# 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


Universitat Oberta de Catalunya

Universitat Oberta de Catalunya

uoc.edu

# 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) |

**Comentado [MS1]:** A la mateixa plantilla suggerien un màxim de 3 paraules per això sols he utilitzat el mínim indispensable.

**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

## 1.1. 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.

## 1.2. 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.

## 1.3. 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.

## 1.4. 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.

## 1.5. Planificació del Treball

**HigiaHealthCode**

TFM - Ciencia de dades
Marc Serret Monserrat

Inici del projecte: 16/02/2025

| TASCA | PROG | INICI | FINAL |
|---|---|---|---|
| **Mòdul 1 - Definició i planificació del treball final** | 100% | 16-2-25 | 9-3-25 |
| Definició del TFM: enunciat i lliurament (M1) | 100% | 16-2-25 | 9-3-25 |
| Inicialitzar GIT | 100% | 16-2-25 | 9-3-25 |
| Instal·lar entorn | 100% | 16-2-25 | 9-3-25 |
| Avaluar dades de l'historia clinica | 100% | 16-2-25 | 9-3-25 |
| **Mòdul 2 - Estat de l'art o anàlisi de mercat del prc** | 100% | 10-3-25 | 30-3-25 |
| Revisió mòdul 1 | 100% | 10-3-25 | 10-3-25 |
| Cerca de models de PLN | 100% | 10-3-25 | 20-3-25 |
| Cerca de estrategia per al DL | 100% | 10-3-25 | 20-3-25 |
| Preparar neteja/extracció de dades | 100% | 21-3-25 | 30-3-25 |
| **Mòdul 3 - Disseny i implementació del treball** | 19% | 31-3-25 | 4-5-25 |
| Revisió mòdul 2 | 0% | 31-3-25 | 31-3-25 |
| Programació de la api | 75% | 1-4-25 | 15-4-25 |
| Entrenament del model | 0% | 16-4-25 | 30-4-25 |
| Avaluacio del model | 0% | 1-5-25 | 4-5-25 |
| **Mòdul 4: Redacció de la documentació del TFM** | 0% | 4-5-25 | 3-6-25 |
| Revisió mòdul 3 | 0% | 4-5-25 | 4-5-25 |
| Redacció de la memòria: lliurament preliminar (M4) | 0% | 4-5-25 | 18-5-25 |
| Redacció de la memòria: lliurament final (M4) | 0% | 19-5-25 | 25-5-25 |
| Presentació audiovisual del treball (M4) | 0% | 25-5-25 | 3-6-25 |
| **Mòdul 5: Defensa del projecte** | 0% | 4-6-25 | 27-6-25 |
| Lliurament de la documentació | 0% | 4-6-25 | 6-6-25 |
| Preparar presentació defensa | 0% | 4-6-25 | 27-6-25 |

## 1.6. 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.

## 1.7. 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.

# 2. 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
   a. Tokenització: Consisteix en dividir el text original en unitats més petites, com ara paraules o frases, facilitant-ne la manipulació posterior.
   b. Normalització: Aquesta esta inclou processos com la conversió a minúscules, eliminació de signes de puntuació i altres transformacions que homogenitzen el text.
   c. Eliminar caràcters sense carrega semàntica, es realitza la supressió de paraules habituals com "el", "de", que no aporten informació significativa per la anàlisis.
   d. 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

És una branca del ML, que utilitza xarxes neuronals amb múltiples capes (arquitectures profundes) per aprendre patrons complexos en grans conjunts de dades. Aquesta xarxes neuronals profundes estan formades per múltiples capes d'unitat de processament (neurones) que poden detectar estrictes complexes i no lineals en les dades, fent-les particularment eficaces per a tasques d'alt nivell com el reconeixement de llenguatge natural, la classificació d'imatges o la predicció de sèries temporals.

En el context del projecte actual, s'aplicarà el Deelp Learning mitjançant l'ús específic de PyTorch, una biblioteca de codi obert.

La implementació del Deep Learning dins del nostre projecte es basarà en la capacitat del model Clinical Longformer per generar representacions vectorial de textos clínics. Aquest vectors numèrics seran la base d'entrada per al nostre model de xarxa neuronal profunda, implementat amb PyTorch, que s'encarregarà específicament de classifica automàticament els codis diagnòstics associats als informes clínics. Aquest enfocament busca obtenir una alta precisió en la codificació dels diagnòstics, contribuint així a la millorar de l'eficiència i la qualitat en processos clínics automatitzats.

## 3. Materials i mètodes

Aquest apartat descriu de manera detallada el conjunt d'eines tecnologies i estratègies metodològiques utilitzades per al desenvolupament del projecte. El treball s'ha estructurat seguint una arquitectura modular, que va des de la gestió de les dades i la seva extracció de la historia clínica, fins a la seva transformació i anàlisi mitjançant un model pre entrenat. A més s'ha desenvolupat una API amb la finalitat de garantir la integració amb sistemes clínics en existents.
En els següents apartats s'exposen les tecnologies utilitzades, les fonts de dades, els processos per pre processament, l'arquitectura del model, les estratègies d'entrenament i validació, i la seva integració operativa en entorns reals.

### 3.1 Tecnologies utilitzades

Per a la fase inicial d'extracció i preparació de dades, s'ha fet ús de l'eina Spoon del paquet de Pentaho Data Integratia, una eina d'ETL(Extracció, Transformació i Carrega) visual que ha permès construir fluxos de dades de manera modular. Mitjançant Spoon, s'ha automatitzat l'obtenció de les dades dels diversos sistemes, les dades clíniques s'han agafat d'un SQL Server, mentre que les dades de la codificació s'han extret de la codificació del CMBD (Conjunt Minim Basic de Dades) que es troben en un altre sistema, aquesta segmentació ha requerit transformacions específiques per garantir la compatibilitat i la integritat de la informació abans d'incorporar-la al Date Warehouse.

A continuació un cop hem tingut les dades aïllades, el desenvolupament de l'eina s'ha centrat en les tecnologies següents.

- PostgreSQL: S'utilitza com a base de dades relacional principal per a l'emmagatzematge i consulta de dades històriques. Es seu ús esta justificat per la seva estabilitat, suport per a consultes complexes i integració amb altres eines analítiques. El DWH conté taules optimitzada amb informació clínica estructurada i no estructurada, ja pre processades per a l'estudi, incloent una columna per diferenciar els diversos conjunts.
- Python: És el llenguatge principal emprat per a la construcció del sistema. Permet la integració fluida de biblioteques especialitzades en tractament de dades, processament de text i aprenentatge profund. Les biblioteques claus són:
  - Pandas i NumPy per a la manipulació de dades
  - Scikit-learn per a transformacions i mètriques d'avaluació
  - Transformers de Huggins Face per accedir al model Clinical Longformer.
  - PyTorch com a fons per a la implantació del model de deep learning, optimitzat per entrenament.
- FastAPI(API RESTful): El sistema utilitza un interfície REST dissenyada amb Fast API, que permet consultes en temps real per part del sistema d'història clínica. Aquesta API està preparada per acceptar dades en format JSON, processar-les mitjançant el model i retornar les prediccions de codis CIM-10.
- GIT: El control de versions es duu a terme mitjançant Git, assegurant traçabilitat i replicabilitat del codi font i facilitant el treball incremental amb diverses etapes de millorar del model.

## 3.2 Font de dades

Les dades utilitzades en el projectes constitueixen un actiu fonamental per al desenvolupament i entrenament del sistema de codificació automàtica. Aquestes dades poden provenir de dues fonts principals. El Data Warehouse corporatiu i els sistemes operatius connectats mitjançant l'API per la interacció a temps real.

### 3.2.1. Data Warehouse (DWH).

El DWH, construït en PostgreSQL, integra la informació provinent de diversos sistemes assistencials de la Xarxa Sanitaria Social i Docent de Santa Tecla. Conté tant dades clíniques estructurades com textos lliures en forma de camps no estructures, extrets directament de l'historia clínica de l'organització (Higia HC).

Les dades han estat organitzades en una taual específica optimitzada per a l'entrenament, validació i prova del model. Aquesta taula inclou:

- Dades estructurades: edat, sexe, codis diagnòstics, tipus d'alta, any d'activitat, servei.
- Dades no estructurades: motiu d'ingrés, malaltia actua, exploració , proves complementaries a l'ingrés, proves complementaries, evolució clínica, antecedents, curs clínic complet.

Aquest conjunt ha estat prèviament aleatoritzat i dividit mitjançant ETL, per garantir la separació adequada.

3.2.2. Entrades en temps real via API.

Per a la fase operativa del projecte, el sistema haurà d'estar preparat per acceptar consultes externes a través una API RESTful. Aquesta interfície permet rebre dades en format JSON, estructurades de manera idèntica a les utilitzades en fase d'entrenament , per garantir la compatibilitat i coherència dels resultats.

D'aquesta manera l'aplicació es podria cridar des de diversos mòduls, serviria per realitzar la codificació inicial del professional de manera automàtica fent que els codificadors sols tinguessin de comprovar la validesa, juntament amb el percentatge de validesa de cadascun dels codis.

Aquesta dualitat de fonts (històrica per entrenament i temps real ) permet validar el model en condicions de producció i facilitat la seva integració dins de qualsevols historia clínica o sistema de codificació extern.

3.3. Pre-processament i transformació dels textos clínics.

És una fase crítica del projecte, ja que té com objectiu preparar la informació no estructurada per tal que pugui ser interpretada per els models. Per tal de netejar aquestes aplicarem els següents processos.

- Normalització a unicode: Convertim el text a codificació utf-8 i eliminem caràcters especials.
- Eliminem marques HTML: Utilitzarem una llibreria de *Python* anomenada *BeautifulSoup* per eliminar possibles etiquetes HTML incrustades en els textos clínics.
- Conversió a minúscules: Unifiquem el text per evitar distorsions durant la tokenització.
- Eliminació de puntuació i caràcters especials no informatius, mitjançant expressions regulars, es filtren símbols i signes que no aporten valor semàntic.
- Eliminació d'espais redundants: Es redueixen múltiples espais consecutius i s'eliminen els espais en blanc per l'esquerra i la dreta.
- Eliminació de paraules que no aporten contingut semàntic. Utilitzo una llibreria que aporta totes aquestes paraules del castellà i una llista manual de les de català.

Aquest processament esta integrat dins la API de manera que independentment de l'origen les dades es tractaran utilitzant la mateixa lògica.

3.4. Arquitectura del model.

El model utilitzat ha estat dissenyat específicament per abordar la classificació múltiple de codis CIE-10 a partir de informes cínics en llenguatge natural. La seva arquitectura es basa en un model preentrenat "Clinical Longformer", personalitzat mitjançant un model de deep learning, que és invocat des d'un mòdul principal, que es el que gestiona la carrega, validació i predicció.

Les característiques tècniques clau són les següents:

3.4.1. Model base – Long Former: Com ja hem dit s'ha escollit per la seva capacitat de processar seqüencies de fins a 4096 claus.

3.4.2. Capçalera dual de classificació: Es tant important codificar els codis com l'ordre d'aquests.
- Classificació de codis: genera una predicció multi etiqueta, ja que cada cas pot tenir fins a 15 codis diagnòstics.
- Ordre dels codis: Classifica els codis segons la seva rellevància clínica. Aquesta funcionalitat s'utilitza per prioritza les prediccions, donant més pes al codis més importants.

3.4.3. Variables categòriques estructurades: Les metadades de cada cas, com són el gènere , el tipus d'alta, el servei mèdic són essencials en la codificació. El model incorpora aquestes dades i les combina amb les variables no estructurades. Un exemple de la importància d'aquests es que hi ha diversos codis que estan limitats per edat, es ha dir una persona de 90 anys no pot tenir diabetis gestacional.

3.4.4. Funció de pèrdua i optimització:
- S'utilitza *BCEWithLogitsLoss*, es una funció idònia per a la classificació multi-etiqueta amb sortides independents per cada codi, integra la *sigmoide* i la *cross-entropy* en una sola operació, millorant l'estabilitat numèrica i el rendiment en la classificació multi etiqueta.
- Es calculen pesos per classe segons la seva freqüència en el conjunt d'entrenament, per compensar la presència desigual de codis, aquesta ponderació es molt important per tal de millorar la sensibilitat del model envers diagnòstics menys freqüents.
- Sols utilitzarà els models prescrits en algun moment, el catàleg ens permet fins a 98000 codis diferents, no obstant en un entorn real rarament s'utilitzen tots.
- Es fa servir el classificador AdamW, una versió millorada de l'optimitzador Adam que separa explícitament la regularització per caiguda de pes. Aquest paràmetre penalitza els pesos grans durant l'entrenament, afavorint solucions més simples i generalitzades, es una mesura per prevenir el sobre entrenament.
- Com a planificador de decreixement de taxa d'aprenentatge utilitzarem *StepLr,* cada època aquest disminueix la taxa per fer un model més estable.
- També com a mesura d'optimització s'ha implementat una eina de parada, si no es detectat una millora significativa en les mètriques de validació en mes de 5 èpoques el sistema pot interrompre per evitar el sobre entrenament i reduir els temps de càlcul.
- S'ha implementat la compatibilitat amb la tecnologia *cuda* de nvidia.

**Comentado [MS2]:** No ho he pogut provar per falta de compatibilitat entre la meva gpu i pytorch

3.4.5. Entrenament incremental cas a cas: Aquesta estratègia permet re entrenar el model amb cada nou cas validat, fet que afavoreix l'aprenentatge continu. Cada cas es procés individualment durant 20 èpoques per millorar la seva incorporació al model sense comprometre la seva estabilitat global.

3.4.6. Persistència i control de versió: El sistema guarda l'estat del model, l'optimitzador, el planificador d'aprenentatge i la llista de codis predit

mitjançant arxius. Aquest gestió garanteix la capacitat de reprendre l'entrenament, fer auditoria del rendiment i torna a començar l'entrenament en cas de fer modificacions.

### 3.5. Entrenament, validació i mètriques.

L'estratègia adoptada es basa en una aproximació incremental i adaptativa El model no s'entrenarà sols amb un conjunt de dades massiu i estàtic, sinó que incorporarà nous casos clínics validats de manera contínua. Això permet mantenir una actualització dinàmica del coneixement del sistema i adaptar-se a l'evolució dels patrons clínics i dels criteris de codificació.

Entrenament incremental:

- Cada nou cas clínic validat és utilitzat per entrenar el model durant un cicle curt de 20 èpoques, permetent una ràpida assimilació d'informació sense comprometre el rendiment global.
- Es fa servir una rutina de parada cada 5 èpoques per detectar si el model millorar significativament durant l'entrenament. En cas contrari, s'atura anticipadament per evitar sobre ajustaments i accelerar el procés.
- El sistema actualitzarà progressivament el conjunt de codis coneguts, ampliant les seves capacitats de predicció a mesura que apareixen noves etiquetes.

Validació continua:

- En qualsevols moment es pot seleccionar un conjunt de dades per tal de validar la capacitat predictiva del model.
- A més de predir els codis amb un llindar de confiança es calculen els codis més probables i es compara amb l'odre real esperat.
- S'utilitza l'índex de Kendall-Tau i l'Accuracy per tal de avaluar la coherència entre els codis, i també l'ordre d'aparició d'aquests.

**Figure 4.** Definitions of Kendall tau

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (1)$$

where $n$ represents the sample size,
$n_c$ is the number of concordant pairs,
$n_d$ is the number of discordant pairs.

*1Font(5)*

Mètriques utilitzades:

- Precisió(*Precision*): per mesurar la proporció de codis predits correctament.
- Sensibilitat(*Recall*): Per detectar fins a qquin punt el model identifica tots els codis reals.
- F1-Score: com a compromís entre precisió i sensibilitat, especialemtn útil en entorns amb codis molt desequilibrats.
- Exactitud (*Accuracy*): Útil per mesurar el grau d'encert global del model.

- Order Accuracy i Kendall-Tau: Avaluent la qualitat de l'ordre dels codis predits en relació amb l'ordre clínic establet pel professional.
- Pèruda (Loss): Es mostren separadament la pèrdua de classificació i la d'ordre, donant visibilitat de la contribució de cada component.

Aquestes mètriques s'usen tant per monitorar l'evolució del model com per predir quan revisar, reajusta o reforçar el seu entrenament amb noves dades.

```
INFO:app.ml.engine:Conjunt de codis entrenats carregat correctament. Total: 5
INFO:app.ml.engine:Codis disponibles per validació: ['E119', 'I680', 'I69', 'O89', 'O90']
Input ids are automatically padded from 11 to 512 to be a multiple of `config.attention_window`: 512
INFO:app.ml.engine:=== VALIDACIÓ CAS: test5 ===
INFO:app.ml.engine:→ Codis reals (2):
INFO:app.ml.engine:   1. I680
INFO:app.ml.engine:   2. E119
INFO:app.ml.engine:→ Codis predits (>90% confiança):
INFO:app.ml.engine:   1. I69 (100.0%)
INFO:app.ml.engine:→ Top 5 codis més probables (entrenats):
INFO:app.ml.engine:   1. I69 (100.0%)
INFO:app.ml.engine:   2. O90 (0.0%)
INFO:app.ml.engine:   3. E119 (0.0%)
INFO:app.ml.engine:   4. I680 (0.0%)
INFO:app.ml.engine:   5. O89 (0.0%)
INFO:app.ml.engine:→ Ordre real:        I680 → E119
INFO:app.ml.engine:→ Ordre predit:      I680 → E119
INFO:app.ml.engine:→ Kendall-Tau:       1.00 → L'ordre predit és molt similar al real
INFO:app.ml.engine:→ Mètriques de Classificació:
INFO:app.ml.engine:   • Accuracy:  0.00
INFO:app.ml.engine:   • Precision: 0.00
INFO:app.ml.engine:   • Recall:    0.00
INFO:app.ml.engine:   • F1 Score:  0.00
INFO:app.ml.engine:→ Mètriques d'Ordre:
INFO:app.ml.engine:   • Order Accuracy: 1.00
INFO:app.ml.engine:   • Kendall-Tau:    1.00
INFO:app.ml.engine:→ Pèrdues:
INFO:app.ml.engine:   • Code Loss:  0.0003
INFO:app.ml.engine:   • Order Loss: 8.9108
```

*Il·lustració 1. Exemple de validació en un conjunt de proves*

### 3.4.7. Consideracions de seguretat i anonimització.

Donat que aquest projecte es desenvolupa en l'àmbit clínic i fa ús de dades sensibles, s'ha posat especial atenció a garantir el compliment de la RGPD. Totes les dades utilitzades en el procés d'entrenament i validació han estat prèviament anonimitzades per evitar qualsevol risc de re identificació dels pacients. Sols hem mantingut el codi del cas per identificar el cas que tractarem i inclús així, no s'utilitza per fer l'entrenament i no queda guardat en cap lloc del programa.

L'aplicació es desplegarà com una API, per el que es poden integrar mesures de protecció i autenticació mitjançant claus i/o ip.

Aquesta capa d'integració assegura que el model pot ser utilitzat en entorns clínics reals, mantenint el compromís amb la seguretat.

# 4. Desenvolupament del model

No se ven be com plantejar aquest apartat, tinc app que funciona amb la versió directament amb base de dades,(sense JSON) no obstant encara no he pogut provar en dades reals, estic esperant a l'actualització de l'equip informàtic ja que en la maquina que disposo actualment es impossible realitzar l'entrenament en l'entorn real.

# 5. Anàlisis de resultats

Fins que no tingui dades reals no m'atreveixo, tampoc se ven be com plantejar l'estudi, aproximadament disposo de 100k casos per realitzar l'entrenament però no se com plantejar el resultat o si podré fer l'entrenament complet abans de que acabi el període d'entrega. Estic pensant en alguna llibreria per monitoritzar l'app a temps real, pero no ho acabo de veure clar.

# 6. 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.

# 7. 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.

# 8. 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

5.Kim M, Jung Y, Jung D, Hur C. Investigating the Congruence of Crowdsourced Information With Official Government Data: The Case of Pediatric Clinics. J Med Internet Res. 3 de febrero de 2014;16(2):e29.

# 9. Annexos

Git-Hub - https://github.com/mserretm/HigiaHealthCode

He obert el GIT de manera publica per tal de si li vols donar una ullada.