2022**	**Fecha Fin**	**11/11/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
	con los diagnósticos.					
H. U. 02	Creación de un algoritmo para colocar las etiquetas correspondientes a los términos que sean la continuación de una entidad médica relacionada con diagnósticos.	16	09/11/2022	10/11/2022	Macarena Lizbeth Cruz Caiza	Finalizado
H. U. 02	Generación de pruebas del sistema con el algoritmo para comprobar el proceso de etiquetado en los términos.	8	11/11/2022	11/11/2022	Macarena Lizbeth Cruz Caiza	Finalizado

Resultado del Sprint

Los resultados obtenidos durante el proceso de desarrollo, ejecución y finalización del sprint. Una vez creado el algoritmo que permite etiquetar los términos médicos reconociendo a qué entidad pertenece sea que tengas los términos médicos de diagnóstico que pertenecen a un código, o que sea la continuación de esa entidad o simplemente no tenga relación alguna.

Los términos que se encontraban en texto plano dentro del corpus ahora deben estar etiquetados conforme el algoritmo desarrollado para el etiquetamiento IOB para los términos que pertenecen a diagnósticos.

Sprint 02: Creación de un dataset para procedimientos

El Sprint dos, toma en cuenta a la Historia de usuario H. U. 03 que se encuentra en la Tabla 2, donde menciona que debe crearse un dataset de los procedimientos para luego entrenar con un modelo de Machine Learning para procesamiento de lenguaje natural.

Historia de Usuario Detalladas

La creación del dataset viene con la Historia de Usuario H. U. 03 del sistema de predicción de diagnósticos, la cual especifica la persona designada de su desarrollo y cuáles serán los elementos de aprobación para crear el conjunto de datos (términos con códigos de procedimientos). La Tabla 9 indica la información que debe llevar la Historia de Usuario detallada.

Tabla 9

Historia de Usuario para la Creación del conjunto de datos. (Procedimientos)

Historia de Usuario	
Número: H. U. 03	**Usuario:** Administrador
Nombre Historia: Desarrollo del algoritmo que tome la información de las notas médicas (procedimientos)	**Sprint dos**

Historia de Usuario	
Prioridad: 1	**Riesgo de desarrollo: 3**
Tiempo de estimación (días): 5	**Interacción asignada:** 1
Desarrollador responsable: Luis Lenin Quishpe Rocha	
Descripción: Quiero crear un dataset que contenga los términos médicos de las notas clínicas (procedimiento) ofrecidas en el corpus.	
Validación: ● Se realiza la revisión del archivo con códigos médicos pertenecientes a los diagnósticos que se encuentran dentro de los textos planos en el corpus. ● Se crea un archivo .xml que será el primer paso para luego convertir al formato .csv que se requiere para el dataset.	

Sprint Backlog

El Sprint Backlog especifica las tareas que se realizarán durante el desarrollo del sprint y quién es el responsable de cada ejecución, los días que desea que se ejecute el sprint, el tiempo estimado en horas y el estado actual de donde se encontró cada tarea. Se observa que el Sprint Backlog está completo. La Tabla 10: Sprint Backlog 03 presenta los detalles.

Tabla 10

Sprint Backlog 03

Sprint 1	**Fecha Inicio**	**14/11/2022**	**Fecha Fin**	**18/11/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
H. U. 03	Revisión del corpus y los códigos de los	8	14/11/2022	14/11/2022	Luis Lenin Quishpe Rocha	Finalizado

Sprint 1	**Fecha Inicio**	**14/11/2022**	**Fecha Fin**	**18/11/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
	procedimientos dentro del corpus y sus posiciones.					
H. U. 03	Creación del algoritmo para pasar de un texto normal al formato necesario para trabajar dentro del sistema.	16	15/11/2022	16/11/2022	Luis Lenin Quishpe Rocha	Finalizado
H. U. 03	Almacenar la información que se genera con el algoritmo para guardarlo como un dataset.	16	17/11/2022	18/11/2022	Luis Lenin Quishpe Rocha	Finalizado

Sprint 02: Desarrollo del algoritmo que tome la información de las notas médicas (procedimientos)

Para el desarrollo del Sprint dos, se tomó en cuenta la Historia de Usuario H. U. 04 que se encuentra en la tabla 2, se menciona que se debe desarrollar un algoritmo para tomar la información de diagnósticos que en este caso nos guiamos con el formato BIO/IOB que permite el etiquetado común

para etiquetar tokens en una tarea de fragmentación en lingüística computacional como el reconocimiento de entidad nombradas.

Historia de Usuario Detalladas

El desarrollo de un algoritmo utilizando técnicas de Machine Learning para que tome la información de las notas médicas, se encuentra en la Historia de Usuario H. U. 04 del sistema de predicción de diagnósticos, la cual especifica la persona designada de su desarrollo y cuáles serán los elementos de aprobación para crear el algoritmo (términos con códigos de procedimientos). La Tabla 11 indica la información que debe llevar la Historia de Usuario detallada.

Tabla 11

Historia de Usuario para el desarrollo del algoritmo del sistema de predicción de diagnósticos (términos con códigos de procedimientos)

Historia de Usuario	
Número: H. U. 04	**Usuario:** Administrador
Nombre Historia: Desarrollo del algoritmo que tome la información de las notas médicas (procedimientos)	**Sprint dos**
Prioridad: 1	**Riesgo de desarrollo: 3**
Tiempo de estimación (días): 15	**Interacción asignada:** 1
Desarrollador responsable: Luis Lenin Quishpe Rocha	
Descripción: Quiero crear un algoritmo que tome la información de las notas médicas (procedimientos) para colocar etiquetas en los términos correspondientes.	
Validación: ● Se desarrolla un algoritmo para tomar la información del corpus, en donde, se aplica la evaluación para colocar etiquetas en los términos de las notas clínicas.	

Historia de Usuario
● Se revisa que las etiquetas sean con el formato IOB que se necesita.

Sprint Backlog

El Sprint Backlog especifica las tareas que se realizarán durante el desarrollo del sprint y quién es el responsable de cada ejecución, los días que desea que se ejecute el sprint, el tiempo estimado en horas y el estado actual de donde se encontró cada tarea. Se observa que el Sprint Backlog está completo. La Tabla 12: Sprint Backlog 04 presenta los detalles.

Tabla 12

Sprint Backlog 04

Sprint 1	**Fecha Inicio**	**21/11/2022**	**Fecha Fin**	**09/12/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
H. U. 04	Creación de un algoritmo para colocar las etiquetas correspondientes a los términos médicos que contengan código perteneciente al procedimiento.	16	21/11/2022	22/11/2022	Luis Lenin Quishpe Rocha	Finalizado

Sprint 1	**Fecha Inicio**	**21/11/2022**	**Fecha Fin**	**09/12/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
H. U. 04	Creación de un algoritmo para colocar las etiquetas correspondientes a los términos que no sean médicos y no tengan relación con los procedimientos.	64	23/11/2022	02/11/2022	Luis Lenin Quishpe Rocha	Finalizado
H. U. 04	Creación de un algoritmo para colocar las etiquetas correspondientes a los términos que sean la continuación de una entidad médica	24	05/12/2022	07/12/2022	Luis Lenin Quishpe Rocha	Finalizado

Sprint 1	**Fecha Inicio**	**21/11/2022**	**Fecha Fin**	**09/12/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
	relacionada con procedimientos.					
H. U. 04	Generación de pruebas del sistema con el algoritmo para comprobar el proceso de etiquetado en los términos.	16	08/12/2022	09/12/2022	Luis Lenin Quishpe Rocha	Finalizado

Resultado del Sprint

Los resultados obtenidos durante el proceso de desarrollo, ejecución y finalización del sprint. Una vez creado el algoritmo que permite etiquetar los términos médicos reconociendo a qué entidad pertenece sea que tengas los términos médicos de diagnóstico que pertenecen a un código, o que sea la continuación de esa entidad o simplemente no tenga relación alguna.

Los términos que se encontraban en texto plano dentro del corpus ahora deben estar etiquetados conforme al algoritmo desarrollado para el etiquetamiento IOB para los términos que pertenecen a procedimientos.

Sprint 03: Creación de un dataset con los diagnósticos y procedimientos

El Sprint tres, toma en cuenta a la Historia de usuario H. U. 05 que se encuentra en la Tabla 2, donde menciona que debe crearse un dataset de los diagnósticos y procedimientos para luego entrenar con un modelo de Machine Learning para procesamiento de lenguaje natural.

Historia de Usuario Detalladas

La creación del dataset viene con la Historia de Usuario H. U. 05 del sistema de predicción de diagnósticos, la cual especifica la persona designada de su desarrollo y cuáles serán los elementos de aprobación para crear el conjunto de datos (términos con códigos de diagnósticos y procedimientos). La Tabla 13 indica la información que debe llevar la Historia de Usuario detallada.

Tabla 13

Historia de Usuario para la creación de un dataset (Procedimientos)

Historia de Usuario	
Número: H. U. 05	**Usuario:** Administrador
Nombre Historia: Desarrollo del algoritmo que tome la información de las notas médicas (diagnósticos y procedimientos)	**Sprint tres**
Prioridad: 1	**Riesgo de desarrollo: 3**
Tiempo de estimación (días): 5	**Interacción asignada:** 1
Desarrollador responsable: Luis Lenin Quishpe Rocha	
Descripción: Quiero crear un dataset de diagnósticos y procedimientos que se encuentran etiquetados con IOB con las entidades médicas reconocidas con la ayuda de algoritmos de Machine Learning e implementarlos con el modelo para revisar su funcionamiento con textos médicos nuevos.	
Validación: ● Se realiza la revisión de los dataset de diagnósticos y procedimientos para tener un nuevo corpus etiquetado con las notas clínicas del corpus inicial. ● Se elige un modelo de entrenamiento para implementar el corpus etiquetado.	

Sprint Backlog

El Sprint Backlog especifica las tareas que se realizarán durante el desarrollo del sprint y quién es el responsable de cada ejecución, los días que desea que se ejecute el sprint, el tiempo estimado en horas y el estado actual de donde se encontró cada tarea. Se observa que el Sprint Backlog está completo. La Tabla 14: Sprint Backlog 05 presenta los detalles.

Tabla 14

Sprint Backlog 05

Sprint 1	**Fecha Inicio**	**12/12/2022**	**Fecha Fin**	**16/12/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
H. U. 05	Revisión dentro del corpus etiquetado y revisar sobre cómo se encuentra y si los términos de diagnósticos y procedimiento.	8	12/12/2022	12/11/2022	Luis Lenin Quishpe Rocha	Finalizado
H. U. 05	Creación del dataset general que servirá para implementar dentro del modelo.	16	13/12/2022	14/12/2022	Luis Lenin Quishpe Rocha	Finalizado

Sprint 1	**Fecha Inicio**	**12/12/2022**	**Fecha Fin**	**16/12/2022**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
H. U. 05	Obtener resultados sobre la identificación de los términos de diagnósticos y procedimiento cuando se envíe texto nuevo.	16	15/12/2022	16/12/2022	Luis Lenin Quishpe Rocha	Finalizado

Sprint 03: Reconocimiento de entidades utilizando el modelo implementado.

El Sprint tres, toma en cuenta a la Historia de usuario H. U. 06 que se encuentra en la Tabla 2, donde menciona que debe usar el modelo de Machine Learning utilizado para realizar el reconocimiento de las entidades ingresadas para lo cual ocupa el dataset etiquetado de procedimientos y diagnósticos.

Historia de Usuario Detalladas

La revisión del modelo viene con la Historia de Usuario H. U. 06 del sistema de predicción de diagnósticos, la cual especifica la persona designada de su desarrollo y cuáles serán los elementos de aprobación para la revisión del modelo juntamente con el dataset que contiene los datos de procedimientos y diagnósticos. La Tabla 15 indica la información que debe llevar la Historia de Usuario detallada.

Tabla 15

Historia de Usuario para la implementación del modelo

Historia de Usuario	
Número: H. U. 06	**Usuario:** Administrador
Nombre Historia: Revisión del modelo utilizado para la predicción de diagnósticos.	**Sprint tres**
Prioridad: 1	**Riesgo de desarrollo: 3**
Tiempo de estimación (días): 20	**Interacción asignada:** 1
Desarrollador responsable: Macarena Lizbeth Cruz Caiza	
Descripción: Quiero que el sistema de predicción de diagnósticos médicos realice el reconocimiento de los términos introducidos usando el modelo de Machine Learning que contiene el dataset etiquetado.	
Validación: ● Se realiza la revisión del modelo usado para la predicción y observar si reconocimiento de las entidades médicas. ● Se elige el dataset creado anteriormente como conjunto de datos.	

Sprint Backlog

El Sprint Backlog especifica las tareas que se realizarán durante el desarrollo del sprint y quién es el responsable de cada ejecución, los días que desea que se ejecute el sprint, el tiempo estimado en horas y el estado actual de donde se encontró cada tarea. Se observa que el Sprint Backlog está completo. La Tabla 16: Sprint Backlog 06 presenta los detalles.

Tabla 16

Sprint Backlog 06

Sprint 1	**Fecha Inicio**	**19/12/2022**	**Fecha Fin**	**13/01/2023**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
H. U. 06	Subir el modelo dentro del sitio web “paperspace” para usar una GPU Y CPU adecuadas.	40	19/12/2022	23/12/2022	Macarena Lizbeth Cruz Caiza	Finalizado
H. U. 06	Creación de una cuenta Pro dentro del sitio “paperspace” con el objetivo de tener a disposición un entorno adecuado para el modelo.	40	26/12/2022	30/12/2022	Macarena Lizbeth Cruz Caiza	Finalizado
H. U. 06	Realizar la interfaz web donde mostrará	40	02/01/2023	06/01/2023	Macarena Lizbeth Cruz Caiza	Finalizado

Sprint 1	**Fecha Inicio**	**19/12/2022**	**Fecha Fin**	**13/01/2023**	**Jornada**	**8**
HU ID	Tareas	Horas	Inicio	Fin	Responsable	Estado
	las entidades reconocidas con el modelo trabajando con el dataset.					
H. U. 06	Revisar la salida del modelo y verificar si se identifica correctamente las entidades de procedimientos y diagnósticos.	40	09/01/2023	13/01/2023	Macarena Lizbeth Cruz Caiza	Finalizado

Resultado del Sprint

Los resultados obtenidos durante el proceso de desarrollo, ejecución y finalización del sprint **se** presentan a continuación:

Primeramente, se implementó los modelos de Machine Learning: Artificial Neural Network, con el fin de, predecir entidades médicas, en otras palabras, se proporcionará al modelo entradas que son oraciones relacionadas con el contexto médico y éste retornará como salidas los términos médicos etiquetados como “DIAGNÓSTICO” o “PROCEDIMIENTO”.

Lo siguiente era entrenar el modelo de Machine Learning con el dataset creado en los Sprint explicados anteriormente, para lo cual, se tuvo que subir el modelo de ML implementado al servidor

“paperspace”, principalmente porque en el computador donde se desarrolló el modelo no contaba con los recursos computacionales necesarios.

Con respecto a paperspace, se adquirió la cuenta Pro para después subir el modelo de ML, donde se implementará el dataset que se encuentra dentro de la página de Huggingface como un conjunto de datos disponible con el formato requerido para ser subido y sea público.

El sistema está construido con Python por esto se utiliza Flask para construir la interfaz web, ya que Flask es visto como un frameworks útil dentro del lenguaje de programación usado, que permite hacer páginas web gracias a su conjunto de herramientas necesarias para realizar una aplicación web funcional, a la vez también se dispone de extensiones que al momento de instalar se acoplan bien con Flask y se sigue dotando de funcionalidades.

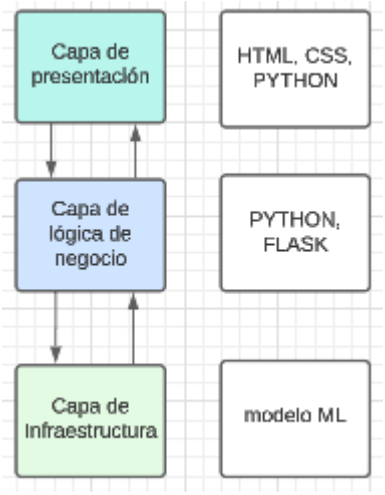
La interfaz web tendrá visualmente un campo que está destinado a ser la entrada, se colocará un texto relacionado al contexto médico que servirá como los datos a que van a ser etiquetados, esta información será procesada por el modelo e identificará los términos que encuentre.

El otro campo contendrá la interfaz que es la salida en esta aparecerá los términos etiquetados que se encontró en el procesamiento del texto que se tuvo como entrada y paso por el modelo, al final se mostrará cuáles son las palabras identificadas como entidades médicas sea de procedimiento o diagnóstico.

En la Figura 6 se muestra la arquitectura para el sistema web que se usará para mostrar el modelo de predicción, se visualiza las capas que contiene el sistema construido.

Figura 6

Capas del sistema web



En la Figura 7 se muestra un fragmento del código que se usó para convertir los archivos proporcionado por el corpus que se encuentran en formato txt a un formato xml, esto se lo hace mediante funciones donde se lee los archivos iniciales y luego se comienza a escribir el archivo xml, tomando en cuenta cuántos archivos se encuentran en el corpus y el nombre de ellos al igual de las palabras médicas que se encuentren.

Figura 7

Código para convertir el texto a un archivo xml

```
> OCTAVO SEMESTRE > TESIS > convert_xml_corpus.py > ...
46  def medicine_files(segments):
57      print(medicine_worlds)
58      return medicine_worlds
59  ##Escribir xml
60  def write_xml(xmlfile,segments, names_files,text_files,medicine_words):
61      with open (xmlfile,"w",encoding="utf8") as f:
62          f.write("<?xml version='1.0' encoding='UTF-8'?>\n")
63          f.write("<Corpus id='v01' annotationDate='2022-12-30+20:00' annotationType='standoff' author='ESPE-2022' d
64
65          for data in zip(names_files,text_files, medicine_words):
66              f.write(f"<document id='{data[0]}'>\n")
67              f.write(f"    <unit id='{data[0]}.u1'>\n")
68              f.write(f"      <text>{data[1]}</text>\n")
69              for word in data[2]:
70                  f.write(f"      <e id='{word[0]}' grp='{word[1]}' len='{word[2]}' offset='{word[3]}' >{word[4]}</e
71              f.write("    </unit>\n")
72              f.write("</document>\n")
73          f.write("</Corpus>")
74      pass
75
76
77  filename = "corpus_v02.txt"
78  corpus = open_corpus(filename)
79  files = each_file_data(corpus)
```

La Figura 8 presenta un fragmento del código realizado en Python que se usó para encontrar las entidades médicas y en caso de ser halladas poder colocar las etiquetas con el formato IOB dependiendo si el término encontrado es perteneciente a un procedimiento, diagnóstico o alguna palabra fuera del contexto médico y signos de puntuación , se toma en cuenta el inicio de las frases, los saltos de línea entre otras cosas que se encuentran presentan en el texto médico del corpus, también se hace uso de las tuplas que ayudan en el proceso de etiquetar las palabras necesarias.

Figura 8

Algoritmo para colocar las etiquetas a las entidades médicas

```
def iobTagger(entities, tokenized):
    for token in tokenized:
        if cont == 0: ##solo si es el inicio de la frase
            startStringInit = 0
        #print("TOKEN: {}".format(token))
        if len(entities) == 0:
            state = "O"
        else:
            startStringCurrent = startStringInit + start_index(token) ##voy a trabajar con esto
            endCurrent = startStringInit + len(token) + 1
            #tupla o término médico en la tupla
            startTupla = entities[0][0]
            endTupla = entities[0][0] + entities[0][1]
            if startTupla == startStringCurrent:
                state = "B-"+ entities[0][2]
                sw = 1
            elif sw == 1 and startStringCurrent < endTupla:
                state = "I-"+ entities[0][2]
            elif sw == 1 and startStringCurrent >= endTupla:
                state = "O"
                sw = 0
            else:
                state = "O"
        tags.append(state)
```

En la Figura 9 se presenta un fragmento de código realizado para el modelo ML donde se visualiza una salida que contiene como columnas un identificador de las palabras, el término que está siendo etiquetado, todos los términos que se usan para el modelo están siendo tomado del dataset creado anteriormente, la etiqueta donde se observa que se encuentra con el formato IOB.

Figura 9

Fragmento del Código del modelo

```
In [2]: #Codificar Dataset
from sklearn.preprocessing import LabelEncoder
Palabra = LabelEncoder()
Resultado = LabelEncoder()
df["TerminoCodificado"] = Palabra.fit_transform(df["Termino"])
df["EtiquetaCodificada"] = Resultado.fit_transform(df["Etiqueta"])
df[['Termino','TerminoCodificado','Etiqueta','EtiquetaCodificada']]
```

Out[2]:

	Termino	TerminoCodificado	Etiqueta	EtiquetaCodificada
0	Describimos	2805	O	4
1	el	9655	O	4
2	caso	7022	O	4
3	de	8593	O	4
4	un	20080	O	4
...	...	...	...	...
166974	función	11286	O	4
166975	de	8593	O	4
166976	los	13716	O	4
166977	órganos	20728	O	4
166978	trasplantados.	19827	O	4

166979 rows × 4 columns

La Figura 10 representa como se visualiza la interfaz en la cual ya se encuentra el dataset junto con el modelo de Ml, se puede ver un campo de entrada donde se va a introducir el texto u oración que se va a ser procesada y luego etiquetada, el botón de “Enviar texto” es el que inicia el proceso de etiquetamiento, finalmente se encuentra la salida donde aparece el texto inicial y se imprime cuáles fueron los términos encontrados dentro del texto y mencionar si la palabra tiene una etiqueta de procedimientos o diagnósticos.

Figura 10

Interfaz del sistema el cual etiqueta las entidades de las oraciones introducidas

La Figura 11 muestra las pruebas que se realizaron con el modelo, para esto se ingresa una oración con relación al contexto médico para observa y revisa cómo se colocan las etiquetas en todas las palabras introducidas, se analiza si los términos tienen las etiquetas correspondientes y si se encuentra algún fallo también tenerlo en cuenta.

Figura 11

Resultado de las pruebas en el modelo

```
etiquetas , palabras = etiquetado(oracion)
for x,y in zip (etiquetas , palabras):
    print(y,"-",x)
```

```
Varón - O
de - O
25 - O
años - O
operado - O
de - O
varicocele - B-DIAGNOSTICO
izquierdo - O
hacia - O
seis - O
meses - O
, - O
que - O
consultó - O
por - O
un - O
problema - O
de - O
infecciones - B-DIAGNOSTICO
urinarias - I-DIAGNOSTICO
de - O
repetición - O
sin - O
patología - O
sexual - O
```

En la Figura 12 se presenta la evaluación de métricas para el modelo, se mostrará cuáles son los valores obtenidos en el accuracy, precisión, recall y F1, estas métricas sirven para ver qué tan adecuado es el modelo que se está usando y si el sistema está dando solución al problema planteado el inicio.

Figura 12

Evaluación de métricas del modelo

```
In [7]: #Entrenar modelo de ML (RANDOM FOREST)
        from sklearn.ensemble import RandomForestClassifier
        clf = RandomForestClassifier()
        clf.fit(x,y)

Out[7]: RandomForestClassifier()

In [8]: #Evaluación Random Forest
        y_pred = clf.predict(x)
        evaluar(y, y_pred)

        -Accuracy: 0.951233388629708 -Precision: 0.9450010795061511 -Recall: 0.951233388629708 -F1: 0.944529748889384

In [9]: #Entrenar Red neuronal
        from sklearn.neural_network import MLPClassifier
        mlp = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(5, 2), random_state=1)
        mlp.fit(x, y)

Out[9]: MLPClassifier(alpha=1e-05, hidden_layer_sizes=(5, 2), random_state=1,
                      solver='lbfgs')

In [10]: #Evaluar Red neuronal
         y_pred = mlp.predict(x)
         evaluar(y, y_pred)

         -Accuracy: 0.9145102078704508 -Precision: 0.9218187124288042 -Recall: 0.9145102078704508 -F1: 0.8736740257232903
```

Capítulo IV

Validación del sistema

En la validación del sistema se revisa el modelo y como usa el dataset, que se encuentra en la página oficial de huggingface, este modelo toma el conjunto de datos creado para realizar las pruebas correspondientes y revisar mediante una matriz de confusión cual es la eficiencia de este modelo con el dataset. Un punto para tener en cuenta para evaluar los modelos de predicción es aplicar las métricas de evaluación que son necesarias para verificar el funcionamiento del sistema.

Definición y aplicación de métricas de evaluación

Las matrices de confusión son uno de los métodos más utilizados para observar la precisión de algunos productos relacionados con la clasificación de datos. También conocida como matriz de error, es una tabla de contingencia utilizada como herramienta estadística para analizar observaciones por pares (López et al., 2018). Para el aprendizaje automático esta matriz es de gran ayuda para visualizar el desempeño de un algoritmo con respecto al aprendizaje supervisado.

La estructura de la matriz se refleja en que cada columna representa un número de predicciones según su clase y las filas representan las instancias de la clase. En la matriz se encuentran 4 resultados posibles como se muestra en la figura 5.

Figura 13

Matriz de confusión

- Verdadero positivo: se refiere a que el valor real es positivo y el valor de la prueba también predijo que es positivo.
- Verdadero negativo: se refiere a que el valor real es negativo y el valor de la prueba predijo que es negativo.
- Falso negativo: se refiere a que el valor real es positivo y el valor de la prueba predijo que es negativo.
- Falso positivo: se refiere a que el valor real es negativo y el valor de la prueba predijo que es positivo.

Tomando en cuenta estas opciones de la matriz de confusión hay que revisar las métricas que son la exactitud, precisión, la sensibilidad y la especificación. Los problemas de machine learning se enfrentan a la decisión de saber cuál es el modelo más adecuado para la solución. La métrica más usada y común es la exactitud (o accuracy en inglés), esta métrica se define como la proporción entre los datos que se clasifican correctamente con respecto al conjunto de datos que existen totalmente.

Si se requiere usar un modelo para determinar si un texto contiene términos médicos, el modelo debe determinar si la palabra es un procedimiento o diagnóstico. Y este problema se resolverá entrenando modelos como una red neuronal, bosque aleatorio o red convolucional. Sea cual sea la elección lo que se debe observar es el sesgo que tendrá el conjunto de datos al momento de clasificar, y si lo hace bien o no.

La matriz de confusión como todo método también tiene sus limitaciones ya que en ocasiones se encuentran dificultades al determinar las ventajas de un modelo u otro, ya que se necesitaría de más información para evaluar el desempeño de los modelos.

A partir de las opciones anteriores de los verdaderos positivos, verdaderos negativos, falsos negativos y falsos positivos, se debe profundizar las métricas que son la exactitud, precisión, la sensibilidad y la especificación.

- La exactitud (Accuracy): muestra qué tan cerca está el resultado de la medida real del valor. Las estadísticas muestran que la precisión está relacionada con el sesgo de estimación. La exactitud es el número de predicciones positivas correctas.

  Se calcula de la siguiente manera:

  (VP+VN) / (VP+FP+FN+VN)

  En donde:

  VP son verdaderos positivos.

  VN son verdaderos negativos.

  FP son falsos positivos.

  FN son falsos negativos.

  Otra métrica es la precisión. Se refiere a la dispersión de un conjunto de valores obtenidos a partir de mediciones repetidas. Este es el porcentaje de casos positivos detectados.

  Su fórmula es la siguiente:

  VP/(VP+FP)

- La sensibilidad (en inglés recall o sensitivity): representa la fracción de los verdaderos positivos.
- La especificidad (especificity): representa la fracción de los verdaderos negativos.
- F1 score: es el resumen de la precisión y la sensibilidad, es muy útil cuando la distribución de las clases es desigual.

  Se calcula de la siguiente manera:

  PuntajeF1= (2*Precisión*Sensibilidad) / (Precisión + Sensibilidad)

Se realizaron las pruebas al modelo de Machine Learning con los datos de entrenamiento la cual se muestra en la Tabla 17.

Tabla 17

Métricas de precisión

Accuracy	Precisión	F1	Recall
0.9145102078704508	0.9218187124288042	0.8736740257232903	0.9145102078704508

Corrección de errores y ajuste de modelos

En los errores se menciona que algunas palabras no se encontraban, por lo cual se coloca más datos en el dataset esto para que el modelo pueda tener mucha más información y luego volver a entrenar el modelo con los nuevos datos para obtener en las métricas de evaluación nuevos resultados favorables y cuando se haga las pruebas se evidencie que las oraciones ingresadas puedan etiquetar las entidades de mejor manera, identificando los procedimientos y diagnósticos correspondientes.

Se recomienda usar datos del corpus de las carpetas no utilizadas ya que contienen términos médicos los cuales volviendo a hacer el proceso de etiquetamiento en los procesos y diagnósticos se pueden unir a los datos actuales en el dataset, de esta manera se consigue un mayor conjunto de datos para usar en el modelo.

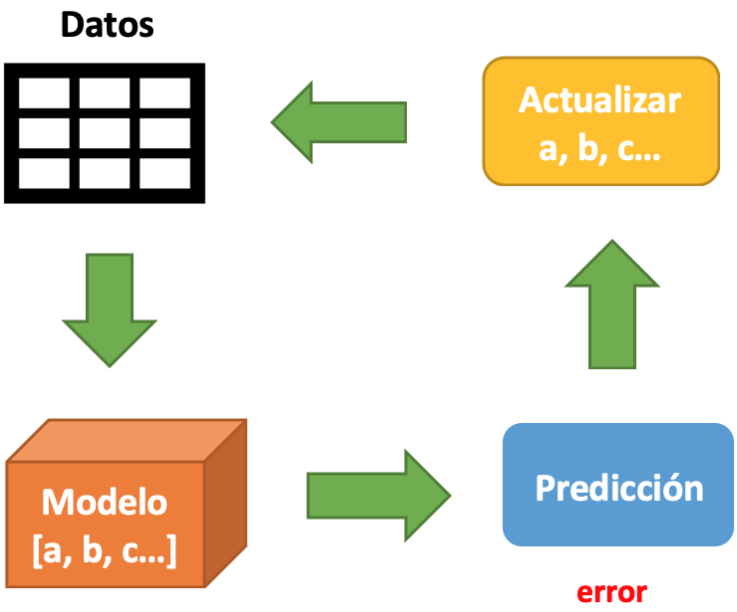
Al trabajar con un conjunto de datos médicos como notas clínicas, registros y descripciones de pacientes el modelo debe contener gran cantidad de información, asegurándose que los errores se reduzcan en lo máximo posible ya que el modelo es la salida de información de nuestro algoritmo.

El aprendizaje automático permite que los modelos se entrenen en conjuntos de datos antes de la implementación. Algunos modelos de aprendizaje automático son en línea y continuos. Este proceso iterativo del modelo en línea mejora los tipos de asociaciones entre elementos de datos. Debido a su complejidad y tamaño, estos patrones y asociaciones podrían haber escapado fácilmente a la observación humana. Una vez que se entrena un modelo, se puede usar para aprender de los datos en

tiempo real. La precisión mejorada es el resultado del proceso de capacitación y la automatización que forma parte del aprendizaje automático, en la Figura 14 se muestra el proceso de un ciclo de vida de un ML.

Figura 14

Ciclo de vida de un modelo de machine learning



Análisis de resultados

Los datos utilizados y convertidos al formato correspondiente se encuentran dentro de la página de Huggingface como un conjunto de datos públicos, esto permite que el modelo usado dentro de paperspace pueda encontrar el conjunto de datos y trabajar de manera correcta.

La construcción del algoritmo se enfoca en colocar el texto al estándar IOB el cual permita manejar el texto de forma fácil y luego el conjunto de datos que se obtuvo se lo pone a disposición en la página de Huggingface como un repositorio.

Para el modelo se observa el desempeño de las redes Transformes y el manejo de los datos principalmente en el procesamiento de lenguaje natural como la predicción de texto, con el modelo

usado se consiguen los siguientes datos en las métricas accuracy de 0.91, precisión de 0.92, F1 de 0.87 y recall de 0.91.

La interfaz web ayuda a facilitar el uso de la herramienta con tecnologías para el desarrollo web como HTML, CSS, FLASK, aplicando buenas prácticas para la construcción de sistemas web que sea escalable y modificable, de esta manera se muestra el modelo de forma amigable al usuario.

Capítulo V

Conclusiones y Recomendaciones

Conclusiones

- Se cumplió con el objetivo de desarrollar un sistema de predicción de diagnósticos médicos en base a notas clínicas de los pacientes, aplicando técnicas y modelos de aprendizaje automático.
- El desarrollo del marco teórico permitió la obtención de conocimientos acerca de los modelos de aprendizaje automático, técnicas y herramientas adecuadas para el desarrollo del sistema.
- El desarrollo del algoritmo para convertir el texto plano a un archivo xml ayudó a tener la información de mejor manera, ya que esto permitió un mejor manejo de datos para poder etiquetar las entidades.
- El algoritmo que se utilizó para etiquetar las entidades hizo uso del archivo xml creado anteriormente con el cual se toma los datos para identificar las entidades médicas de acuerdo si son procedimientos o diagnósticos.
- El uso de métricas de evaluación ayudó al análisis de resultados permitiendo validar la precisión del modelo y evidenciando si se encuentra las entidades de diagnósticos y procedimientos.
- Al utilizar paperspace como entorno se consiguió un mejor rendimiento con respecto al tiempo de ejecución, reduciendo el tiempo que tomaba entrenar el modelo con gran cantidad de datos.
- Al utilizar el modelo implementado con el dataset se evidencio que las oraciones ingresadas etiquetaban las entidades médicas encontradas sean de diagnósticos o procedimientos.

Recomendaciones

- Es esencial contar con un conjunto de datos representativo y adecuado para entrenar y validar el modelo de aprendizaje automático.
- El uso de un conjunto de datos incompleto o desequilibrado puede generar un modelo ineficaz.
- Se deben explorar varias técnicas de aprendizaje automático, como árboles de decisión, redes neuronales, regresión logística para encontrar la mejor solución al problema.
- La validación del modelo con el uso de métricas de rendimiento adecuadas, tales como la precisión, el recall y la F1 score, es una etapa crítica para medir el desempeño del modelo y determinar si se deben realizar ajustes y mejoras.
- La implementación de una interfaz de usuario amigable es esencial para permitir que los médicos utilicen el sistema con facilidad.
- Es importante aplicar técnicas de preprocesamiento de datos para normalizar y limpiar los datos antes de introducirlos en el modelo.
- Realizar una evaluación continua del sistema es fundamental para detectar posibles errores y actualizar el modelo con nuevos datos.

Bibliografía

Aguirre Ascona, Y. D. (2019). Métodos de aprendizaje supervisado para la predicción de diabetes: Una revisión sistemática de la literatura. *Universidad Peruana Unión*. https://repositorio.upeu.edu.pe/handle/20.500.12840/2511

Alfaro, A. D., & Ospina, J. V. D. (2021). *Revisión sistemática de literatura: Técnicas de aprendizaje automático (Machine Learning) | Cuaderno activa*. https://ojs.tdea.edu.co/index.php/cuadernoactiva/article/view/849

Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica. (s. f.). Recuperado 10 de febrero de 2023, de https://revistas.uis.edu.co/visores/Revista_UIS_Ingenierias_Vol_19_Num_4/553768213002/

Arias, V., Salazar, J., & Garicano, C. (2019). Una introducción a las aplicaciones de la inteligencia artificial en Medicina: Aspectos históricos. *Revista Latinoamericana de Hipertensión*, *14*.

Ariza López, F. J., Rodríguez Avi, J., & Alba Fernández, M. V. (2018). Control estricto de matrices de confusión por medio de distribuciones multinomiales. *Geofocus: Revista Internacional de Ciencia y Tecnología de la Información Geográfica*, *21*, 6.

CESAR, P. L., & DANIEL, S. G. (2007). *Minería de datos. Técnicas y herramientas: Técnicas y herramientas*. Ediciones Paraninfo, S.A.

Cruz, I., Martinez, S. S., & Abed, A. R. (s. f.). *Redes neuronales recurrentes para el an8lisis de secuencias*.

Elgueta Morales, J. A.- jorelgue@gmail com. (2017). *Comparación de rendimiento de técnicas de aprendizaje automático para análisis de afecto sobre textos en español*. http://repobib.ubiobio.cl/jspui/handle/123456789/1772

Escrivá, J. V. S., Peyró, C. F., Vayá, M. de la I., Montell, J. A., & Fabra, M. J. E. (2020). Aplicación de la Inteligencia Artificial con Procesamiento del Lenguaje Natural para textos de investigación

cualitativa en la relación médico-paciente con enfermedad mental mediante el uso de tecnologías móviles. *Revista de Comunicación y Salud*, *10*(1), Art. 1. https://doi.org/10.35669/rcys.2020.10(1).19-41

Fabregat Marcos, H. (2021). *Biomedical Information Extraction: Exploring new entities and relationships*. http://e-spacio.uned.es/fez/view/tesisuned:ED-Pg-SisInt-Hfabregat

Frayre, C. D. A., & Martínez, L. F. F. (2022). Redes transformers: Generación de consultas SQL a través de lenguaje natural: 4CP22-22. *Memorias Científicas y Tecnológicas*, *1*(1), Art. 1.

Gonzales Balcázar, M. A. (2022). *Uso de la metodología aula invertida para mejorar el nivel de conocimiento sobre algoritmos secuenciales en estuiantes de ingeniería de sistemas de la Universidad Tecnológica del Perú, 2019*.

González, H. A. B. (2017). Aplicación del análisis de regresión lineal simple para la estimación de los precios de las acciones de Facebook, Inc. *REICE: Revista Electrónica de Investigación en Ciencias Económicas*, *5*(10), Art. 10. https://doi.org/10.5377/reice.v5i10.5535

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 900-903. https://doi.org/10.1109/UKRCON.2017.8100379

Guardiola González, C. (2020). *Clasificador de textos mediante técnicas de aprendizaje automático* [Proyecto/Trabajo fin de carrera/grado, Universitat Politècnica de València]. https://riunet.upv.es/handle/10251/133840

Hema, V., Thota, S., Kumar, S. N., Padmaja, C., Krishna, C. B. R., & Mahender, K. (2020). Scrum: An Effective Software Development Agile Tool. *IOP Conference Series: Materials Science and Engineering*, *981*(2), 022060. https://doi.org/10.1088/1757-899X/981/2/022060

Jara, F. A., & Lobato, D. H. (2018). *APRENDIZAJE NO-SUPERVISADO CON MODELOS GENERATIVOS PROFUNDOS*.

Kayes, I., Sarker, M., & Chakareski, J. (2016). Product backlog rating: A case study on measuring test quality in scrum. *Innovations in Systems and Software Engineering*, *12*(4), 303-317. https://doi.org/10.1007/s11334-016-0271-0

Kuz, A., Falco, M., & Giandini, R. S. (2018). Comprendiendo la Aplicabilidad de Scrum en el Aula: Herramientas y Ejemplos. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, *21*, Art. 21. https://doi.org/10.24215/18509959.21.e07

Lalaleo Achachi, D. F. (2021). *Diseño de un algoritmo utilizando Machine Learning para la predicción de la radiación solar en el sector de Lasso.* [MasterThesis, Ecuador: Latacunga: Universidad Técnica de Cotopaxi: UTC.]. http://repositorio.utc.edu.ec/handle/27000/8014

Lemus-Delgado, D., & Navarro, R. P. (2020). Ciencia de datos y estudios globales: Aportaciones y desafíos metodológicos. *Colombia Internacional*. https://doi.org/10.7440/colombiaint102.2020.03

Luna, M., & Lorenzo, G. (2011). Minería de datos: Cómo hallar una aguja en un pajar. *Ciencia - Academia Mexicana de Ciencias*, *62*(3), Art. 3.

Meseguer Esbri, P. (2022). *Diseño y desarrollo de un sistema automático para la predicción de la colitis ulcerosa aplicando técnicas de aprendizaje profundo débilmente supervisado sobre Whole-Slide Images*. https://riunet.upv.es/handle/10251/188277

Pérez Guerrero, J. (2020). *Redes recurrentes*. https://idus.us.es/handle/11441/115230

Pérez Ortiz de Landaluce, M. (2021). *Clasificación de imágenes mediante algoritmos de Deep Learning: Mascarillas de COVID-19*. https://idus.us.es/handle/11441/127041

Pineda, J. M. (2022). Modelos predictivos en salud basados en aprendizaje de maquina (machine learning). *Revista Médica Clínica Las Condes*, *33*(6), 583-590. https://doi.org/10.1016/j.rmclc.2022.11.002

Pisner, D. A., & Schnyer, D. M. (2020). Chapter 6—Support vector machine. En A. Mechelli & S. Vieira (Eds.), *Machine Learning* (pp. 101-121). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Rad, N. K., & Turley, F. (2019). *Los Fundamentos de Agile Scrum*. Van Haren.

Ramírez, P. E., Grandón, E. E., Ramírez, P. E., & Grandón, E. E. (2018). Predicción de la Deserción Académica en una Universidad Pública Chilena a través de la Clasificación basada en Árboles de Decisión con Parámetros Optimizados. *Formación universitaria*, *11*(3), 3-10. https://doi.org/10.4067/S0718-50062018000300003

Repetur, A. E. (s. f.). *Redes Neuronales Artificiales*.

Riquelme Santos, J. C., Ruiz, R., & Gilbert, K. (2006). *Minería de Datos: Conceptos y Tendencias*. https://idus.us.es/handle/11441/43290

Rodríguez, C., & Dorado Vicente, R. (2015). ¿Por qué implementar Scrum? *Revista ONTARE*, *3*(1), 125-144.

Rodríguez Suárez, Y., Suárez, Y. R., & Amador, A. D. (2011). Herramientas de minería de datos. *Revista Cubana de Ciencias Informáticas*, *3*(3-4). https://rcci.uci.cu/?journal=rcci&page=article&op=view&path[]=78

Samy, D. (2021). *Reconocimiento y clasificación de entidades nombradas en textos legales en español*. https://doi.org/10.26342/2021-67-9

Sarmiento-Ramos, J. L. (2020). Aplicaciones de las redes neuronales y el deep learning a la ingeniería biomédica. *Revista UIS Ingenierías*, *19*(4), 1-18. https://doi.org/10.18273/revuin.v19n4-2020001

Sedano, T., Ralph, P., & Péraire, C. (2019). The Product Backlog. *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*, 200-211. https://doi.org/10.1109/ICSE.2019.00036

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, *134*, 93-101. https://doi.org/10.1016/j.eswa.2019.05.028

Valero Moreno, A. I. (2017). *Técnicas estadísticas en minería de textos*. https://idus.us.es/handle/11441/63197

Vásquez, A. C., Quispe, J. P., & Huayna, A. M. (s. f.). *Procesamiento de lenguaje natural | Revista de investigación de Sistemas e Informática*. Recuperado 10 de febrero de 2023, de https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Villena, F., Dunstan, J., Villena, F., & Dunstan, J. (2019). Obtención automática de palabras clave en textos clínicos: Una aplicación de procesamiento del lenguaje natural a datos masivos de sospecha diagnóstica en Chile. *Revista médica de Chile*, *147*(10), 1229-1238. https://doi.org/10.4067/s0034-98872019001001229

Anexos