|  |  |
| --- | --- |
| **HigiaHealthCode** Eina de codificació d’històries clíniques amb CIE-10 | |
|  | |
|  | **Marc Serret Monserrat**  Màster Universitari en Ciència de Dades  Àrea 3: Machine Learning and Computer Vision in Healthcare and Medical Applications  **Tutor/a de TF**  Susana Pérez Álvarez  **Professor/a responsable de l’assignatura**  Laia Subirats Maté    Data Lliurament  Diumenge, 25 de maig de 2025 |

[Llicència de Creative Commons](http://creativecommons.org/licenses/by-nc-nd/3.0/es/)

Aquesta obra està subjecta a una llicència de [Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons](http://creativecommons.org/licenses/by-nc-nd/3.0/es/)

Fitxa del Treball Final

|  |  |
| --- | --- |
| **Títol del treball:** | HigiaHealthCode |
| **Nom de l’autor/a:** | Marc Serret Monserrat |
| **Nom del Tutor/a de TF:** | Susana Pérez Álvarez |
| **Nom del/de la PRA:** | Laia Subirats Maté |
| **Data de lliurament:** | 05/2025 |
| **Titulació o programa:** | Màster Universitari en Ciència de Dades |
| **Àrea del Treball Final:** | Àrea 3: Machine Learning and Computer Vision in Healthcare and Medical Applications |
| **Idioma del treball:** | Català |
| **Paraules clau** | (DL)*Deep Learning*, (ML)*Machine Learning*, PLN (Processament de llenguatge natural) |

|  |
| --- |
| **Resum del Treball** |
| El projecte es basa en el desenvolupament d’un sistema d’ajuda a la codificació d’altes mèdiques mitjançant tècniques de processament de llenguatge natural (PLN) i *deep learning*. L’objectiu principal és agilitzar la codificació de diagnòstics en CIM-10 a partir dels textos clínics redactats pels professionals assistencials, entrenant el model amb la codificació realitzada per experts en codificació mèdica. Aquesta eina busca reduir el temps dedicat a la codificació manual i millorar la coherència i precisió del codis assignats.  Per la implementació és farà servir una Pytorch com a eina principal per al desenvolupament dels models de *deep learning*. Els textos clínics emprats inclouen informació rellevant com la malaltia actual, evolució del pacient i altres dades clíniques recollides des de l’ingrés fins l’alta.  El sistema resultant ha de ser una eina de suport que faciliti la identificació i qualitat de la informació clínica codificada. |
| **Abstract** |
| The project is based on the development of a support system for medical discharge coding using natural language processing (NLP) techniques and deep learning. The main objective is to streamline the coding of diagnoses in ICD-10 from clinical texts written by healthcare professionals, training the model with coding performed by medical coding experts. This tool aims to reduce the time spent on manual coding and improve the consistency and accuracy of the assigned codes.  Pytorch will be used as the main tool for developing the deep learning models. The clinical texts used include relevant information such as the current illness, patient evolution, and other clinical data collected from admission to discharge.  The resulting system is intended to be a support tool that facilitates the identification of diagnoses and improves the quality of the coded clinical information. |

Índex

[1. Introducció 1](#_Toc194156460)

[1.1. Context i justificació del Treball 1](#_Toc194156461)

[1.2. Objectius del Treball 1](#_Toc194156462)

[1.3. Impacte en sostenibilitat, ètic-social i de diversitat 2](#_Toc194156463)

[1.4. Enfocament i mètode seguit 3](#_Toc194156464)

[1.5. Planificació del Treball 4](#_Toc194156465)

[1.6. Breu sumari de productes obtinguts 5](#_Toc194156466)

[1.7. Breu descripció dels altres capítols de la memòria 5](#_Toc194156467)

[2. Base teòrica i fonaments 6](#_Toc194156468)

[2.1 Sistema d’Informació Sanitari i la Gestió de les dades. 6](#_Toc194156469)

[2.2 Processament del Llenguatge Natural 7](#_Toc194156470)

[2.3 Deep Learning 8](#_Toc194156471)

[3. Materials i mètodes 8](#_Toc194156472)

[3.1 Tecnologies utilitzades 8](#_Toc194156473)

[3.2 Font de dades 8](#_Toc194156474)

[4. Resultats 10](#_Toc194156475)

[5. Conclusions i treballs futurs 11](#_Toc194156476)

[6. Glossari 12](#_Toc194156477)

[7. Bibliografia 12](#_Toc194156478)

[8. Annexos 13](#_Toc194156479)

# Introducció

## Context i justificació del Treball

El projecte HigaHealthCode sorgeix com a resposta a una necessitat detectat dins l’empresa on treballo actualment: la Xarxa Sanitària, Social i Docent de Santa Tecla. Aquesta entitat, que gestiona un ampli conjunt de centres sanitaris a l’àrea del Tarragonès i Baix Penedès, així com centres de serveis d’atenció intermèdia, residència i centres d’atenció primària, s’enfronta a un volum molt elevat de codificació d’altes mèdiques.

En l’actualitat, la codificació d’aquests diagnòstics, basats en la CIM-10 (*International Classification of Diseases, Tenth Revision, Clinical Modifications*) es realitza amb l’estructura i els recursos disponibles, però el gran nombre de centres i la quantitat d’altes generen una càrrega de treball considerable. Degut a la demanda i a la necessitat de revisar més tipus d’activitats fa que la quantitat de treball hagi estat augmentant, fen molt difícil mantenir el nivell de qualitat exigit en la codificació d’altes clíniques.

En resum la justificació del projecte es basa amb els següents punts:

* **L’impacte en la gestió clínica i administrativa**: Una codificació automàtica i més precisa per millora la qualitat de la informació clínica, essencial per a la presa de decisions i la gestió hospitalària.
* **L’oportunitat de millorar processos**.La implementació d’una eina tecnològica avançada permetrà reduir els temps de processament i minimitzar error, contribuint a una gestió més eficient dels recursos.

## Objectius del Treball

L’objectiu d’aquest projecte és desenvolupar un sistema d’ajuda a la codificació d’altes mèdiques basa en tècniques de processament de llenguatge natural (PLN) i *deep learning,* que permeti automatitzar la classificació de diagnòstic en CIM-10.

S’han establert els següents objectius:

1. Objectiu principal:
   * Desenvolupar un model de deep learning capaç d’automatitzar la codificació d’altes mèdiques a partir de textos clínics, millorant la precisió i l’eficiència del procés en un entorn real.
2. Objectius secundaris
   * Recollida i pre-processament de dades:
     + Extreure textos clínics d’una historia clínica, garantint el compliment dels requisits de seguretat i privacitat.
     + Realitzar una neteja, normalització dels textos, així com la tokenització i vectorització utilitzant models de PLN preentrenats.
   * Desenvolupament i entrament del models:
     + Implementar i entrenar diverses arquitectures de *deep learning* ( xarxes neuronals recurrents i transformadors) mitjançant Pytorch.
     + Ajustar els hiperparàmetres del model per optimitzar el rendiment, utilitzant tècniques de validació creuada per evitar el overfitting*.*
   * Validació i comparativa
     + Comparar els resultats obtinguts amb la codificació manual realitzada per experts, utilitzant mètriques com la precisió, el *recall* i el *F1-score*.
     + Realitzar un anàlisi d’errors per identificar àrees de millorar i validar la robustes del model.
   * Integració i avaluació pràctica.
     + Desenvolupar una API que permeti la integració del sistema dins del flux clínic de una historia clínica.
   * Futures implementacions.
     + Analitzar possibles millores del model, processar textos en diversos idiomes i integrar-ho dins el model, o ajudar amb la codificació dels procediments (CIM10-SCP), explorar altres tecnologies emergents que puguin optimitzar aquest procès.

## Impacte en sostenibilitat, ètic-social i de diversitat

* Sostenibilitat:

La implementació d’un sistema automatitzat permetrà una optimització dels recursos humans dins l’empresa, en reduir la dependència del procés manual, els professionals dedicats a aquest àmbit podran invertir més temps en aquells casos que realment ho necessitin a més que podran dedicar més temps a formar-se, fet que de manera intrínseca farà millorar el sistema. Aquest enfocament afavorirà pràctiques més qualitatives en la gestió documental i administrativa.

* Ètic-social:

El desenvolupament d’aquesta eina ha de complir rigorosament amb els estàndards ètics i de seguretat, per sobre de tots en l’àmbit de protecció de dades personals. Per això es garanteix el compliment del reglament general de protecció de dades (RGPD), assegurant que les dades tractades siguin tractades amb la màxima confidencialitat i seguretat, important dir que després del anàlisis de cada cas aquesta informació mai es guardarà dins el model. També cal tenir en compte un dels riscos més grans, al tractar-se d’una eina de (ML), pot induir a la falsa predicció de codis diagnòstics degut a biaixos en les dades d’entrada, de manera que sols servirà com una ajuda a la codificació i mai com a sistema autònom. (1)

* Diversitat:

En el context sanitari i la zona geogràfica on ens trobem l’eina ha de reconèixer i adaptar-se a les variabilitats lingüístiques, culturals i regionals. Aquesta adaptabilitat garantirà que la solució sigui inclusiva i aplicable a tots els professionals independentment de la llengut utilitzada.

En definitiva, el projecte busca una gestió més sostenible dels recursos, un tractament ètic i segur de la informació i la promoció d’una pràctica inclusiva que té en compte la diversitat dins l’empresa.

## Enfocament i mètode seguit

L’enfocament adoptat per al desenvolupament és basa en una gestió integral del projecte, ja que es tracta d’un projecte nou des de zero. Utilitzarem una metodologia àgil basada en Scrum, de manera que s’aniran realitzat entregues parcials rebent comentaris i propostes de millora per part de la tutora del treball i aplicant les modificacions en cadascuna de les iteracions.

Utilitzarem part de la metodologia Srum:

* Sprints curts.

El projecte es dividirà en cicles curs de treball, cadascun amb objectius clars i definits. Al final de cada cicle s’avaluaran els resultats i es realitzarà l’ajust sobre la planificació.

* Revisions i retrospectives

Cada cicle conclourà amb una revisió per tal de valorar les millores i els inconvenients que vagin apareixent.

Estratègia de recerca:

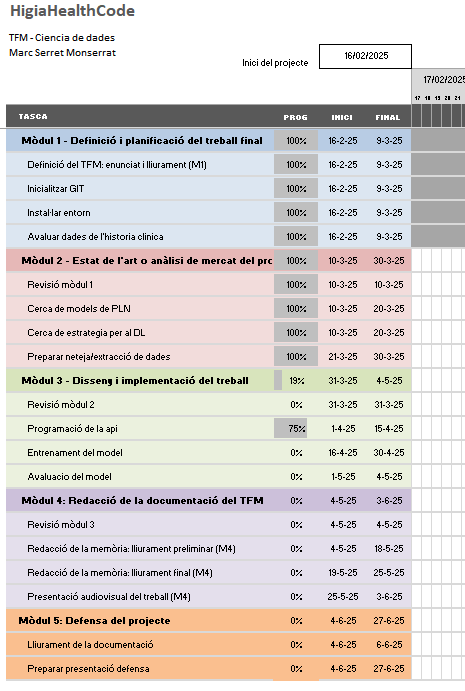
Es fonamenta en utilitzar una base sòlida en la teoria amb la finalitat de desenvolupar una aplicació completa.

* Revisió de documentació sistemàtica

Es realitzarà una revisió de documentació contínuament durant el desenvolupament del projecte, te com a finalitat la cerca de la millor estratègia per a desenvolupar les eines basades en llenguatge natural i *deep learning.*

* S’avaluarà el model utilitzant *train-test split amb hold-out validation*, com a metodologia inicial, però amb un enfocament dinàmic i iteratius per millorar contínuament el rendiment del model. El model inclourà la reintroducció dels casos validats dins del model perquè aquest pugui aprendre progressivament i adaptar-se als nous patrons.

## Planificació del Treball



## Breu sumari de productes obtinguts

El projecte generarà els següents productes.

* Model *deep learning* entrenat:

Es desenvoluparà i entrenarà un model de *deep learing* basat en arquitectures de xarxes neuronals que serà capaç de processar i analitzar textos clínics per assignar codis CIM-10 amb un alt grau de precisió.

* Validacions i avaluacions del model.

Es realitzarà un estudi que realitzarà una validació comparant els codis generats amb els codificats per un expert, utilitzant mètriques com la precisió, el *recall* i els F1-score.

* Documentació tècnica i manuals d’usuari.

Elaboració d’una documentació tècnica per a la implementació del programari.

## Breu descripció dels altres capítols de la memòria

* Materials i mètodes:

Aquest capítol descriu de manera detallada la metodologia emprada en el desenvolupament del treball. L’enfocament s’ha centrat en tres àrees clau: la gestió i pre-processament de les variables, la implementació dels models de processament de llenguatge natural (PLN) i del disseny del model de *deep learning.*

* Resultats

Aquest apartat presenta l’anàlisi de resultats obtinguts després de l’entrenament i validació del model. S’explica com es compara els codis generats automàticament amb la codificació manual realitzada per tècnics en documentació clínica, mitjançant diverses mètriques.

* Conclusions

Finalment es resumeixen les conclusions obtingudes del projecte. A més es proposen línies futures de recerca i millores relaciones amb el ràpid avanç d’aquestes tecnologies.

# Base teòrica i fonaments

### 2.1 Sistema d’Informació Sanitari i la Gestió de les dades.

En l’actualitat, els sistemes d’informació sanitari juguen un paper fonamental en la presa de decisions clíniques i en la gestió administrativa. La complexitat i el volum d’informació generada en aquests entorns requereixen de processos robustos per l’emmagatzematge, gestió i anàlisis, els quals són assolits mitjançant Data Warehouses que es nodreixen mitjançant, sistemes d’extracció, transformació i carrega.

Un Data Warehouse és un sistema d’emmagatzematge de dades centralitzat, ens permet consolidar dades que provenen de diversos fonts o sistemes, transformant-ho en un format homogeni que facilitat el seu anàlisis.

Els sistemes d’informació sanitari han de gestionar dades que provenen de fonts molt diverses, així la integració de dades estructurades i no estructurades és un repte clau. L’ús de ETL és fonamental per transformar dades de diferents formats en un conjunt homogeni i coherent, capaç de donar suport en el anàlisis de dades. Aquest procés pot incloure la neteja de dades, la normalització de formats i la validació de la informació.

Una altra consideració important en la gestió de dades sanitàries és la seguretat i la privacitat. Donat que les dades contenen informació sensible dels pacients, és essencial aplicar tècniques d’anonimització i encriptació per complir amb les normatives de la RGPD.

En resum, la gestió de les dades en el sector sanitari requereix d’un enfocament integral que combini tecnologies d’emmagatzematge amb processos rigorosos per a l’extracció de les dades sempre mantenint totes les mesures de seguretat per complir amb la normativa de seguretat i privacitat d’aquestes.

En el cas del projecte actual es farà servir d’un DWH que ja esta implementat dins la Xarxa Sanitaria, Social i Docent de Santa Tecla i que ja conte les dades carregades prèviament mitjançant processos de ETL i que compleix amb la normativa referent a la llei de protecció de dades.

### 2.2 Processament del Llenguatge Natural

Es una disciplina informàtica que s’encarrega de tractar computacionalment les llengües, combina tècniques de intel·ligència artificial, lingüística i estadística per permetre que les maquines comprenguin, analitzin i generin text en llenguatge humà.

Per entendre com funciona un sistema de PLN podem definir 3 fases.

1. Pre-processament del text
   1. Tokenització: Consisteix en dividir el text original en unitats més petites, com ara paraules o frases, facilitant-ne la manipulació posterior.
   2. Normalització: Aquesta esta inclou processos com la conversió a minúscules, eliminació de signes de puntuació i altres transformacions que homogenitzen el text.
   3. Eliminar caràcters sense carrega semàntica, es realitza la supressió de paraules habituals com “el”, “de”, que no aporten informació significativa per la anàlisis.
   4. Lematització/stemming; Es redueixen les paraules a la seva forma base o arrel, facilitant l’agrupació de termes semànticament similars.
2. Representació vectorial: En aquesta face, el text Preprocessat es transforma en una representació numèrica (vectors), imprescindible perquè pugui ser interpretat per models d’aprenentatge automàtic. Aquesta conversió habitualment es realitza mitjançant tècniques que tenen en compte l’entorn, basant-se amb el seu context.
3. Modelatge del llenguatge. Amb els vectors d’entrada ja disponibles, un model de ML o DL s’encarrega d’entendre el context i generar una resposta, classificació, predicció.

Un exemple d’aquest models és BERT que ha suposat un gran canvi dins els món del PLN. Aquest utilitza una arquitectura basada en “transformers” amb mecanismes d’autoatenció, capaços d’analitzar el context complet d’una paraula dins d’una frase, millorant notablement respecte models anteriors, *chatGPT* ha fet servir models basats en “transformers”.

Malgrat els avantatges de BERT, aquest model presenta dues limitacions per al nostre projecte.

* Té una capacitat limitada per processar seqüencies llargues, amb un màxim de 512 claus, insuficients per a textos clínics extensos.
* Esta entrenat amb textos generals, es ha dir wikipedia i llibres, fet que limita la seva efectivitat amb textos altament especialitzats com els clínics.

Per superar aquesta limitació, s’ha seleccionat el model Clinical Longformer(3,4), específicament dissenyat i preentrenat amb textos clínics reals. Aquest model ofereix:

* La capacitat d’analitzar seqüencies més llargues, de fins a 4096 claus, sens especialment adequat per a documents clínics extensos.
* Un entrenament específica amb terminologia mèdica, millorant considerablement la seva eficàcia en els nostre context.

Aquesta elecció creec que garanteix la correcta interpretació dels textos amb la finalitat de realitzar una codificació clínica automàtica basada en el processament del llenguatge natural.

### 2.3 Deep Learning

# Materials i mètodes

Aquest apartat descriu les dades que utilitzarem així com les tecnologies aplicades, per al desenvolupament de l’eina.

### 3.1 Tecnologies utilitzades

* PostgreSQL: Utilitzat com a base de dades, permet l’emmagatzematge i consulta eficient de les dades històriques així com guardar les estadístiques de rendiment del model.
* Python: La principal eina per l’anàlisi i processament de dades, així com per al desenvolupament del model de *deep learning*. Amb biblioteques com Pandas, NumPy, Pytorch.
* API (RESTful): Implementada per a gestionar les consultes en temps real, Aquesta interfície s’encarregarà d’executar els mètodes.
* GIT

### 3.2 Font de dades

L’origen de tot el projecte es basa en la qualitat de les dades utilitzades per al seu entrenament. En aquest projecte les dades poden arribar de dos orígens:

* + Data Warehouse (DWH) de l’empresa. Utilitzada com a font principal per a les fases d’entrenament del model, s’ha preparat una taula amb les dades històriques ja tractades i optimitzada per l’estudi, en aquesta mateixa taula ja s’ha realitzat l’aleatorització per al conjunt d’entrenament, test i validació.
  + API amb missatges JSON: Utilitzada com a funcionalitat operativa per a la consulta i codificació en temps real.
    1. Data Warehouse

El DWH, esta implementat en PostgreSQL, serveix com la base de dades d’origen per a la creació del model. Aquest magatzem conté informació que te l’origen en la mateixa historia clínica que s’utilitza en tota la organització. Les dades les podem agrupar en informació estructurada i informació no estructurada.

* + - Informació clínica estructurada:

Dins de les dades estructurades podem trobar els codis diagnòstics, edat, gènere i any de l’activitat

* + - Informació clínica no estructurada:

Aquí trobarem un conjunt de textos lliures escrits per els diversos professionals de la salut que és la base fonamental del projecte que estic realitzant, antecedents, motiu d’ingrés, recomanacions de tractament i el curs clínic.

Amb la finalitat de reproduir un entorn real, les dades no s’han netejat prèviament intentant simular la funcionalitat en un entorn real, serà la API la que implementarà mètodes de neteja.

* + 1. La API

Encara que el projecte es basa en el desenvolupament d’un model de predicció de dades, cal tenir en compte que el que busquem es una eina que ens permeti codificar a temps real les altes dels pacients i que sigui fàcil de implementar en qualsevol sistema.

Les dades de les dos entrades han de mantenir la mateixa estructura.

# Resultats

Detalleu en aquest apartat els resultats obtinguts utilitzant la metodologia descrita a l’apartat anterior.

# Conclusions i treballs futurs

Aquest capítol ha d’incloure:

* Una descripció de les conclusions del treball:
  + Un cop s’han obtingut els resultats quines conclusions s’extreu?
  + Aquests resultats són els esperats? O han estat sorprenents? Per què?
* Una reflexió crítica sobre l’assoliment dels objectius plantejats inicialment:
  + Hem assolit tots els objectius? Si la resposta és negativa, per quin motiu?
* Una anàlisi crítica del seguiment de la planificació i metodologia al llarg del producte:
  + S’ha seguit la planificació?
  + La metodologia prevista ha estat prou adequada?
  + Ha calgut introduir canvis per garantir l’èxit del treball? Per què?
* Dels impactes previstos a 1.3 (ètic-socials, de sostenibilitat i de diversitat), avaluar/esmentar si s'han mitigat (si eren negatius) o si s'han aconseguit (si eren positius).
* Si han aparegut impactes no previstos a 1.3, avaluar/esmentar com s'han mitigat (si eren negatius) o què han aportat (si eren positius).
* Les línies de treball futur que no s’han pogut explorar en aquest treball i han quedat pendents.

# Glossari

BERT - Bidirectional Encoder Representations from Transformers

DL – Deep Learning  
DWH – Data Warehouse.

ML – Machine Learning.

PLN – Processament de llenguatge natural

RGPD – Reglament General de Protecció de Dades.

# Bibliografia

1. Deep Learning Deep ethics: Ètica per a l’ús de la intel·ligència artificial en medicina | Institut Borja de Bioètica [Internet]. [citado 18 de marzo de 2025]. Disponible en: https://www.iborjabioetica.url.edu/ca/blog-de-bioetica-debat/deep-learning-deep-ethics-etica-lus-de-la-intelligencia-artificial-en-medicina

2. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

3. Beltagy I, Peters ME, Cohan A. Longformer: The Long-Document Transformer [Internet]. arXiv; 2020 [citado 29 de marzo de 2025]. Disponible en: http://arxiv.org/abs/2004.05150

4. Tinn R, Cheng H, Gu Y, Usuyama N, Liu X, Naumann T, et al. Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing [Internet]. arXiv; 2021 [citado 29 de marzo de 2025]. Disponible en: http://arxiv.org/abs/2112.07869

# Annexos

Llistat d’apartats que són massa extensos per incloure dins la memòria i tenen un caràcter autocontingut (per exemple, manuals d’usuari, manuals d’instal·lació, etc.)

Depenent del tipus de treball, és possible que no calgui afegir cap annex.