

Escuela Superior Politécnica del Litoral

Facultad de Ciencias Naturales y Matemáticas

Modelo automático de codificación basado en CIE-10 de diagnósticos médicos
usando modelos de machine learning

MATE-208

Proyecto integrador

Previo a la obtención del Título de:

Ingeniero en Estadística

Presentado por:

Miguel Alfonso Flores Boada

GUAYAQUIL – ECUADOR

Año: 2024

Dedicatoria

Dedico este proyecto, principalmente a mi madre, Catalina y a mi padre, Julio, que han sido los pilares fundamentales para este logro, con su trabajo diario, que me ha inspirado a centrarme, teniendo constancia y disciplina. Su ayuda ha sido primordial en esta etapa de mi vida.

También quisiera dedicar este trabajo a mis abuelas, ya que han sabido enseñarme la fortaleza personal y, de manera especial, por mostrarme un estilo de vida que he adoptado como propio. Esto va para que sus nombres no caigan en el olvido, Francia y Marina.

Miguel Flores Boada

Agradecimientos

Quisiera presentar un agradecimiento a Dios y a mis padres, por su ayuda y apoyo en este proceso. De igual forma a mis hermanos, Sofía y Sebastián, que de una u otra forma me han brindado soporte.

A mis profesores, por impartir una educación de calidad, con calidez. A mi Universidad, por acogerme y por ser como un hogar en una ciudad diferente a la mía.

A Anahí, por ser alguien con quien puedo contar siempre y por su respaldo en los momentos complicados de mi carrera.

A mis amigos de la Universidad, por hacer más llevadera esta experiencia.

En general agradezco a todos los que han aportado en algo durante este proceso.

Miguel Flores Boada

Declaración Expresa

Yo Miguel Alfonso Flores Boada acuerdo y reconozco que:

La titularidad de los derechos patrimoniales de autor (derechos de autor) del proyecto de graduación corresponderá al autor o autores, sin perjuicio de lo cual la ESPOL recibe en este acto una licencia gratuita de plazo indefinido para el uso no comercial y comercial de la obra con facultad de sublicenciar, incluyendo la autorización para su divulgación, así como para la creación y uso de obras derivadas. En el caso de usos comerciales se respetará el porcentaje de participación en beneficios que corresponda a favor del autor o autores.

La titularidad total y exclusiva sobre los derechos patrimoniales de patente de invención, modelo de utilidad, diseño industrial, secreto industrial, software o información no divulgada que corresponda o pueda corresponder respecto de cualquier investigación, desarrollo tecnológico o invención realizada por mí durante el desarrollo del proyecto de graduación, pertenecerán de forma total, exclusiva e indivisible a la ESPOL, sin perjuicio del porcentaje que me/nos corresponda de los beneficios económicos que la ESPOL reciba por la explotación de mi/nuestra innovación, de ser el caso.

En los casos donde la Oficina de Transferencia de Resultados de Investigación (OTRI) de la ESPOL comunique al autor que existe una innovación potencialmente patentable sobre los resultados del proyecto de graduación, no se realizará publicación o divulgación alguna, sin la autorización expresa y previa de la ESPOL.

Guayaquil, 10 de octubre del 2024.



Miguel Alfonso Flores

Boada

Evaluadores

Sandra Lorena García Bustos

Profesora de Materia Integradora

Omar Honorio Ruiz Barzola

Tutor de proyecto

Resumen

La innovación tecnológica ha alcanzado a todos los ámbitos de la humanidad, la estadística y el análisis de datos son una parte fundamental de este avance. El campo de la medicina es un beneficiario importante del progreso tecnológico. Este proyecto tiene como objetivo clasificar automáticamente enfermedades con código CIE-10, a través de modelos estadísticos de aprendizaje y clasificación, contribuyendo de esta forma a la Red de Salud Pública Nacional. Existen diferentes formas en las que se codifica cada diagnóstico en el Ecuador. Por el momento se usa un algoritmo determinista que funciona a través de las reglas implementadas en el manual de la CIE-10. Los datos fueron obtenidos del registro estadístico de camas y egresos hospitalarios. Se utilizó Excel, Rstudio y Python para el procesamiento de datos y creación de los modelos estadísticos. La red neuronal implementada tuvo una precisión de entre el 80% y 86%, en menos de 15 minutos de codificación automática. Los modelos desarrollados permiten reducir el tiempo de codificación manual, empleando los recursos tecnológicos de mejor manera en una institución estatal.

Palabras clave: automático, clasificación, diagnósticos, redes neuronales.

Abstract

Technological innovation has reached every aspect of humanity, with statistics and data analysis playing a fundamental role in this progress. The field of medicine is a significant beneficiary of technological advancements. This project aims to automatically classify diseases using the ICD-10 coding system through statistical learning and classification models, thus contributing to the National Public Health Network. In Ecuador, there are different methods for coding each diagnosis. Currently, a deterministic algorithm is used, based on the rules outlined in the ICD-10 manual. The data was obtained from the statistical registry of hospital beds and discharges. Excel, Rstudio and Python were utilized for data processing and model development. The implemented neural network achieved an accuracy of 80% to 86% in less than 15 minutes of automatic coding. The developed models significantly reduce the time required for manual coding, optimizing the use of technological resources in a state institution.

Keywords: *automatic, classification, diagnoses, neural networks.*

Índice General

Resumen.....	I
Abstract.....	II
Índice General.....	III
Abreviaturas.....	VI
ÍNDICE DE FIGURAS.....	VII
Índice de tablas	VIII
CAPÍTULO I: INTRODUCCIÓN.....	1
1.1 Descripción del problema:	3
1.2 Justificación del problema:	4
1.3 Objetivos:.....	5
1.3.1 Objetivo General:.....	5
1.3.2 Objetivos Específicos:	5
1.4 Marco Teórico:.....	5
1.4.1 Registro estadístico de camas y egresos	5
1.4.2 Codificación.....	6
1.4.3 Clasificación Internacional de enfermedades (CIE-10).....	6
1.4.4 Diagnóstico médico	6
1.4.5 Inteligencia artificial (IA)	6
1.4.6 Procesamiento de Lenguaje Natural (PLN)	7
1.4.7 Evaluación de modelos	7

1.4.8 Random Forest	8
1.4.9 Redes neuronales	8
1.4.10 Error Out-of-bag	9
1.4.11 One-hot-encoding	9
1.4.11 Estado del arte.....	9
CAPÍTULO II: METODOLOGÍA	11
2.1. Diseño del estudio.....	12
2.2. Población y muestra.....	12
2.3. Técnicas de recolección de datos	13
2.3.1. Instrumentos utilizados	13
2.4 Variables de estudio.....	14
2.5 Procedimiento de análisis de datos	15
2.5.1 Procesamiento inicial	15
2.5.2 Análisis estadístico.....	16
2.5.3 Validación de resultados	17
2.6 Limitaciones del estudio	17
CAPÍTULO III: RESULTADOS.....	18
3.1 Análisis exploratorio.....	19
3.1.1 Análisis de características demográficas	19
3.1.2 Análisis de los diagnósticos	21
3.1.3 Análisis de características del texto	23

3.2 Modelo de Random Forest.....	24
3.2.1 Optimización.....	24
3.2.2 Evaluación del modelo.....	25
3.3 Red Neuronal	28
3.3.1 Optimización del modelo	28
3.3.2 Evaluación del modelo.....	28
3.4 Comparación entre los modelos.....	30
3.4.1 Comparación de métricas.....	30
3.4.2 Capacidad predictora	31
3.5 Discusión.....	31
CAPÍTULO IV: CONCLUSIONES Y RECOMENDACIONES	34
4.1 Conclusiones	35
4.2 Recomendaciones	35
BIBLIOGRAFÍA	37
ANEXOS	39

Abreviaturas

CIE-10:	Clasificación internacional de enfermedades – tomo 10
ESPOL:	Escuela Superior Politécnica del Litoral
IA:	Inteligencia artificial
INEC:	Instituto Nacional de Estadística y Censos
OPS:	Organización Panamericana de la salud
PLN:	Procesamiento de lenguaje natural
RC:	Red complementaria
RPIS:	Red Pública Integral de Salud

ÍNDICE DE FIGURAS

Figura 1: Funcionamiento de red neuronal	8
Figura 2: Distribución de pacientes por sexo en la sección de traumatismos	19
Figura 3: Distribución de edades de pacientes en la sección de traumatismos.	20
Figura 4: Distribución de edades de pacientes en la sección de obstetricia	21
Figura 5: Diagnósticos más frecuentes en la sección de traumatismos	22
Figura 6: Diagnósticos más frecuentes en la sección de obstetricia	22
Figura 7: Distribución de la longitud de palabras en la sección de traumatismos	23
Figura 8: Distribución de la longitud de palabras en la sección de obstetricia.....	24
Figura 9: Comparación de métricas de traumatismos.....	25
Figura 10: Error out-of-bag en la sección de traumatismos.....	26
Figura 11: Comparación de métricas de obstetricia.....	27
Figura 12: Error out-of-bag en la sección de obstetricia.....	27
Figura 13: Comparación de métricas de entrenamiento vs prueba de la red neuronal	29
Figura 14: Curva de aprendizaje de la red neuronal	30

Índice de tablas

Tabla 1: Variables utilizadas.....	14
Tabla 2: Resumen de métricas de entrenamiento	30
Tabla 3: Resumen de métricas de prueba	31

CAPÍTULO I: INTRODUCCIÓN

Con el paso del tiempo, la expansión tecnológica ha aportado de gran manera a la automatización de actividades cotidianas en todos los ámbitos, desde la ingeniería o matemáticas hasta ciencias sociales o medicina, precisamente es en esta última en la que los avances han llegado a puntos sorprendentes, ya sea nuevas formas de elaborar prótesis adecuadas para la implementación en humanos, hasta vacunas creadas en tiempo récord, visto en la pandemia de la COVID-19. Ninguno de estos aportes a la humanidad sería capaz sin la captación, almacenamiento, tratamiento y presentación de grandes volúmenes de datos.

Con la creciente cantidad de datos generados en el área de salud, se precisa necesario encontrar métodos más eficientes para resolver, examinar y usar este tipo de información. Bajo este contexto, los diagnósticos médicos y su codificación bajo estándares internacionales, como el sistema de la Clasificación Internacional de Enfermedades, por sus siglas CIE-10, son fundamentales para cumplir este propósito. Esta actividad, que generalmente ha sido realizada de forma manual, no solo consume tiempo de los profesionales médicos, sino que también es propensa a errores humanos por el desconocimiento de la extensa normativa de esta clasificación, impacta negativamente en la calidad de la gestión médica y la atención sanitaria.

La automatización de la codificación CIE-10 mediante procesos tecnológicos avanzados, como el aprendizaje de máquinas (machine learning) y el procesamiento de lenguaje natural (PLN), simboliza una solución promisorio a estos retos. Estos métodos permiten analizar descripciones de diagnósticos médicos y asignar códigos con rapidez y precisión, optimizando el creciente trabajo en instituciones de salud en el Ecuador y el mundo. Asimismo, esta tecnología puede contribuir significativamente a elevar la calidad de los registros médicos, los cuales son esenciales para la investigación, el diseño de políticas públicas y la administración en el ámbito hospitalario.

No obstante, el desarrollo de modelos predictivos para esta labor enfrenta varios desafíos, desde la necesidad de datos de calidad alta y bien estructurados, hasta la forma en que

los algoritmos comprenden las expresiones técnicas y las claves características del CIE-10. En este contexto, la investigación en esta área no solo se enfoca en desarrollar modelos operativos, sino también en analizar cómo estas herramientas pueden ajustarse a diversos entornos médicos, lenguajes y sistemas de salud, optimizando su aplicación a nivel mundial. Es importante señalar que el contexto ecuatoriano influye de manera significativa, es decir, la investigación estará enfocada en encontrar soluciones y sobrepasar desafíos para el contexto local.

1.1 Descripción del problema:

La calidad de la información es deficiente en el ámbito médico, se puede tener una simple idea con la mala caligrafía en las recetas médicas e historias clínicas, lo que dificulta la captación de datos, si es que un paciente no entiende una receta, sería difícil seguirla al pie de la letra. En el ámbito hospitalario pasa de una manera similar, cuando los médicos escriben de forma manual los diagnósticos en las historias clínicas, muchas veces cometen errores ya sea ortográficos o caligráficos por la velocidad en la que se escribe.

La mejor forma de afrontar esta problemática sería digitalizando toda la información, pero también se puede tener errores al momento de la digitación, esto es un problema al momento de realizar comparaciones de estadísticas con otros países para conocer la realidad del país y la salud de sus habitantes. Dado que la estandarización siempre ha sido un problema persistente por la pluriculturalidad mundial, la forma más fácil de homogeneizar la información es mediante códigos, esto se ha visto por ejemplo en la química con la tabla periódica; o en las matemáticas con las fórmulas basadas en letras.

En el Ecuador el organismo público encargado de generar estadísticas oficiales es el Instituto Nacional de Estadística y Censos, por sus siglas INEC. En la mayoría de sus investigaciones, para poder sintetizar y estandarizar los datos, se usan codificaciones, existen

de 3 tipos: manuales, asistida por computador y automática. En gran parte de los casos se utiliza una codificación manual asistida por computador. (Moreno, 2023)

El registro estadístico de camas y egresos hospitalarios (INEC, 2023), recaba información sobre la morbilidad hospitalaria en la Red Pública Integral de Salud y Red Complementaria, es decir en todos los entes hospitalarios del Ecuador que tengan servicio de internación, ya sean hospitales, clínicas, etc. Dejando de lado los centros de salud del país. En este registro se levantan datos sobre los diagnósticos médicos y uno de los trabajos del INEC es codificar cada diagnóstico para realizar las respectivas comparaciones con otros países, mediante una codificación aceptada a nivel mundial llamada Clasificación Internacional de Enfermedades (CIE-10), el problema es simple: no existe, todavía, una forma eficiente de hacerlo de manera automática.

1.2 Justificación del problema:

Existen diferentes formas en las que el INEC se las ha ingeniado para codificar lo mejor posible cada diagnóstico. Por el momento se usa un algoritmo determinista que funciona a través de las reglas implementadas en el manual de la CIE-10, funciona limpiando el texto de modo “manual” para que el algoritmo asigne un código de manera semiautomática. Al principio de la aplicación de este sistema, la correcta codificación fue de un impresionante 70% y con el paso de los años fue mejorando hasta casi alcanzar una cota superior cercana al 90% y ahora por más cambios que se realicen no se puede superar este valor.

El tiempo ocupado para esto es otro indicador para tomar en cuenta, la creación de este algoritmo determinista ha tomado mucho tiempo, lo mismo con su perfeccionamiento y si bien este proceso ha mejorado mucho, su trabajo necesita demasiados recursos temporales del personal del INEC.

El aprendizaje de máquinas para resolver problemas relacionados con la clasificación ha tenido varios éxitos, bajo estas referencias se ha propuesto en este trabajo realizar 2 modelos

de clasificación, el primero, uno de random forest para saber cómo se catalogan los códigos bajos este modelo; el segundo, una red neuronal que buscará adaptar el poder computacional a los requerimientos necesarios.

1.3 Objetivos:

1.3.1 Objetivo General:

Clasificar automáticamente enfermedades con código CIE-10, a través de modelos estadísticos de aprendizaje y clasificación, contribuyendo de esta forma a la Red de Salud Pública Nacional.

1.3.2 Objetivos Específicos:

Diseñar un modelo de clasificación para codificar automáticamente diagnósticos médicos mediante CIE-10, usando random forest.

Diseñar un modelo de clasificación para codificar automáticamente diagnósticos médicos mediante CIE-10, usando redes neuronales.

Comparar la efectividad entre diferentes algoritmos para procesamiento de lenguaje

1.4 Marco Teórico:

1.4.1 Registro estadístico de camas y egresos

Este informe presenta información sobre morbilidad hospitalaria, es decir, la frecuencia de aparición de alguna enfermedad con referencia a la población de cierto sector. La información se origina en los establecimientos de la Red Pública Integral de Salud y de la Red complementaria, en establecimientos que cuentan con internación hospitalaria. En el Ecuador, este registro es la única fuente disponible de datos de morbilidad. En el 2023 se reportó información procedente de 627 establecimientos. (INEC, 2023)

1.4.2 Codificación

(Moreno, 2023) establece que en el Ecuador y otros países se utilizan 3 tipos de codificaciones: la manual, que es realizada por personas a las que se las ha entrenado para esta actividad, esto es costoso en términos de tiempo por el gran volumen de datos; la codificación asistida, en esta existe un software que ayuda al codificador a escoger la codificación que mejor se adapte al trabajo realizado; por último está la codificación automática, aquí se asignan los códigos mediante un programa de computación si necesidad de la intervención humana, se reduce considerablemente el tiempo y los errores humanos.

1.4.3 Clasificación Internacional de enfermedades (CIE-10)

La OPS, (Organización Panamericana de la Salud, 2015) establece una forma de clasificar mediante códigos a los diagnósticos médicos, esta clasificación fue aprobada en la Asamblea Mundial de la Salud en 1989. Esta categorización pretende estandarizar a las enfermedades para ayudar a médicos a sobrepasar fronteras de idiomas y apreciaciones subjetivas al momento de generar información a través de estadísticas, el manual se conforma por una serie de reglas para la correcta separación de diagnósticos.

1.4.4 Diagnóstico médico

Para (Díaz Novás, Gallego, & León, 2006), “el diagnóstico médico es una de las tareas fundamentales de los médicos y la base para una terapéutica eficaz.” Establece también que es un medio indispensable para establecer el tratamiento más adecuado para cada paciente. El diagnóstico se basa en el análisis de datos seguros, no tiene validez si es que no se basa en hechos precisos y nociones exactas.

1.4.5 Inteligencia artificial (IA)

(Rouhiainen, 2018) instituye varias pautas para entender de una mejor manera la inteligencia artificial. Manifiesta que dar una definición de IA es complicado porque es un tema complejo, pero intenta dar una visión general cuando indica que la inteligencia artificial es “la

habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana”, como generar algoritmos, aprender de datos y tomar decisiones tal y como lo haría un ser humano.

1.4.6 Procesamiento de Lenguaje Natural (PLN)

“Por PLN se entiende a la habilidad de la máquina para procesar la información comunicada, no simplemente las letras o los sonidos del lenguaje.” (Gelbukh, 2010) Tomando en cuenta el siguiente ejemplo: un programa que busque palabras en un documento no sería un PLN, pero si en vez de buscar literalmente por palabra, buscaría una idea de las palabras escritas, eso sí es PLN. Un programa que se usa mucho en la actualidad es ChatGPT. El procesamiento de lenguaje natural es una forma de inteligencia artificial en la rama de la lingüística computacional.

1.4.7 Evaluación de modelos

Para evaluar la eficacia de los modelos propuestos, después del entrenamiento y la prueba, existen varias métricas, las más aceptadas son estas 5: la precisión, exactitud, la sensibilidad, el F1 score y el índice Kappa. (Moreno, 2023) Otro concepto útil es la matriz de confusión que contiene 4 elementos: verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN).

1.4.7.1 Precisión. Mide qué tan bien el modelo clasifica las observaciones. Es el porcentaje de predicciones correctas sobre el total de predicciones.

1.4.7.2 Exactitud. Esta métrica mide la relación de verdaderos positivos entre todas las observaciones verdaderas.

1.4.7.3 Sensibilidad. Mide la proporción de verdaderos positivos entre todas las observaciones positivas.

1.4.7.4 F1 Score. Combina la precisión y sensibilidad en una sola métrica.

1.4.7.5 índice Kappa. Mide el nivel de concordancia entre las predicciones del modelo y las observaciones reales. (Moreno, 2023)

1.4.8 Random Forest

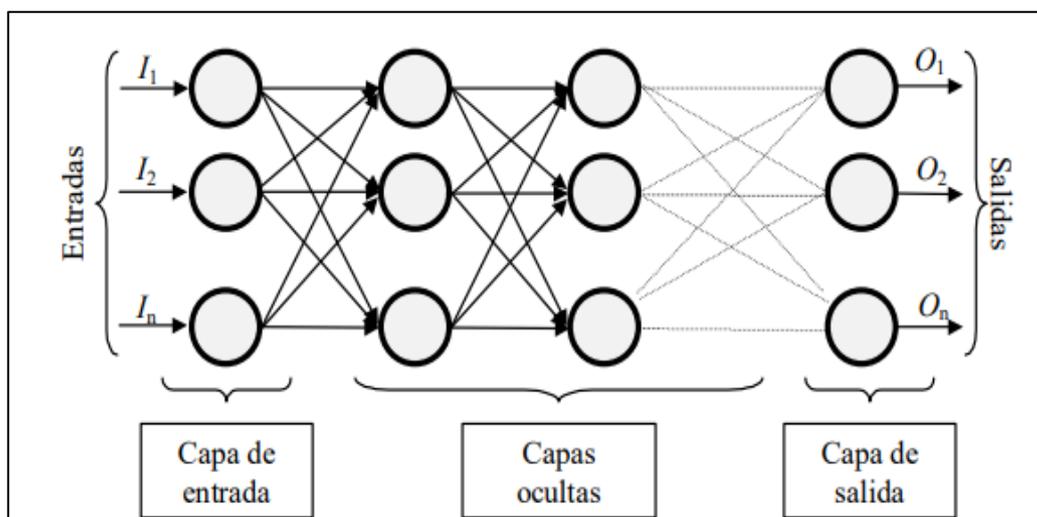
Los árboles de decisión proporcionan son herramientas de clasificación potentes que son fáciles de entender, prácticamente brindan 6 opciones: clasificación, segmentación, predicción, reducción, identificación-interrelación y recodificación. Random forest es un algoritmo de machine learning que se compone de varios árboles de decisión. (Bouza & Santiago, 2012)

1.4.9 Redes neuronales

Las redes neuronales son otra forma de emular características humanas en un ordenador. En la **figura 1** se puede observar cómo funciona una red neuronal. (Ruiz & Basualdo, 2001)

Figura 1

Funcionamiento de red neuronal (Ruiz & Basualdo, 2001)



Básicamente usan algoritmos y conexiones entre nodos para procesar información y aprender de ella, se pasa primero por la capa de entrada, después, mediante cálculos

matemáticos en las capas ocultas se obtienen los resultados de la capa de salida. Hay diferentes tipos como: redes neuronales artificiales, recurrentes, feedforward, etc.

1.4.10 Error Out-of-bag

La necesidad de datos para evaluar el modelo no siempre se puede satisfacer, el error out of bag sirve para evaluar el desempeño del modelo sin la necesidad de un conjunto de datos de evaluación, es decir solo usando los datos de entrenamiento. (Bouza & Santiago, 2012)

1.4.11 One-hot-encoding

El método one-hot-encoding permite representar varias palabras, por ejemplo, una oración, como una lista de ceros y unos de la misma longitud que todo el vocabulario tomado en cuenta. El 0 se digita cuando la palabra no aparece en el texto y el 1 cuando si aparece esa palabra en el texto. Dando como resultado en una oración una secuencia de ceros y unos. (González, 2021)

1.4.11 Estado del arte

En este punto se resumen algunas investigaciones referentes a la codificación automática:

En el artículo “Desarrollo de un aplicativo para recolección de datos de diagnóstico automático de desórdenes mentales” (Chauca, 2019) se propone el desarrollo de una página web que utiliza la metodología UML-Based Web Engineering y la aplicación del algoritmo kNN, para asistir a los profesionales del área de salud en el diagnóstico de enfermedades mentales. Esta herramienta busca simplificar el proceso de diagnóstico al abordar las principales dificultades, como la amplia diversidad de trastornos mentales descritos en la CIE-10, la ambigüedad en la conexión entre síntomas y enfermedades, y la ausencia de registros de casos previos que puedan servir como referencia para nuevos diagnósticos, entre otros desafíos.

En el artículo “Desarrollo de un sistema de clasificación multi-etiqueta basado en Transformers para la clasificación de códigos eCIE-O-3.1” (Valencia, 2021) se plantean

diversas arquitecturas basadas en Transformers con el objetivo de abordar un problema de clasificación multietiqueta enfocado en los códigos morfológicos de la CIE-O-3.1, específicamente relacionados con las neoplasias. Los modelos utilizados están basados en diversas variantes de BERT preentrenadas con distintos conjuntos de datos y en diferentes idiomas: algunos están específicamente diseñados para español, mientras que otros son multilingües. Además, ciertos modelos han sido entrenados con textos médicos, mientras que otros se basan en textos de carácter general. Se llevó a cabo un preprocesamiento de los textos médicos para adaptarlos al formato de entrada requerido por BERT. Finalmente, se realizó una evaluación detallada del sistema de clasificación utilizando el conjunto de prueba de CANTEMIST para medir su desempeño.

En el artículo “Estudio preliminar de la anotación automática de códigos CIE-10 en informes de alta hospitalarios” (Almagro M. , Martínez, Fresno, & Montalvo, 2018) se explica que actualmente, la codificación de informes médicos requiere una cantidad significativa de recursos. La reciente adopción del estándar CIE-10 en el sistema sanitario español ha intensificado los desafíos, ya que aumenta el número de códigos CIE posibles para cada informe, mientras que la cantidad de datos disponibles y previamente codificados sigue siendo limitada. Este artículo analiza los retos asociados a esta tarea y presenta una aproximación inicial que combina diversas técnicas para desarrollar un sistema capaz de recomendar automáticamente códigos CIE-10 a los codificadores a partir de un informe médico.

CAPÍTULO II: METODOLOGÍA

En este capítulo se describe paso a paso el proceso utilizado en la estructuración del estudio. Se analizará la forma en que el proceso creativo fue plasmado, de igual manera, se expondrán conceptos teóricos de evaluación para el capítulo de resultados.

2.1. Diseño del estudio

Tomando en cuenta que la propuesta de este proyecto radica en diseñar y evaluar un modelo de codificación automática utilizando técnicas de Machine learning, se enmarca dentro de una investigación de tipo experimental, puesto que su propósito es crear, entrenar y evaluar un modelo estadístico.

Predomina un enfoque analítico, dado que busca reconocer algún tipo de relación entre las características de los diagnósticos médicos y los códigos CIE-10. Es de carácter aplicado ya que pretende resolver un problema de la vida real con resultados que pueden implementarse en el sistema de codificación nacional. Es un estudio retrospectivo debido a que se utilizó datos ya existentes recolectados en 2022 y 2023. Por último, la presente investigación es cuantitativa porque su análisis está relacionado con datos numéricos inherentes a las métricas de evaluación del modelo.

2.2. Población y muestra

La población objetivo de este estudio son todos los registros de personas que hayan ingresado a los establecimientos de salud públicos y privados con internación hospitalaria, es decir, los establecimientos de la Red Pública Integral de Salud (RPIS) y Red Complementaria (RC), de igual manera, los registros de pacientes que egresaron de hospitalización a nivel nacional.

Los criterios de inclusión utilizados son, registros que contengan necesariamente el sexo, la edad, el diagnóstico y el código CIE-10 del paciente. Básicamente, si el establecimiento de salud tiene un buen registro, en todos los casos debe cumplirse este criterio.

En el caso de los códigos, todos están incluidos para extraer mejores resultados, exactamente son 16645 códigos.

Existen algunos registros en que es necesario no solo poner un código CIE-10, ya que según esta clasificación un dictamen médico puede generar varios códigos dependiendo la situación, por ejemplo: una fractura de hueso metacarpiano tiene un código de S623, pero puede haber varias razones para haberse fracturado un hueso, ahí es necesario otro código como el X580 que significa “Exposición a otros factores especificados, vivienda”, Esto quiere decir que la fractura fue ocasionada en el hogar, en el presente estudio se analizarán solo los registros que posean un código único para facilitar el proceso y análisis.

El tamaño de la población es de 18066 registros, es una cantidad grande, aunque para los métodos actuales de machine learning, con ese número no es imposible trabajar, el verdadero inconveniente surge con la gran cantidad de códigos únicos y códigos que nunca son usados, esto afecta demasiado a las conclusiones generadas a través del análisis estadístico. Este es el motivo por el cual se seleccionaron 2 muestras para trabajar, los diagnósticos relacionados a traumatismos y relacionados a obstetricia, los modelos serán entrenados con esos datos y estos mismos se pueden extrapolar para la generalidad de registros.

2.3. Técnicas de recolección de datos

2.3.1. Instrumentos utilizados

El principal recurso es el Registro estadístico de camas y egresos hospitalarios, este recaba información sobre la morbilidad hospitalaria, la utilización de camas hospitalarias de dotación normal y camas disponibles de los establecimientos de salud que presten internación hospitalaria de la Red Pública integral de salud y de la Red complementaria (INEC, 2023). En 2023 se registraron 1170813 egresos hospitalarios, 23645 camas disponibles en 627 establecimientos a nivel nacional. La periodicidad de este registro es anual.

Precisamente, los datos de esta investigación fueron obtenidos al recabar información presente en este registro, en total fueron 18066 observaciones. Cabe destacar que se firmó un acuerdo de confidencialidad para protección de los datos dado que son propiedad del Estado ecuatoriano a través del Instituto Nacional de Estadística y Censos.

Como último instrumento, se utilizó una base de datos que contiene información sobre todos los códigos de la clasificación CIE-10, que consta de 16645 códigos.

Se utilizó diferentes softwares para la realización del proyecto, principalmente R para el procesamiento y análisis estadístico, Excel para manejar las bases de datos y el entorno colaborativo de Python en Google para el trabajo relacionado con la red neuronal.

2.4 Variables de estudio

Dentro de la base de datos recopilada, se obtuvieron 13 variables de estudio, que se resumen en **tabla 1**:

Tabla 1

Variables utilizadas

Variable	Descripción	Tipo
Sexo	Sexo del paciente con codificación 1 si es masculino y 2 si es femenino.	Categórica
Código_edad	Número entre 1 y 4 siendo 1 los días, 2 las semanas, 3 los meses y 4 los años.	Categórica
Edad	Número entero que se combina con el código para formar la edad completa del paciente	Numérica
Internación	Días que el paciente pasó internado.	Numérica
Estado_paciente	forma en que el paciente salió del establecimiento de salud, ya sea vivo o no.	Categórica

Especialidad	Especialidad del médico tratante	Categoría
Diag_def	Diagnóstico principal	Textual
Diag_sec1	Diagnóstico secundario 1	Textual
Diag_sec2	Diagnóstico secundario 2	Textual
Cau_ext	Causa externa para el diagnóstico	Textual
Cod_CIE10	Código CIE-10 del diagnóstico principal	Categoría
Cod_sec1	Código CIE-10 del diagnóstico secundario1	Categoría
Cod_cau_ext	Código CIE-10 de la causa externa	Categoría

La variable dependiente es el código CIE-10, se evaluará las variables que son más importantes mediante selección de características por regulación.

2.5 Procedimiento de análisis de datos

2.5.1 *Procesamiento inicial*

La base de datos obtenida no contuvo datos faltantes, dado que fue obtenida por parte del Instituto ecuatoriano de estadísticas y censos. Ya que se estaba trabajando con registros que contenían bastantes palabras, se tuvo que efectuar un procesamiento del lenguaje natural, con librerías de R tales como “text2vec” y “tm”. Estos paquetes son esenciales para trabajar con grandes volúmenes de texto y así convertirlos en un lenguaje que la máquina pueda entender. Las principales labores fueron eliminar espacios innecesarios, poner todo en minúsculas, eliminar signos de puntuación y tildes, deshacerse de palabras vacías como el, la, de, lo, etc. Se redujo palabras a su forma de raíz para unificar diagnósticos y, por último, se agrupó palabras derivadas con significado igual.

2.5.2 Análisis estadístico

El análisis estadístico se dividió en 3 partes, primero un poco de estadística descriptiva para entender mejor a los datos, se utilizó:

La siguiente parte comprendió en realizar dos modelos de random forest para cumplir el objetivo general de clasificar los códigos CIE-10 en los data sets de traumatismos y de obstetricia respectivamente. Primero se prepararon los datos de diagnósticos, aplicando el procesamiento de lenguaje natural. Después se creó una matriz TF-IDF (term frequency-Inverse document frequency), que es una representación utilizada para cuantificar y analizar la relevancia de cada término en un corpus, que es un conjunto de documentos. Luego se vectorizó el diccionario de vocabulario. Finalmente se entrenó al modelo usando las variables más importantes que fueron: edad (días), sexo del paciente, diagnóstico y código CIE-10. Para el entrenamiento se utilizó un 80% de observaciones para entrenamiento y el restante para prueba. El número de árboles utilizados preliminarmente fue de 100.

Para la red neuronal, la metodología utilizada fue: usar el entorno colaborativo de Python en Google dado que R tenía dificultades al correr el modelo por la naturaleza propia de los datos. Las librerías utilizadas fueron: numpy, para trabajar con cálculos numéricos; Pandas, que es utilizado para manejar y manipular bases de datos; Scikit-learn, que sirve para el aprendizaje automático; Tensor Flow, diseñada para facilitar la creación y entrenamiento de modelos de machine learning. Se utilizó una manera parecida de crear el modelo, aunque se utilizó un 80% de observaciones para entrenamiento. El optimizador usado para compilar el modelo fue el optimizador Adam, que ajusta la tasa de aprendizaje de manera individual para cada parámetro del modelo en función del promedio móvil del gradiente y sus cuadrados. La función de pérdida utilizada fue la de `categorical_crossentropy`, esta es ideal para problemas multiclase y asegura que el modelo genere probabilidades cercanas a la realidad. Finalmente se utilizó `early stopping` para prevenir el sobreajuste.

2.5.3 Validación de resultados

Para los modelos de random forest, las métricas utilizadas para validación son: precisión, índice Kappa, sensibilidad, especificidad, el f1-score y el error out-of-bag. Para la red neuronal se utilizó: la precisión y función de pérdida.

2.6 Limitaciones del estudio

El tamaño de muestra es una limitación, ya que en trabajos que utilizan modelos de machine learning, se necesita una cantidad muy grande de registros para que el entrenamiento tenga calidad, según (Shaip, 2024), “la cantidad del conjunto de datos de entrenamiento necesarios debe ser diez veces mayor que cada parámetro del modelo”, refiriéndose a que la cota inferior para el presente modelo es de al menos 70 registros, no obstante para análisis profundos se requiere que todas las posibles clases estén bien representadas.

La cantidad de diagnósticos que tienen códigos CIE-10 únicos son un gran problema, ya que alrededor de 36% de clases están subrepresentadas y hacen que el algoritmo no trabaje de la mejor manera.

El tiempo computacional es una limitación, ya que entrenar modelos de redes neuronales complejos requiere de una gran capacidad computacional, por lo que se requirió mucho tiempo para correr estos modelos.

Finalmente, la interpretabilidad del modelo presenta un desafío importante de señalar, esto se debe a que mientras más complejo el modelo, más difícil es su interpretación y presentación a las personas.

CAPÍTULO III: RESULTADOS

Este capítulo se centrará en exponer los resultados obtenidos a partir de los modelos creados y del análisis exploratorio realizado, asimismo al final de este capítulo se comparará esta investigación con otras similares. El objetivo de este capítulo es dar a conocer los hallazgos alcanzados en esta investigación.

3.1 Análisis exploratorio

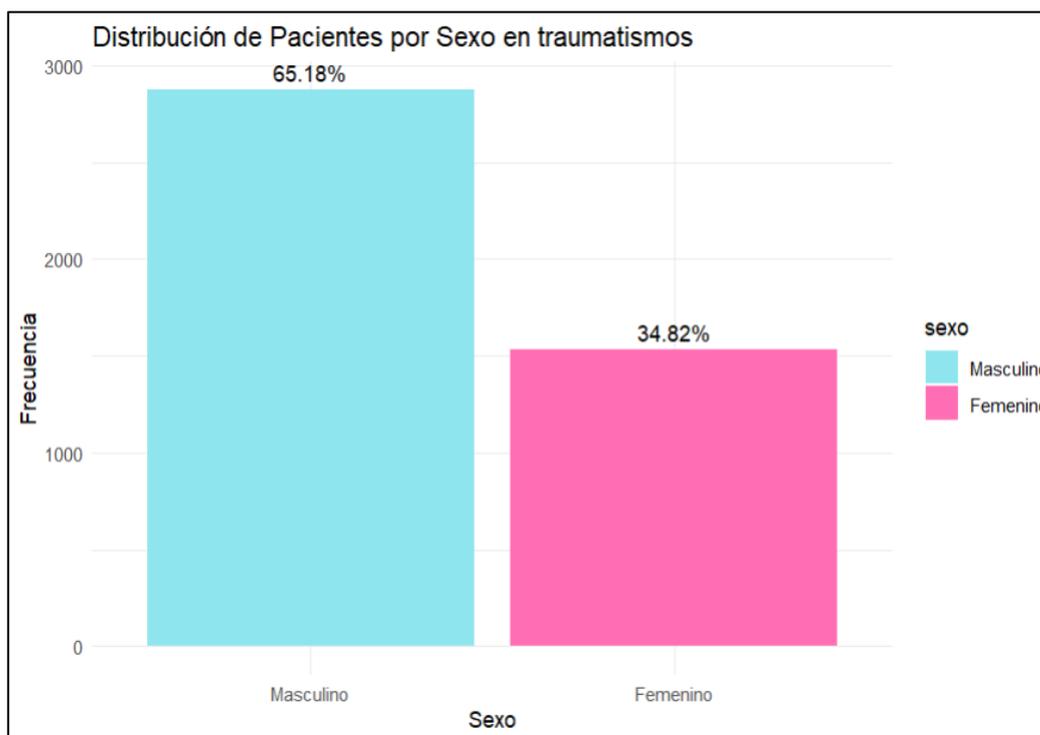
Para entender las características de la población estudiada, se realizó un análisis exploratorio introductorio, la información encontrada se presentará en 2 partes siempre, primero la sección de traumatismos y luego la sección de obstetricia.

3.1.1 Análisis de características demográficas

En la **figura 2**, se observan los porcentajes de cada género en la sección de traumatismos, la mayor cantidad de personas es del sexo masculino con un 65.18% de registros.

Figura 2

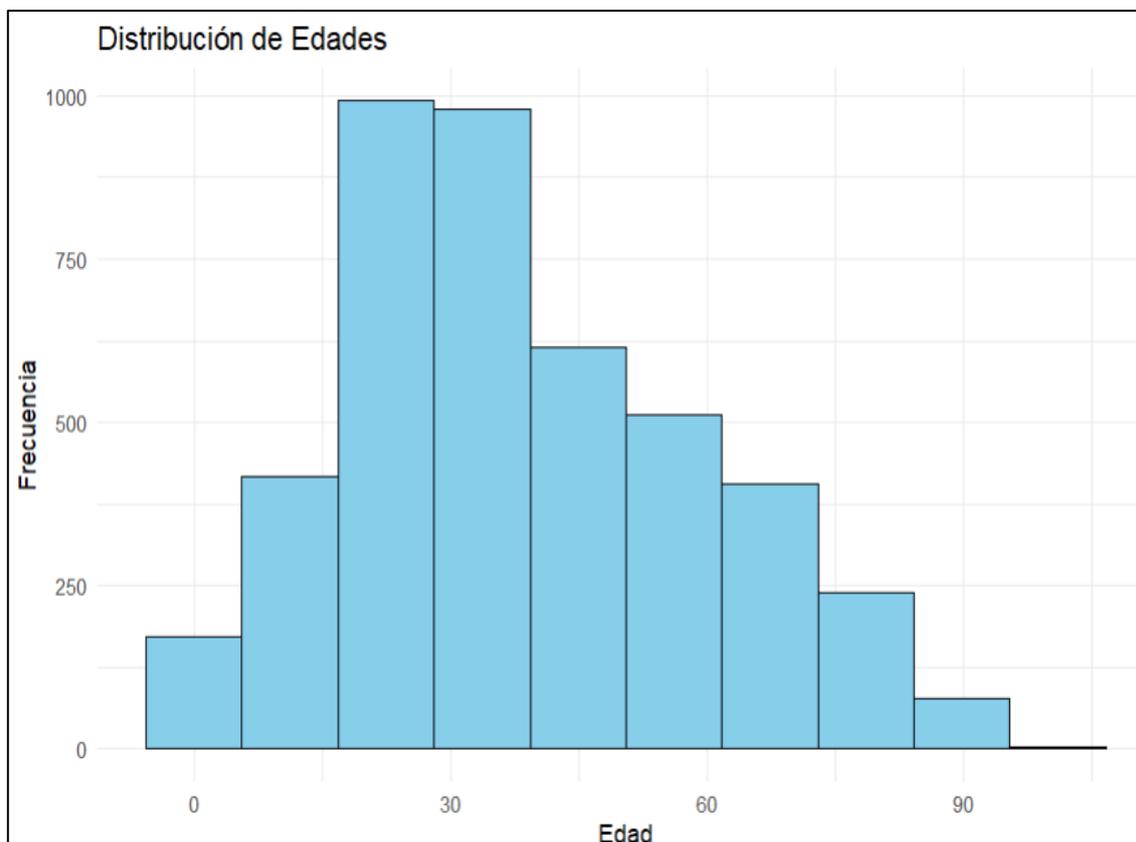
Distribución de pacientes por sexo en la sección de traumatismos



En la **figura 3** se aprecia la distribución de edades de los pacientes, en este caso se encuentra un promedio de 35 años, la edad mínima es de un recién nacido (0 años) y la edad máxima es de 101 años. Esta variable tiene una desviación estándar de 20.7 años y una mediana de 35. La mayoría de pacientes tienen una edad entre los 20 y 40 años.

Figura 3

Distribución de edades de pacientes en la sección de traumatismos.



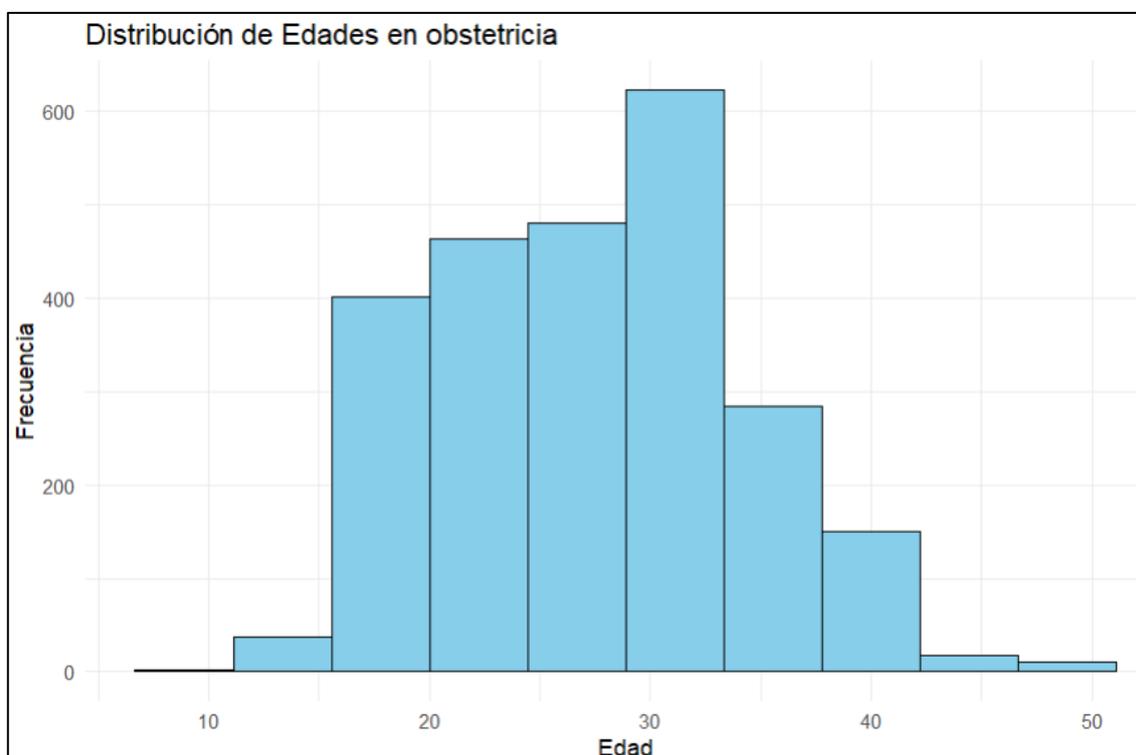
Para la sección de obstetricia, esta fue la información encontrada: en la distribución de pacientes por sexo, el 100% está representado por mujeres dado que esta rama de la medicina se ocupa del cuidado de la mujer durante el embarazo, parto y post parto.

En la **figura 4** se puede apreciar la distribución de edades, esta sección se presenta de una manera muy diferente que en la sección traumatismos, con un promedio y una mediana de 27 años, una edad mínima de 10 años y una máxima de 50 años, entre estas edades esta la ventana fértil en la vida de una mujer, se puede contrastar con la de traumatismos ya que un

golpe, fractura o lesión puede ocurrir en cualquier instante de la vida. La desviación estándar es de 6.66 años.

Figura 4

Distribución de edades de pacientes en la sección de obstetricia



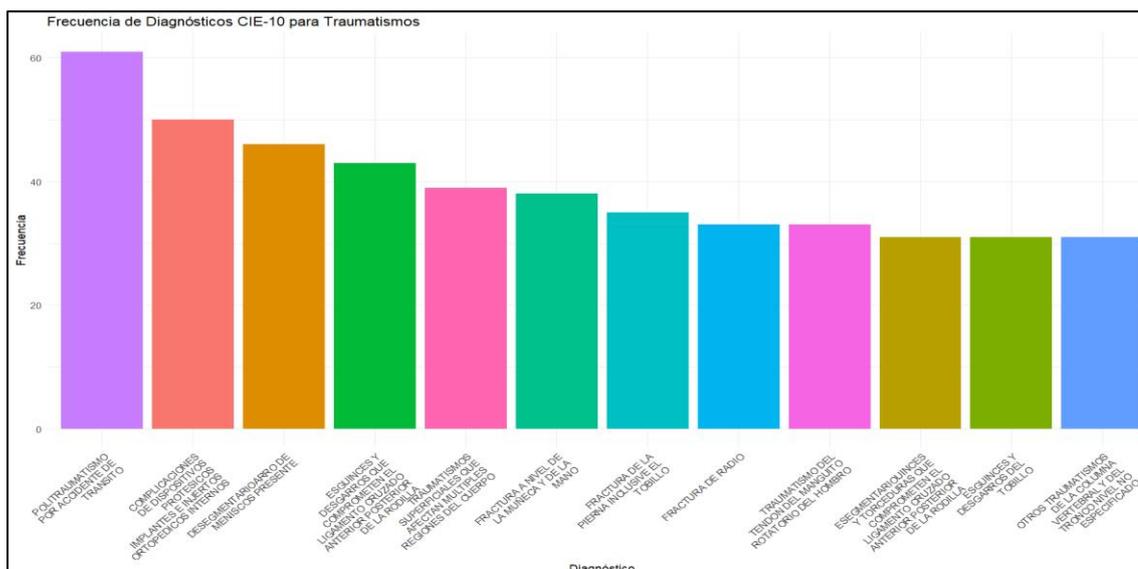
3.1.2 Análisis de los diagnósticos

En cuestión de diagnósticos se determinó cuáles fueron los 10 diagnósticos más comunes en el registro de traumatismos, como se observa en la **figura 5**, el politraumatismo por accidentes de tránsito es el diagnóstico más común, con más de 60 registros, las fracturas en brazos y piernas también es bastante común, las complicaciones por injertos o implantes tiene mucha incidencia.

Cabe destacar que los problemas relacionados a traumatismos representan un 13.12% del total de diagnósticos presentes. Es un porcentaje importante, reflejando que los golpes, lesiones y fracturas están muy presentes en el día a día de los ecuatorianos.

Figura 5

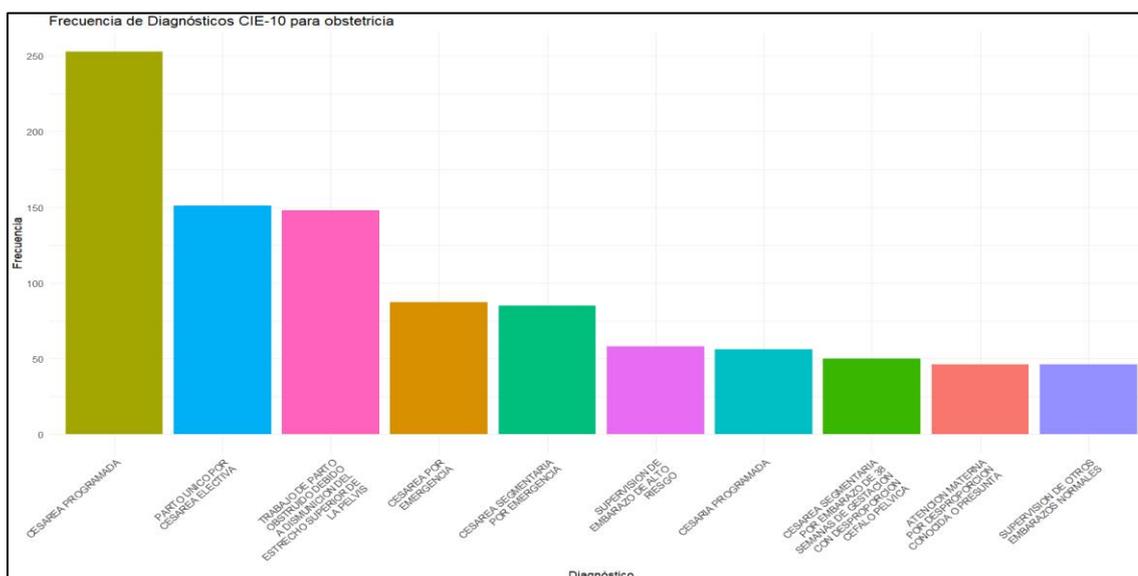
Diagnósticos más frecuentes en la sección de traumatismos



Por otra parte, en el segmento de diagnósticos relacionados con problemas obstétricos, las evaluaciones médicas más comunes se muestran en la **figura 6**. La cesárea programada es el diagnóstico más común, y esta información cobra sentido cuando se entiende que en el Ecuador el número de nacidos vivos es de más de 600 al día. (INEC, 2023). Los diagnósticos obstétricos representan el 2.94% del total.

Figura 6

Diagnósticos más frecuentes en la sección de obstetricia

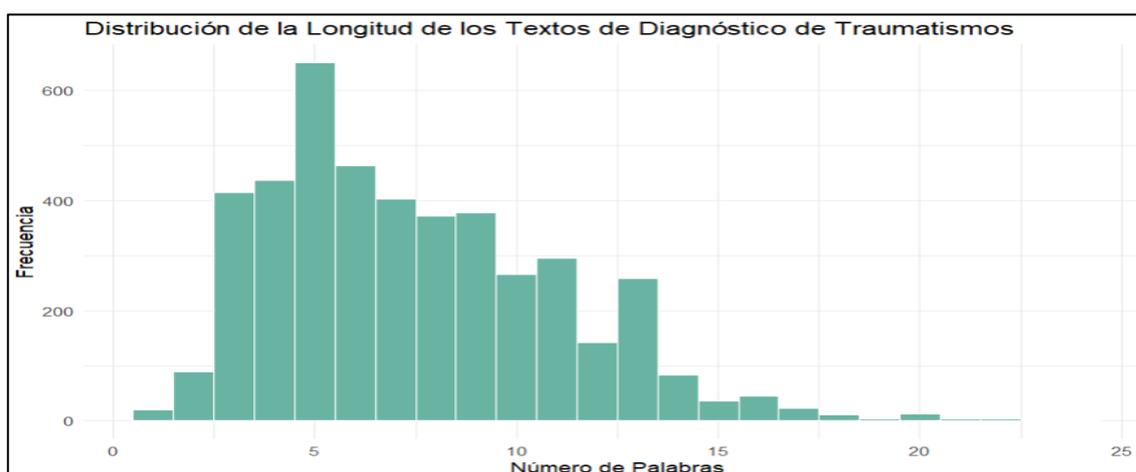


3.1.3 Análisis de características del texto

Se examinó las características de la variable más importante de este estudio, el diagnóstico médico. En la **figura 7**, se registra la distribución de la longitud de palabras en la unidad de traumatismos, teniendo como promedio 7 palabras y 49 letras. Las palabras más frecuentes son: fractura, mano, traumatismo, golpe y múltiples. (Más información en la sección de anexos)

Figura 7

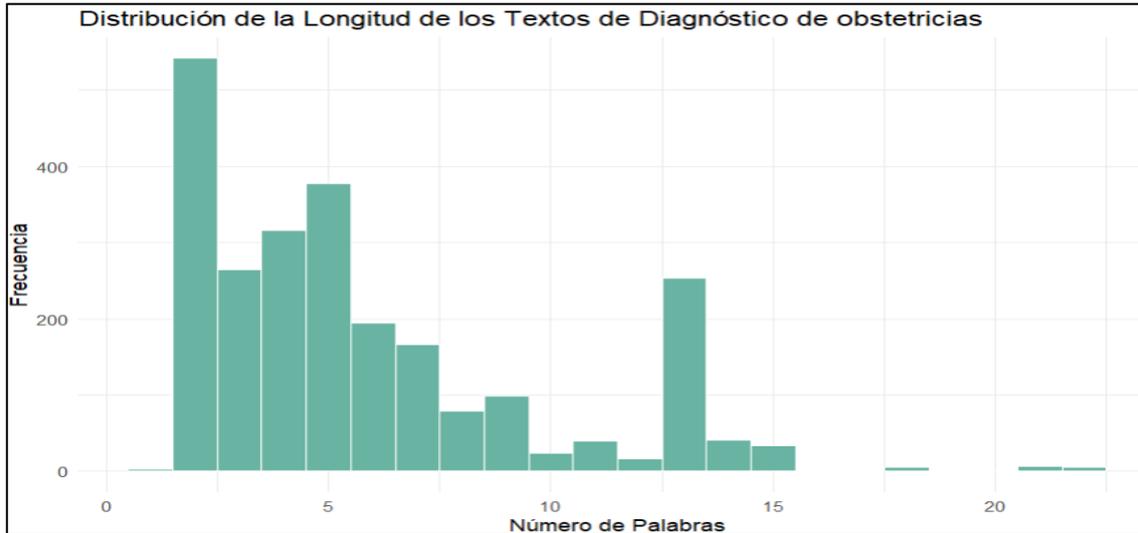
Distribución de la longitud de palabras en la sección de traumatismos



Para el apartado de obstetricia, el promedio de palabras es de 6 por diagnóstico y de 42 letras respectivamente, como se ve en la **figura 8**. Comparado con la sección anterior, para obstetricia se usan menos palabras y letras. Los términos más frecuentes son: cesárea, parto, programada y embarazo. (Más información en la sección de anexos)

Figura 8

Distribución de la longitud de palabras en la sección de obstetricia



3.2 Modelo de Random Forest

El primer modelo entrenado fue el modelo de random forest, en esta sección se analizará los resultados obtenidos a partir de este modelo, como es general en este trabajo, primero se analizará la parte de traumas y después la parte de obstetricia.

3.2.1 Optimización

Aunque el número de variables no es tan grande, se decidió hacer una selección de variables no tan específica. Esto se explica gracias a que Random forest es, en sí mismo, un modelo que ayuda a la selección de características (variables) más importantes en sí mismo. Sin embargo, se utilizó la importancia basada en la reducción de impureza, este método se calcula observando la disminución de la entropía. Las variables más importantes fueron el sexo, la edad y el diagnóstico.

Al principio se intentó correr el modelo de random forest con 100 árboles para todos los 18000 datos, el entrenamiento de ese modelo duró alrededor de 7 horas. La duración de los modelos de traumatismos y de obstetricia fue de casi 2 y 3 horas aproximadamente. Gracias a las optimizaciones generadas en el código, que van desde arreglar el preprocesamiento de texto

hasta reducir la dimensionalidad con la selección de características previo, se logró reducir el tiempo a solo minutos en los 2 segmentos de datos.

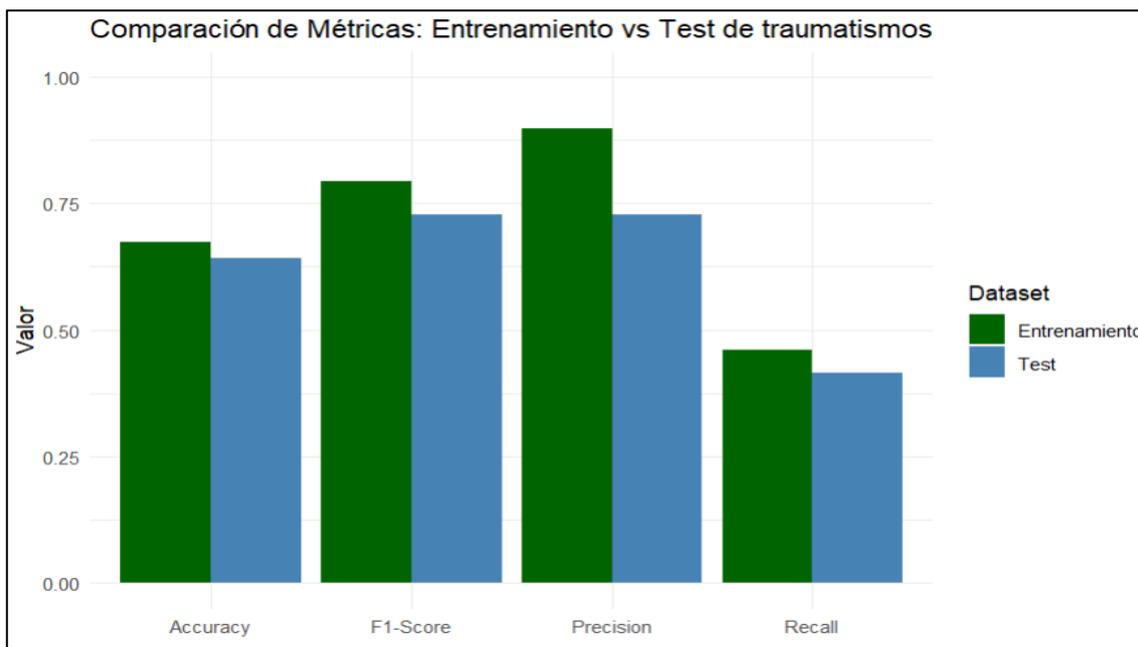
3.2.2 Evaluación del modelo

Como se explicó en el capítulo anterior, la forma de realizar una evaluación de los modelos es en base a sus métricas, en la **figura 9** se puede observar una comparación entre las métricas de entrenamiento y las de prueba. El primer punto a analizar es que las barras siguen la misma tendencia, esto nos quiere decir que no se está cayendo en un sobreajuste ni en un subajuste.

La precisión ronda entre los 0.65 esto quiere decir que alrededor de 65% de diagnósticos son correctamente clasificados, la sensibilidad es de alrededor de 0.5 siendo un poco baja. El F1-score es más de 0.75 indicando que tiene un buen equilibrio entre precisión y sensibilidad. La precisión es de alrededor de 0.8, esto minimiza los falsos positivos.

Figura 9

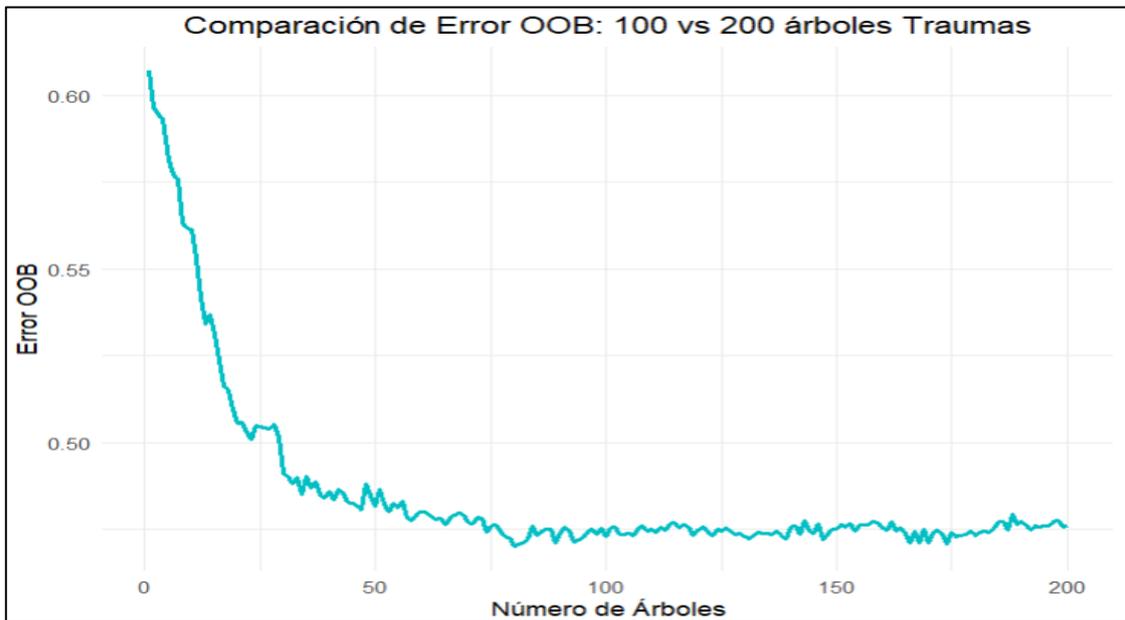
Comparación de métricas de traumatismos



Otra forma intrínseca de evaluar la capacidad predictiva del modelo es el error out-of-bag, en la **figura 10** se puede observar que con más árboles decrece el error, estabilizándose alrededor de 0.45 por el árbol número 50.

Figura 10

Error out-of-bag en la sección de traumatismos

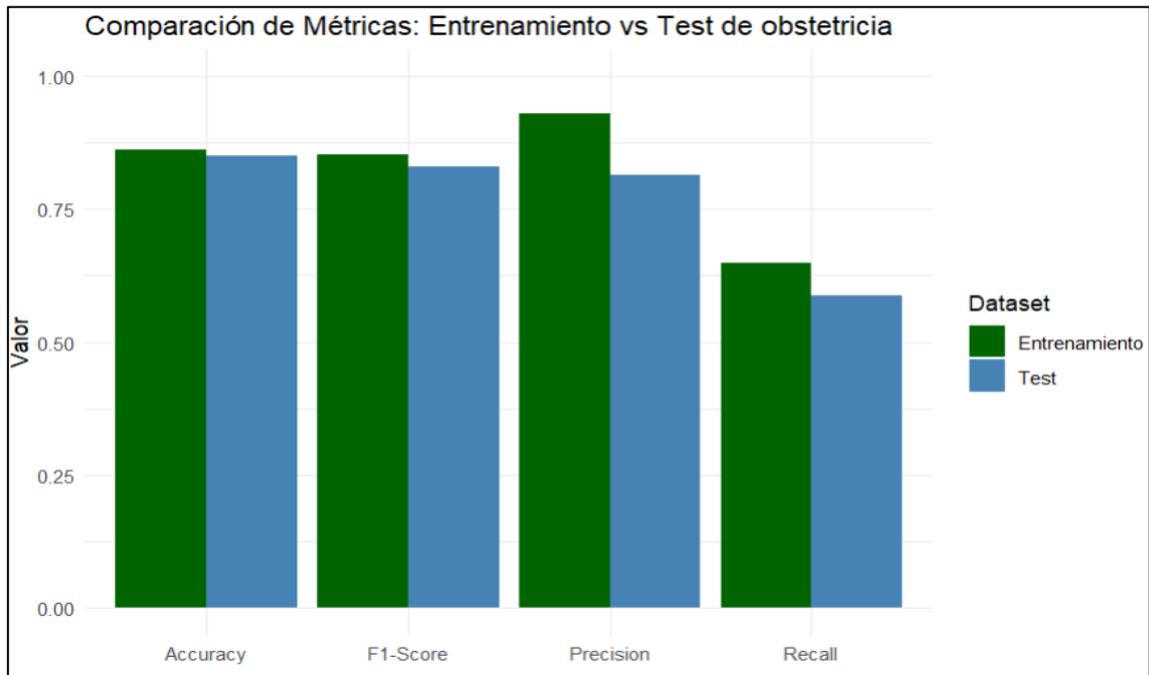


Para la base de datos de obstetricia se calculó exactamente las mismas métricas, en la **figura 11** se puede observar la comparación de indicadores, de igual manera, se puede apreciar que las barras siguen la misma tendencia, descartando preliminarmente el sobre y sub ajuste.

De nuevo el recall, es decir, la sensibilidad, es la métrica más baja, aunque comparado a la sección de traumatismos, todas las métricas aumentaron. La exactitud y el F1-score rondan los 0.85, la precisión es de nuevo el indicador más alto de todos.

Figura 11

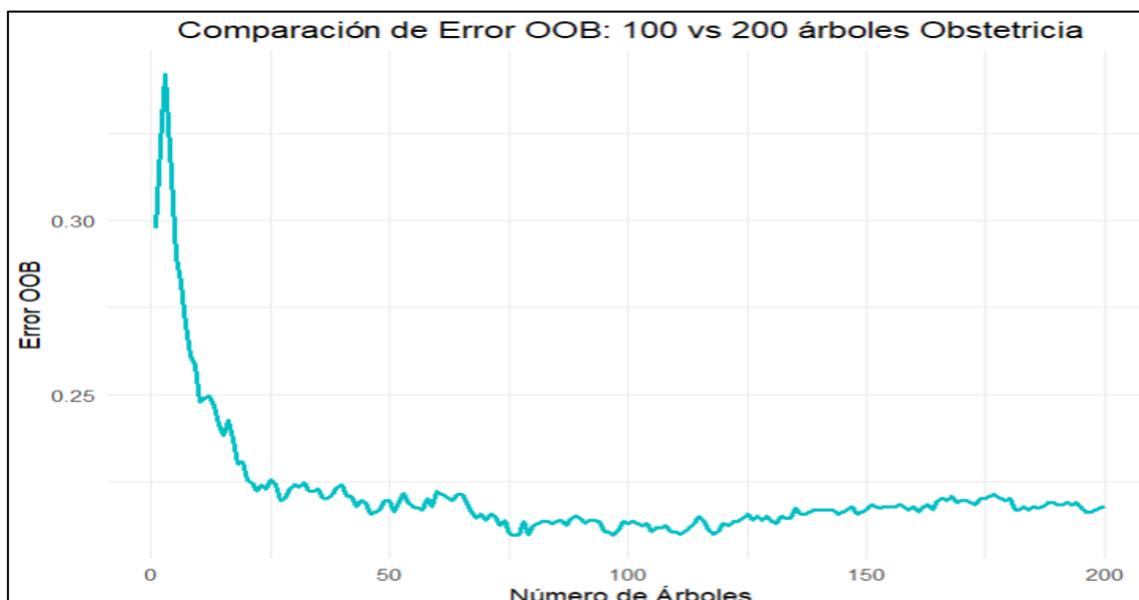
Comparación de métricas de obstetricia



En la **figura 12** se puede notar la evolución del error out-of-bag, en este caso el error es mucho menor que en la sección anterior, dando a entender una mejor capacidad predictiva de este modelo, y de igual manera se estabiliza cerca del 0.2 aproximadamente en el árbol 75.

Figura 12

Error out-of-bag en la sección de obstetricia



3.3 Red Neuronal

Esta investigación fue iniciada principalmente para desarrollar una red neuronal que satisfaga los objetivos de esta misma, en esta sección se analizará los resultados de este modelo.

3.3.1 Optimización del modelo

La red neuronal se estableció con las variables seleccionadas en el modelo de random forest, empezó trabajando con el optimizador Adam (adaptive moment estimation) que es un algoritmo utilizado ampliamente en redes neuronales, ajusta los pesos de la red neuronal durante el entrenamiento usando información del gradiente y los momentos de primer y segundo orden. Este optimizador funcionó bien pero todavía quedaban asuntos por mejorar.

Para que todo el entorno mejore, se realizaron los siguientes cambios: se mejoró la arquitectura aplicando normalización por lotes para estabilizar y acelerar el entrenamiento, se aumentó la profundidad de la red con una capa oculta adicional; se mejoró el preprocesamiento gracias a la aplicación de n-gramas en la matriz TF-IDF para capturar mejor el contexto del diagnóstico; se usó técnicas de regularización como pesos por clase para manejar el desbalance de estas y se optimizó el entrenamiento gracias a la validación cruzada k-fold.

Gracias a todas estas herramientas se mejoró considerablemente el trabajo realizado por la red neuronal, que, aunque ahora necesita más tiempo computacional, nos brinda mejores resultados de clasificación, evidenciándose esto en la mejoría de las principales métricas.

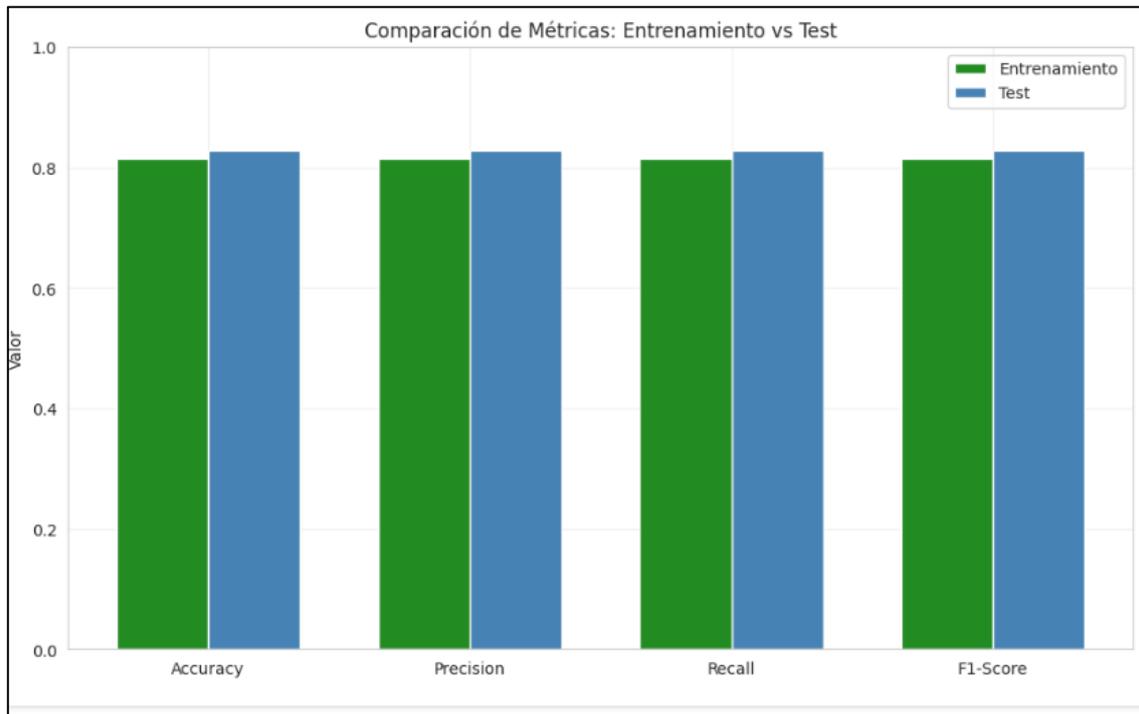
3.3.2 Evaluación del modelo

Al igual que en el modelo de random forest, en el modelo actual se utilizaron las métricas como método de evaluación, en la **figura 13** se observa una comparación de las métricas de entrenamiento vs las de prueba.

Todas las métricas están sobre los 0.8, se ve que las barras siguen la misma tendencia, así que el modelo no está ni sobre ni sub ajustado. Existe una importante diferencia que radica en que la red juntó las dos bases de datos así que solo hay un gráfico de métricas.

Figura 13

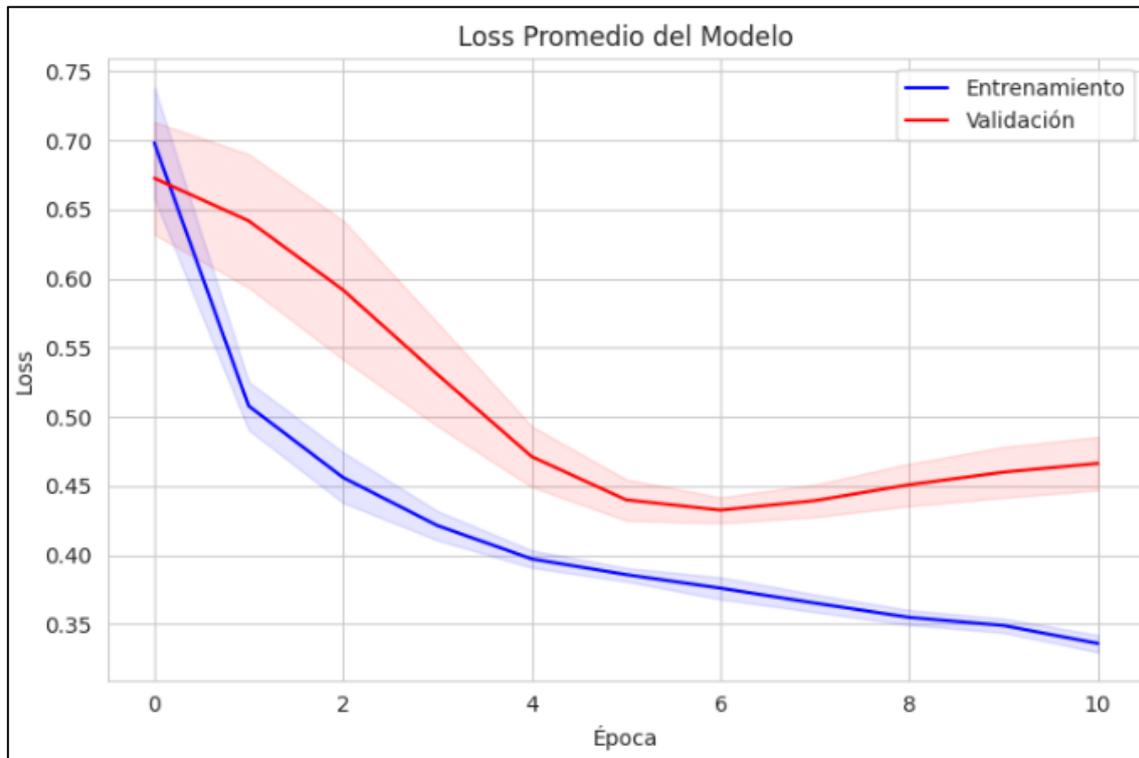
Comparación de métricas de entrenamiento vs prueba de la red neuronal



En la figura 14 se muestra una curva de aprendizaje que explica cómo evoluciona la función de pérdida en cuanto a cada época, este gráfico es de un promedio de todos los folds de la validación cruzada, en el gráfico se nota que la función decrece, ajustándose mejor a los datos.

Figura 14

Curva de aprendizaje de la red neuronal



3.4 Comparación entre los modelos

Una vez entrenados los modelos, el último paso es su comparación, se realizará a través de dos mecanismos, las métricas de cada uno y su capacidad predictora.

3.4.1 Comparación de métricas

En la **tabla 2** se observa el resumen de las métricas de entrenamiento de los modelos creados.

Tabla 2

Resumen de métricas de entrenamiento

Modelo	Exactitud	Sensibilidad	Precisión	F1-score
Random forest Traumas	0.67	0.46	0.89	0.79
Random forest Obstetricia	0.86	0.65	0.93	0.85
Red neuronal	0.7812	0.7999	0.7967	0.7811
TOTAL (promedio)	0.7704	0.6367	0.8722	0.807

En la **tabla 3** se muestra el resumen de las métricas de prueba de los modelos creados.

Tabla 3

Resumen de métricas de prueba

Modelo	Exactitud	Sensibilidad	Precisión	F1-score
Random forest Traumas	0.65	0.42	0.73	0.73
Random forest Obstetricia	0.85	0.59	0.83	0.82
Red neuronal	0.8015	0.8124	0.8097	0.8001
TOTAL	0.7671	0.6075	0.7899	0.7833

Hallando información en las métricas, se puede entender que el mejor modelo está entre el random forest de la sección obstétrica y la red neuronal, tomando como referencia el F1-score.

3.4.2 Capacidad predictora

Para finalizar las comparaciones se utilizó a las predicciones, ingresando 100 datos aleatorios, como “Fractura de coxis”; 8 años; sexo masculino, se analizó qué modelo dio la mejor clasificación. El modelo de red neuronal optimizado fue el mejor clasificando correctamente casi el 80% de entradas. Una curiosidad es que la red neuronal clasificó como “no posible” a los datos obstetras masculinos, a diferencia de los modelos de random forest que intentaban dar un código.

El mejor modelo entonces es la red neuronal, es entendible debido a la naturaleza de los datos y que tenía algunas clases únicas, haciendo una compensación entre la métrica F1-score y la capacidad predictora.

3.5 Discusión

En este último apartado se realiza un análisis acerca de trabajos y resultados relacionados con la investigación elaborada. Se analizaron 4 trabajos y finalmente se compararon con el presente proyecto.

El primer paper es “Estudio preliminar de la anotación automática de códigos CIE-10 en informes de altas hospitalarios” (Almagro M. , Martínez, Fresno, & Montalvo, 2018), su investigación, realizada en Madrid, se basó prácticamente en crear igual una clasificación automática para los códigos CIE-10, usaron un Support vector machine (SVM) que es un modelo de aprendizaje supervisado, especialmente utilizado para resolver problemas de clasificación binaria, sus resultados reflejan problemas en la clasificación de códigos poco frecuentes, ya que el modelo se desbalancea y no cumple bien su función.

El siguiente trabajo es la “Codificación automática de enfermedades CIE-10 ES para episodios asistenciales con técnicas de Procesamiento de Lenguaje Natural” (Valdivieso, 2018), este trabajo, realizado en Valencia, busca diseñar, validar e implementar un prototipo de sistema de codificación automático para una clínica en España. Se implementó un modelo analítico de autoaprendizaje que, partiendo de un texto libre no estructurado, automatice la categorización CIE-10, usando procesamiento de lenguaje natural. Sus resultados son de alrededor de un 88% de diagnósticos codificados correctamente, generando un ahorro económico y de tiempo.

La investigación “Publicly Available Clinical BERT Embeddings”, traducido como Embeddings clínicos BERT públicamente disponibles (Alsentzer, et al., 2019). En esta investigación, realizada en Minneapolis, se usó el modelo BERT que es un modelo de lenguaje basado en transformers, su característica principal es que puede entender el texto de manera bidireccional, es decir, usando el contexto de las palabras anteriores y siguientes. Usaron un modelo embedding que transformó los diagnósticos en vectores numéricos y se clasificó dentro de un grafo, estableciendo una clasificación de hasta un 93% de aciertos.

Un texto importante para analizar, ya que se elaboró en un contexto local, es la tesis “Desarrollo de dos modelos de clasificación usando boosting y redes neuronales para codificar las actividades económicas y ocupaciones de investigaciones sociodemográficas del INEC”

(Méndez, 2023), esta tesis aplica dos modelos para desarrollar la clasificación automática de algunos indicadores necesitados por el INEC en la ciudad de Quito. Primero utilizó el algoritmo de Xgboost (extreme gradient boosting) que es una optimización de un modelo boosting basado en árboles de decisión, uniendo varios árboles débiles y creando un modelo fuerte. El f1-score de este modelo rondó el 91% en entrenamiento y el 82% en validación. También utilizó una red neuronal feedforward, que consiste en una red en donde la información viaja a través de una sola vía sin retroalimentación ni conexiones entre las capas pasadas. Es una red simple pero que tiene una gran cantidad de aplicaciones. El f1-score generado en esta red es de alrededor de 96% con los datos de entrenamiento y un 86% con los datos de prueba. La diferencia de ese trabajo con este es el área de codificación, las actividades codificadas en esa investigación no tiene tantas clases únicas como la de los códigos CIE-10, mejorando el rendimiento de los modelos.

Se han analizado 4 trabajos en los cuales se ha mostrado que en otros países se está más avanzado en temas de codificación automática utilizando algoritmos de machine learning. Sin embargo, los resultados no son tan dispares en relación a los presentados en el presente proyecto, indicando que existe una idea clara del panorama general de la clasificación automática.

CAPÍTULO IV: CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

- La naturaleza de los datos complicó el desarrollo y entrenamiento de los modelos dado que poseía muchas clases únicas, dificultando su clasificación oportuna, este problema se encontró, también, en trabajos internacionales similares.

- La incorporación de procesamiento inicial de texto como la matriz TF-IDF, mejoró significativamente la capacidad de los modelos de interpretar y catalogar los términos médicos.

- Se creó, entrenó y validó 2 modelos de aprendizaje automático para clasificar códigos CIE-10, estos modelos fueron: el primero de random forest, el segundo una red neuronal optimizada. Los dos tuvieron un funcionamiento similar, destacando su efectividad para la codificación automática basado en diagnósticos de la Red de salud pública nacional y de la red complementaria.

- Las métricas establecidas para evaluación de modelos y la capacidad predictora fueron los indicadores clave para escoger el mejor modelo. Principalmente la exactitud y el F1- score dado que los datos están desbalanceados. El análisis demuestra que la red neuronal optimizada es el mejor modelo para clasificar códigos CIE-10.

- Los modelos desarrollados permitirán reducir el tiempo de codificación manual del personal del INEC, empleando los recursos tecnológicos de buena manera en una institución estatal.

4.2 Recomendaciones

- Sería recomendable ampliar el conjunto de datos de entrenamiento para así poder tener más ejemplos de las categorías más representadas y que el modelo no esté tan desbalanceado, esto ayudaría a mejorar la precisión de clasificación.

- Se recomienda leer toda la literatura relacionada con los códigos CIE-10 para entender de mejor manera los temas tratados en esta investigación.

- En trabajos futuros, se podrían evaluar los modelos con métricas extras, como la curva ROC o analizar la matriz de confusión. Estas métricas no fueron realizadas en el presente trabajo debido a la premura en la entrega de resultados.

- La evaluación continua y la actualización de los modelos es indispensable para garantizar una eficacia independiente del tiempo en el que se usen estos modelos.

- El desarrollo de una interfaz gráfica amigable con el usuario sería un gran paso para empezar a implementar estos modelos dentro de establecimientos de la Red de salud pública nacional y red complementaria.

BIBLIOGRAFÍA

Almagro, M., Martínez, R., Fresno, V., & Montalvo, S. (2018). *Estudio preliminar de la anotación automática de códigos CIE-10*. Madrid: Universidad Nacional de Educación a Distancia (UNED).

Almagro, M., Martínez, R., Fresno, V., & Montalvo, S. (2018). *Estudio preliminar de la anotación automática de códigos CIE-10 en informes de alta hospitalarios*.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). *Publicly Available Clinical BERT Embeddings*. Minneapolis: ClinicalNLP.

Bouza, C., & Santiago, A. (2012). LA MINERÍA DE DATOS: ÁRBOLES DE DECISIÓN Y SU APLICACIÓN EN ESTUDIOS MÉDICOS. *Modelación matemática de fenómenos del medio ambiente y la salud*, 64-78.

Chauca, D. (2019). *Desarrollo de un aplicativo para recolección de datos de diagnóstico automático de desórdenes mentales*. Quito: Universidad Politécnica Salesiana.

Díaz Novás, J., Gallego, B., & León, A. (2006). El diagnóstico médico: bases y procedimientos. *Revista Cubana de Medicina General Integral*, 22.

Gelbukh, A. (2010). Procesamiento de Lenguaje Natural y sus aplicaciones. *Komputer Sapiens*, 6-11.

González, A. (2021). *El procesamiento de textos jurídicos: natural language processing & legal tech*. <https://repositorio.comillas.edu/rest/bitstreams/437255/retrieve>.

INEC. (2023). *Registro estadístico de camas y egresos hospitalarios*. Ecuador.

Méndez, D. (2023). *Desarrollo de dos modelos de clasificación usando boosting y redes neuronales para codificar las actividades económicas y ocupaciones de investigaciones sociodemográficas del INEC*. Quito: Escuela Politécnica Nacional.

Moreno, D. C. (2023). *Desarrollo de dos modelos de clasificación usando boosting y redes neuronales para codificar las actividades económicas y ocupaciones de investigaciones sociodemográficas del INEC*. Quito: Escuela Politécnica Nacional.

Organización Panamericana de la Salud. (2015). *Clasificación Estadística Internacional de Enfermedades y Problemas relacionados con la Salud 10ma Revisión*. Washington D. C.: World Health Organization.

Palomino, M. (2023). *Los Grandes Modelos del Lenguaje basados en Transformers : revisión y aplicación práctica con ChatGPT*. Universidad Pontificia Comillas.

Rouhiainen, L. (2018). *Inteligencia artificial*. Madrid: Alienta Editorial.

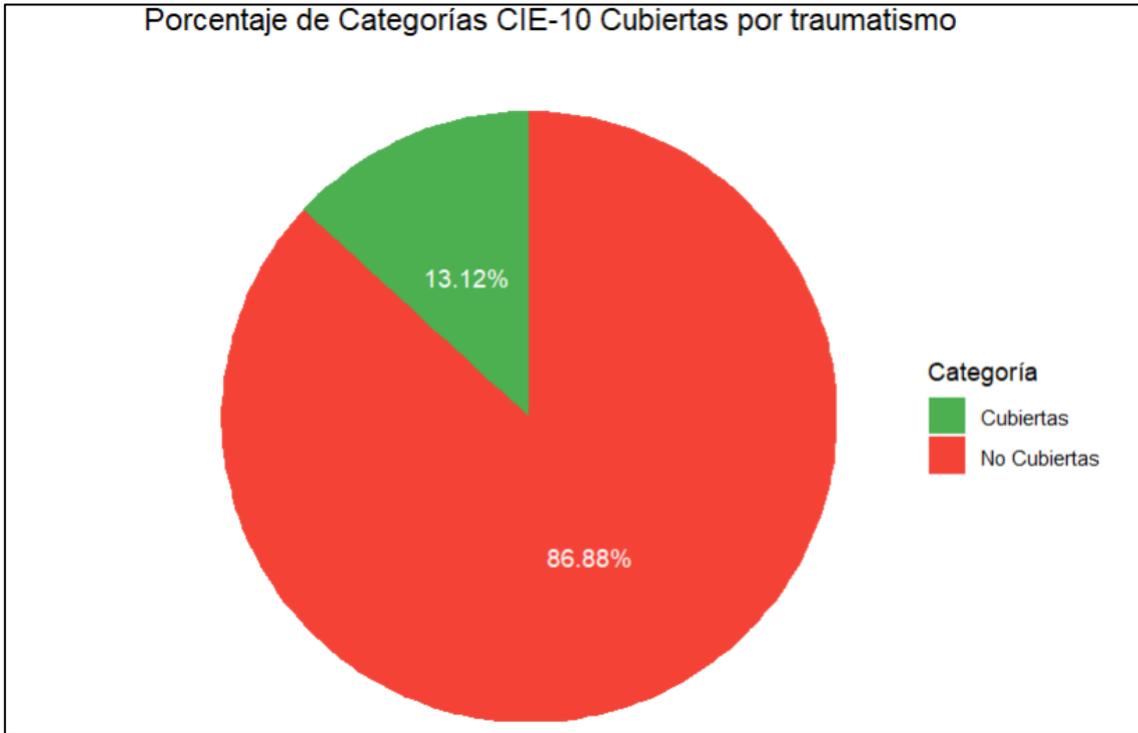
Ruiz, C., & Basualdo, M. (2001). *Redes neuronales: conceptos básicos y aplicaciones*. Rosario: Universidad Tecnológica Nacional.

Valdivieso, B. (2018). *Codificación automática de enfermedades CIE-10 ES para episodios asistenciales con técnicas de Procesamiento de Lenguaje Natural*. Valencia: IIS La Fe.

Valencia, L. (2021). *Publicación: Desarrollo de un sistema de clasificación multi-etiqueta basado en Transformers para la clasificación de códigos eCIE-O-3.1*. Madrid.

ANEXOS

Anexo I: Porcentaje de categorías CIE-10 cubiertas por la sección de traumatismos



Anexo II: Porcentaje de categorías CIE-10 cubiertas por la sección de obstetricia

