Universitat Autònoma
de Barcelona

Treball Final de Grau

Grau en Matemàtica Computacional i Anàlisi de Dades

Analitzant malalties neurològiques des de la perspectiva de la teoria de grafs

Judit Yebra Valencia

Supervisor
Jordi Casas-Roma i Carlos Boned Riera

Any
2024

Convocatòria
Septembre

To my colleagues,

...

Abstract

This project has the primary objective of investigating the alterations in brain connectivity among patients diagnosed with multiple sclerosis (MS) and how these changes correspond to the cognitive condition of the patients. The research will use magnetic resonance imaging (MRI) data obtained from the Hospital Clínic de Barcelona, and afterwards also data obtained from Naples, Italy, to analyze the structural connectivity of the brain in both MS patients and healthy individuals which will be referred as patients and control.

The motivation behind this project stems from the potential to identify distinct patterns of brain connectivity associated with MS, which could facilitate the monitoring of each patient's condition and potentially predict the progression of the disease on an individual level. The proposed methodology involves employing graph theory to model and examine the MRI images. What this approach aims to ascertain is whether there are specific connectivity patterns linked to the disease or not.

As added in the bibliography, the project idea is supported by three scientific articles that have addressed similar issues using various techniques, including topological data analysis, multiplex networks and classification utilizing support vector machines (SVM), with the help of graph theory. This project aims to expand upon and apply some of these methodologies within the context of multiple sclerosis.

Resum

Aquest projecte té com a objectiu principal investigar les alteracions en la connectivitat cerebral entre els pacients diagnosticats amb esclerosi múltiple (EM) i com aquests canvis corresponen a la condició cognitiva dels pacients. La recerca utilitzarà dades de ressonància magnètica (IRM) obtingudes de l'Hospital Clínic de Barcelona i posteriorment també de Nàpols, Itàlia, per analitzar la connectivitat estructural del cervell tant en pacients amb EM com en individus sans, que seran referits com a pacients i control.

La motivació darrere d'aquest projecte prové del potencial per identificar patrons distints de connectivitat cerebral associats amb l'EM, el que podria facilitar el seguiment de la condició de cada pacient i potencialment predir la progressió de la malaltia a nivell individual. La metodologia proposada implica l'ús de la teoria de grafs per modelar i examinar les imatges d'IRM. El que aquest enfocament pretén assolir és determinar si hi ha patrons de connectivitat específics vinculats a la malaltia o no.

Com s'ha afegit a la bibliografia, la idea del projecte està suportada per tres articles científics que han tractat temes similars utilitzant diverses tècniques, incloent l'anàlisi de dades topològiques, xarxes múltiplex i la classificació utilitzant màquines de vector de suport (SVM), amb l'ajuda de la teoria de grafs. Aquest projecte té com a objectiu expandir i aplicar algunes d'aquestes metodologies dins del context de l'esclerosi múltiple.

Preface

Aquest treball de final de grau es centra en l'estudi de les connexions cerebrals en pacients amb esclerosi múltiple (EM) i la seva relació amb les capacitats cognitives d'aquests pacients. Mitjançant l'anàlisi de dades de ressonància magnètica obtingudes a l'Hospital Clínic de Barcelona i a un hospital de Nàpols, es pretén identificar patrons de connectivitat cerebral que puguin diferenciar els pacients d'individus sans.

L'interès en aquest projecte neix de la possibilitat de trobar noves formes de comprendre l'impacte de l'EM en el cervell, utilitzant eines com la teoria de grafs per modelar i analitzar la complexitat de les dades obtingudes. Aquest enfocament ofereix una perspectiva única que podria contribuir significativament al camp de la neurologia.

El treball es divideix en diverses parts, que inclouen una revisió de la literatura existent, la descripció de la metodologia emprada, l'anàlisi dels resultats i la discussió de les conclusions. L'ús de tècniques com la teoria de grafs, la classificació amb màquines de vector de suport (SVM) i la regressió logística permet aprofundir en l'estudi dels canvis estructurals en el cervell relacionats amb l'EM. Posteriorment s'ha fet també ús d'*embeddings* i de *graph neural networks*.

Vull expressar el meu agraïment als meus tutors, Jordi Casas-Roma i Carlos Boned Riera, pel seu suport i guia durant tot aquest procés, així com a totes les persones que m'han ajudat d'una manera o altra. Aquest projecte representa la culminació del meu esforç acadèmic i espero que reflecteixi l'aprenentatge d'aquests anys.

Barcelona, Septembre 2024

Judit Yebra Valencia

Contents

Chapter 1

Introducció i marc teòric

1.1 Context del treball

1.1.1 Esclerosi múltiple

L'esclerosi múltiple (EM) és una malaltia neurològica crònica que afecta el sistema nerviós central, especialment el cervell i la medul·la espinal. Aquesta malaltia es caracteritza per la desmielinització, un procés on la mielina, la capa protectora que envolta les fibres nervioses, és danyada. Això provoca una interrupció en la comunicació entre el cervell i la resta del cos, la qual cosa pot conduir a una àmplia gamma de símptomes, incloent-hi problemes motors, visuals, sensorials i cognitius.

Els símptomes de l'esclerosi múltiple poden variar àmpliament entre els individus i poden incloure fatiga, dificultat per caminar, problemes de visió, espasticitat muscular, problemes d'equilibri i coordinació, dolor, i problemes cognitius i emocionals. Aquests símptomes poden aparèixer i desaparèixer en períodes coneguts com a brots, seguit de períodes de remissió, o poden empitjorar de manera progressiva amb el temps, segons el tipus d'esclerosi múltiple.

Com s'ha vist hi ha diferents tipus d'esclerosi múltiple, que en el treball estan classificats en tres subtipus de la malaltia i ens voluntaris sans o controls, que són els zeros. Aquests tres subtipus són els següents: la RRMS, la SPMS, i la PPMS. Cada tipus té característiques clíniques i trajectòries diferents:

- **Relapsing-Remitting Multiple Sclerosis (RRMS):** És el tipus més comú i es caracteritza per brots clarament definits de nous símptomes o agreujament dels símptomes existents, seguits per períodes de remissió parcial o completa.
- **Secondary Progressive Multiple Sclerosis (SPMS):** Aquesta forma sol desenvolupar-se en pacients que inicialment tenen RRMS. En SPMS, hi ha una progressió constant de la discapacitat amb o sense brots superposats.
- **Primary Progressive Multiple Sclerosis (PPMS):** És una forma menys comuna, caracteritzada per una progressió constant de la discapacitat des del començament on no es produeixen períodes de remissió.

El diagnòstic precoç i la classificació precisa dels tipus d'EM són crucials per proporcionar un tractament adequat i millorar la qualitat de vida dels pacients. Les tècniques tradicionals de diagnòstic inclouen la ressonància magnètica (MRI), l'anàlisi del líquid cefalorraquidi mitjançant punció lumbar, i les proves neurofisiològiques. El problema que hi ha és que aquestes tècniques poden ser invasives, costoses, i en alguns casos, no suficientment precises per detectar les diferències subtils entre els tipus d'esclerosi múltiple en les primeres etapes de la malaltia.

Motiu pel qual hi ha un interès creixent en el desenvolupament de mètodes no invasius basats en dades per millorar la precisió del diagnòstic. Les mètriques de grafs, que analitzen la connectivitat cerebral, ofereixen una oportunitat prometedora per identificar patrons específics associats amb diferents tipus d'EM. Aquest enfocament permet una comprensió més profunda de les alteracions en la xarxa neuronal i pot contribuir a diagnòstics més precisos i accelerats.

1.1.2 Fractional Anisotropy

L'anisotropia fraccional (en anglès, *fractional anisotropy*, FA), és una mesura que es deriva de les imatges de difusió per ressonància magnètica (DTI). Aquesta tècnica permet mesurar el moviment de les molècules d'aigua en el teixit cerebral. L'anisotropia fraccional mesura la direccionalitat del moviment de l'aigua, la qual cosa proporciona informació sobre la integritat de les fibres de matèria blanca en el cervell.

Aquest moviment de les molècules d'aigua es veu restringit per les membranes cel·lulars i altres estructures microscòpiques, i aquest moviment és comunament més restringit en una direcció en comparació amb altres. L'anisotropia fraccional proporciona un valor entre 0 i 1, on 0 indica que el moviment de l'aigua és igual en totes les direccions (isotròpic) i 1 indica que el moviment és totalment restringit a una direcció (anisotròpic). Aquesta mesura és útil per avaluar si les vies de matèria blanca estan danyades per alguna malaltia neurològica.

Les fibres de matèria blanca connecten diferents regions del cervell, i que no estiguin danyades és essencial per a la comunicació neuronal efectiva. En pacients amb esclerosi múltiple, la desmielinització i el dany axonal poden afectar l'anisotropia fraccional, la qual cosa fa que sigui una mesura valuosa per a l'efectuació de l'estudi d'aquesta malaltia.

1.1.3 Gray Matter

La matèria grisa (en anglès, *grey matter*, GM) és un tipus de teixit cerebral que conté la major part dels cossos neuronals, dendrites, i terminals axonals, així com cèl·lules, com per exemple, les cèl·lules glials (cèl·lules del sistema nerviós) i capil·lars. La matèria grisa es troba principalment a la superfície del cervell, és a dir, al còrtex cerebral i en diverses estructures profundes.

El còrtex cerebral és responsable de moltes funcions complexes, incloent-hi la percepció sensorial, el pensament, el raonament i el moviment voluntari. Les alteracions a la matèria grisa poden afectar aquestes funcions i estan implicades en diverses malalties neurològiques.

En el context de l'esclerosi múltiple, la pèrdua de volum de matèria grisa és un marcador important de la progressió de la malaltia. L'anàlisi de la matèria grisa permet identificar aquestes pèrdues mitjançant tècniques d'imatge per ressonància magnètica i correlacionar-les amb els símptomes clínics i el curs de la malaltia.

1.1.4 Resting State (RS)

L'estat de repòs (en anglès, *resting state*, RS) es refereix a l'activitat cerebral que es mesura mentre el subjecte està, com indica el seu nom, en un estat de repòs, és a dir, no acompleix cap tasca específica.

La imatge per ressonància magnètica funcional (fMRI) en estat de repòs és una tècnica utilitzada per examinar la connectivitat funcional del cervell, ja que detecta els canvis en el flux sanguini cerebral associats amb l'activitat neuronal. Aquesta tècnica permet identificar xarxes neuronals que es mantenen actives de manera coherent durant el repòs.

En l'esclerosi múltiple, la connectivitat funcional pot estar alterada, i l'anàlisi de l'estat de repòs pot revelar patrons específics de desconnexió entre regions cerebrals. Això pot proporcionar informació molt valuosa sobre el funcionament de la malaltia i les seves manifestacions a les diferents regions del cervell.

1.2 Objectius

El present treball final de grau s'ha articulat al voltant de diversos objectius que han estat fonamentals per a la seva realització. Els objectius principals han estat:

- El primer objectiu primari és la investigació de les alteracions en la connectivitat cerebral en pacients amb esclerosi múltiple, centrada en l'anàlisi exhaustiva de les alteracions que es produeixen en les xarxes neuronals dels pacients diagnosticats amb aquesta malaltia, utilitzant mètriques de grafs basades en les dades proporcionades.

- El segon proposit fonamental és l'anàlisi de la variabilitat en els resultats segons les tècniques i canvis en les dades, explorant com diferents tècniques i ajustaments en les dades influeixen en els resultats obtinguts en la predicció de si son pacients o controls.

Els objectius secundaris que han motivat aquest projecte han estat:

- Primerament, la contribució a la millora dels mètodes diagnòstics de l'esclerosi múltiple, justificant la necessitat de mètodes més senzills per a la diagnosi, d'aquesta forma es pot suposar una millora significativa a la vida dels pacients, ja que es fan ús de tècniques menys invasius.

- Relacionat amb el punt anterior, els nous mètodes son més economics el que facilita que es facin més diagnòstics i amb més facilitat.

- Per últim, aquests diagnòstics també són més eficients. i proposant que els models desenvolupats podrien oferir una eina útil per a la detecció precoç de la malaltia. Això podria suposar una millora significativa en la qualitat de vida dels pacients, ja que una detecció més ràpida i precisa permetria intervenir de manera més efectiva i personalitzada.

D'aquesta manera, l'estudi no només contribueix al coneixement científic en l'àmbit de la neurociència, sinó que també ofereix noves perspectives a la practicitat d'utilitzar mètodes d'aprenentatge automàtic a la medicina moderna

1.3 Estructura del document

Primerament al capítol 1 hi ha la continuació del resum on es presenta el context del treball amb nomenclatura i les seves definicions importants per poder seguir-lo, aquestes inclouen l'esclerosis múltiple, l'anisotropia fraccional, la matèria grisa i l'estat de repòs. A continuació en aquest mateix capítol es detallen els objectius del treball els quals ja es mencionaven en el resum o *abstract* però de forma més extensa.

El capítol 2 inclou una explicació de com funcionen i que són realment els diferents models de *Machine Learning* que s'han utilitzat al treball, entre els quals s'inclouen la màquina de suport vectorial (SVM) i la regressió logística. En la segona part del capítol es defineixen i s'explica el funcionament de tècniques avançades en el camp de l'aprenentatge automàtic i el processament de dades estructurades en forma de grafs com poden ser el *random walk*, el *deep walk* o les GNNs.

En el capítol 3 es detallen les diferents dades que s'han utilitzat per a aquest projecte el seu treball i la seva estructura. En la primera secció s'expliquen les diferents dades generals com poden ser els pacients i els controls, informació important sobre ells, i les dades FA, RS i GM. També s'explica com s'han unit les dades obtingudes de dues fonts diferents. En la segona secció com indica el seu nom es parla sobre processament de les dades, és a dir, com es converteixen les matrius a graf i quins filtres s'apliquen.

El capítol 4 es centra en la descripció de la metodologia, i es divideix en quatre seccions: la descripció general de la metodologia, les diferents mètriques de teoria de grafs, els embeddings and randomwalks i deepwalk i per últim les GNNs. Aquest capítol segueix la mateixa estructura que el treball i comenta tots els passos que s'han fet al codi de forma resumida i utilitzant els conceptes que s'han definit anteriorment.

El capítol 5 presenta els resultats detallant quina ha estat la resposta de cadascun dels passos que s'han anat explicant a la metodologia i seguint al treball.

Finalment, al capítol 6, es discuteixen les conclusions on es resumeixen les principals claus obtingudes i es proposen futures línies de recerca.

Chapter 2

Estat de l'art

2.1 Models de Machine Learning

El *Machine Learning*, o en català, aprenentatge automàtic, és una branca de la intel·ligència artificial que se centra en el desenvolupament d'algorismes i models que permeten als ordinadors aprendre a partir de dades i fer prediccions o prendre decisions sense ser programats explícitament per a cada tasca. Aquests models analitzen grans volums de dades, identifiquen patrons i ajusten les seves prediccions en funció de l'experiència, millorant amb el temps i amb l'acumulació de dades. El *machine learning* s'aplica en àrees com la visió per computador, el processament del llenguatge natural, i els sistemes de recomanació, entre altres.

Els models de *Machine Learning* són algorismes o conjunts d'algorismes dissenyats per a aprendre patrons a partir de dades. L'objectiu d'aquests models és fer prediccions o prendre decisions basades en noves dades, utilitzant el coneixement adquirit durant la fase d'entrenament. Aquests models poden ser de diferents tipus d'aprenentatge:

- **Aprenentatge Supervisat:** aprenen a partir d'un conjunt de dades etiquetades, on cada exemple d'entrenament està associat amb una etiqueta o resultat esperat. Són molt útils per a fer classificació, com era necessària en aquest cas.

- **Aprenentatge No Supervisat:** treballen amb dades sense etiquetar, buscant estructures o patrons ocults en les dades. Són molt útils per fer *clustering*, és a dir, agrupar dades segons les seves similituds.

- **Aprenentatge Semi-Supervisat:** combina un petit conjunt de dades etiquetades amb un gran conjunt de dades sense etiquetar. L'objectiu és millorar la precisió de l'aprenentatge utilitzant tant dades etiquetades com sense etiquetar. És una barreja entre l'aprenentatge supervisat i el no supervisar i és molt útil per classificar grans dades on només hi ha una part etiquetada.

- **Aprenentatge per Reforç:** implica un agent que aprèn a prendre decisions mitjançant la interacció amb un entorn, rebre recompenses o penalitzacions per les accions que realitza, i millorar les seves decisions al llarg del temps.

2.1.1 Support Vector Machines (SVM)

Les màquines de vector de suport són un algoritme d'aprenentatge automàtic supervisat, ja que aprèn a partir d'un conjunt de dades etiquetades, on cada exemple d'entrenament està associat amb una etiqueta o classe coneguda. És utilitzat per a tasques de classificació i regressió. L'objectiu principal d'una SVM és trobar un hiperplà òptim que separi les diferents classes en un espai de característiques. L'algorisme només pot trobar aquest hiperplà en problemes que permeten la separació lineal; en la majoria dels problemes pràctics, l'algorisme maximitza el marge flexible permetent un petit nombre de classificacions errònies. Aquest hiperplà es selecciona per maximitzar la distància (marge) entre els punts de dades de les diferents classes més propers, coneguts com a vectors de suport.

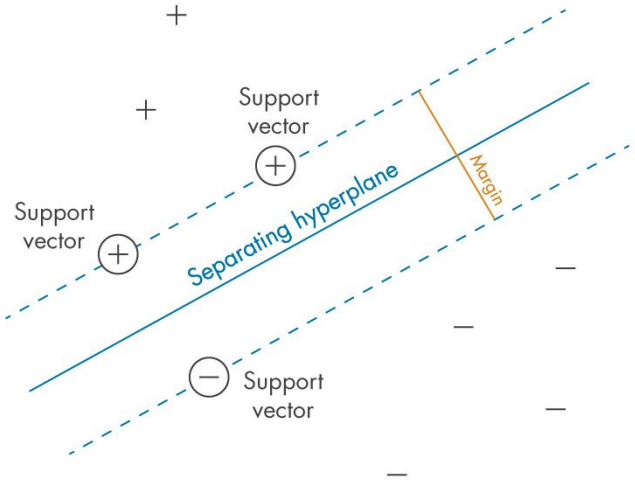


Figure 2.1: Teoria de la SVM

En la figura 2.1 es pot observar com l'algorisme intenta separar dues classes diferents de dades, representades per símbols + i -, utilitzant una línia divisòria que és l'hiperplà separador. Aquest hiperplà separador, representat per una línia blava contínua al centre de la imatge, és la frontera que divideix les dues classes de dades. És, per tant, l'objectiu principal de l'algorisme SVM trobar l'hiperplà que separi les dues classes amb el màxim marge possible. Aquest marge és la distància entre l'hiperplà separador i els punts de dades més propers a ell, que són coneguts com a vectors de suport.

Els vectors de suport, marcats a la imatge, són els punts de dades més propers a l'hiperplà i són especialment importants perquè determinen la posició exacta de l'hiperplà. Les línies discontínues blaves que passen per aquests vectors de suport defineixen els límits del marge. L'algorisme busca maximitzar aquest marge, perquè com més gran sigui, més clara i robusta serà la separació entre les dues classes.

L'algorisme SVM estàndard està formulat per a problemes de classificació binària; els problemes multiclasse normalment es redueixen a una sèrie de problemes binaris. Les SVM són especialment efectives en espais d'alta dimensió i destaquen per la seva utilitat en situacions on el nombre de dimensions és superior al nombre de mostres.

Les SVM són famosament útils perquè poden utilitzar diferents funcions de nucli (kernel functions) les quals assignen les dades a un espai dimensional diferent, que sovint és superior, amb l'expectativa que resulti més fàcil separar les classes després d'aquesta transformació, simplificant potencialment els límits de decisió complexos no lineals per fer-los lineals en l'espai dimensional de característiques superior assignat. En aquest procés, no cal transformar explícitament les dades, ja que això suposaria una alta càrrega computacional. Aquest mètode és conegut com el truc del nucli.

Hi ha diferents tipus de màquines de vector de suport segons la seva finalitat:

- **SVM Lineal:** s'utilitza quan les dades són linealment separables, és a dir, es pot traçar una línia (o un hiperplà en dimensions superiors) que separa clarament les dues classes. És l'original i, per tant, al qual se li aplica tota la teoria anterior.

- **SVM No Lineal:** s'utilitza quan les dades no són linealment separables. En aquest cas, s'utilitza una funció de nucli (kernel) per transformar les dades en un espai de característiques superior on es poden separar linealment.

- **SVM amb Nucli:** és una extensió de l'SVM que utilitza diferents funcions de nucli per projectar les dades en un espai de característiques de major dimensió, on les dades es poden separar millor. Els nuclis més comunament utilitzats són el nucli lineal, el nucli polinòmic, el nucli radial basis function (RBF) o gaussià i el nucli sigmoide. S'utilitza per tractar problemes on la separació no és possible en l'espai original de les dades, per exemple les dades no es poden separar de forma lineal.

- **SVM de marge suau (soft-margin):** permet algunes violacions de la separació correcta de les dades, és a dir, permet que alguns punts caiguin al costat incorrecte de l'hiperplà. Això és útil quan les dades són sorolloses o no perfectament separables, ja que prefereix un model que generalitzi millor en lloc de separar perfectament totes les dades d'entrenament.

- **SVM de marge dur (hard-margin):** requereix que totes les dades siguin correctament classificades sense cap error. Aquest enfocament només funciona bé si les dades són completament linealment separables i no hi ha soroll. És necessari per fer una classificació exacta.

- **SVM multiclasse:** és una adaptació de l'SVM original, que és de classificació binària, per poder fer classificació multiclasse. Hi ha dos tipus, el One-vs-One (OvO), el qual crea un model SVM per a cada possible parella de classes, i la classe que guanya més comparacions és seleccionada i el One-vs-Rest (OvR), el qual crea un model SVM per a cada classe, on la classe objectiu es compara contra totes les altres classes combinades.

- **Support Vector Regression (SVR):** és una extensió de l'SVM per a tasques de regressió, on l'objectiu és predir un valor continu en lloc de classificar dades en categories.

2.1.2 Logistic Regression

La regressió logística és un model estadístic utilitzat per a tasques de classificació, especialment quan l'objectiu és preveure l'etiqueta de classe binària (per exemple, sí/no o vertader/fals) d'una variable dependent. A diferència de la regressió lineal, que prediu valors continus, la regressió logística estima la probabilitat que una mostra pertanyi a una de les classes, de forma que pot oferir informació addicional sobre la certesa del model.

El model utilitza una funció logística (també coneguda com a funció sigmoide) per *mapejar* qualsevol valor real en un interval comprès entre 0 i 1. La funció logística té la forma:

$$\mathrm{P}(Y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}}$$

On:

- $(P(Y = 1|X))$ és la probabilitat de l'esdeveniment d'interès.
- $(\beta_0, \beta_1, \ldots, \beta_n)$ són els coeficients del model que s'aprenen durant l'entrenament.
- $(X_1, \ldots, X_n)$ són les variables independents o característiques.

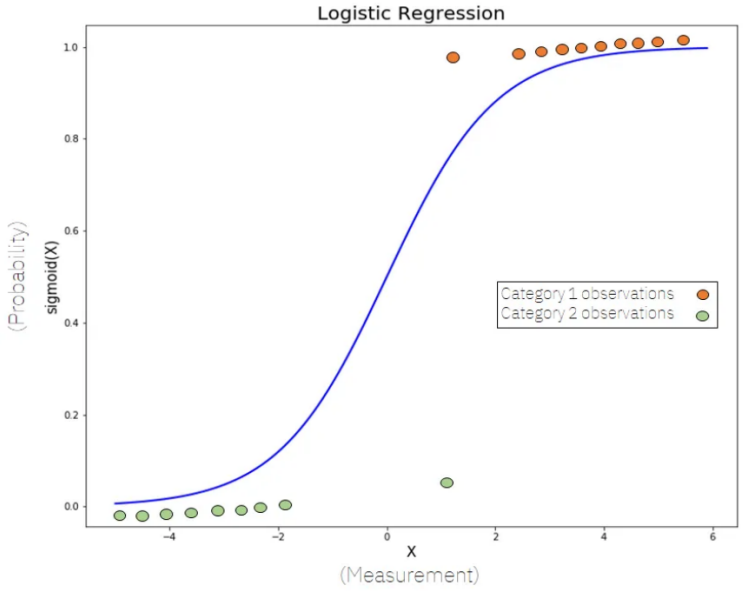


Figure 2.2: Funció sigmoide ajustada a unes dades.

A la figura 2.2 es pot observar la funció sigmoide d'una regressió logística aplicada a unes dades. Com s'ha explicat anteriorment, els models de regressió logística són models de classificació generalment binària motiu pel qual es poden percebre dues categories diferents, aquestes dues categories són dues observacions completament distintes. D'altra banda l'eix Y va de 0 a 1 perquè la funció sigmoide sempre té com a màxim i mínim aquests dos valors, cosa que s'adapta molt bé a l'objectiu de classificar mostres en dues categories diferents. En calcular la funció sigmoide de X (que és una suma ponderada de les característiques d'entrada, igual que en la regressió lineal), s'obté una probabilitat (evidentment entre 0 i 1) que una observació pertanyi a una de les dues categories.

Hi ha diferents tipus de regressió logística:

- **Regressió Logística Binària:** és l'original i més bàsica. S'encarrega de problemes de classificació on hi ha dues classes (0 o 1, sí o no, positiu o negatiu, etc.).

- **Regressió Logística Multiclasse (o Multinomial):** és una extensió de la regressió logística binària que s'utilitza quan hi ha més de dues classes. A diferència de la regressió logística binària, aquest tipus de regressió pot manejar múltiples categories simultàniament. Hi ha diferents tipus com el One-vs-Rest (OvR) o One-vs-All (OvA) on es crea un model de regressió logística binària per a cada classe, on la classe objectiu és comparada contra totes les altres classes combinades. Finalment, se selecciona la classe amb la predicció de probabilitat més alta i el Softmax (o Regressió Logística Multinomial) el qual és una extensió natural de la regressió logística que calcula probabilitats per a totes les classes i escull la classe amb la probabilitat més alta. És comú en models de xarxes neuronals.

- **Regressió logística ordinal:** s'utilitza quan les categories tenen un ordre natural, però les distàncies entre les categories no són necessàriament iguals.

- **Regressió logística multi-etiqueta (multilabel):** la regressió logística multietiqueta permet que un exemple pertanyi a múltiples classes simultàniament. Això es fa entrenant múltiples models de regressió logística binària, un per cada classe, i cada model prediu si una etiqueta s'aplica o no a un exemple. És útil per a problemes on els elements poden tenir diverses categories alhora, com en classificació de textos o etiquetatge d'imatges.

- **Regressió logística lasso i ridge:** són versions regularitzades de la regressió logística. La regularització s'utilitza per prevenir el sobreajustament (overfitting) i millorar la generalització del model. Hi ha dos tipus la *lasso* (L1) la qual pot portar a la selecció de variables, ja que tendeix a reduir a zero els coeficients de variables menys rellevants i la *ridge* (L2) la qual redueix la magnitud dels coeficients, però no porta a coeficients exactament iguals a zero.

2.2 Graph learning

L'aprenentatge en grafs, o en anglès *graph larning*, és una branca del *machine learning* que se centra en l'anàlisi i interpretació de dades representades en forma de graf. Un graf és una col·lecció de nodes (o vèrtexs) i arestes, on els nodes representen entitats i les arestes les relacions o interaccions entre aquestes entitats. Aquesta estructura és útil per modelar xarxes complexes en àmbits com les xarxes socials, xarxes biològiques i xarxes de comunicació.

Una xarxa neuronal és un model computacional inspirat en el funcionament del cervell humà, dissenyat per reconèixer patrons i prendre decisions basades en dades. Està formada per capes de neurones artificials (que són els nodes o vèrtex) interconnectades, on cada connexió té un pes associat. Les dades s'introdueixen a la capa d'entrada, passen per les capes ocultes on es processen, i finalment produeixen un resultat a la capa de sortida. Les xarxes neuronals són àmpliament utilitzades en tasques com reconeixement d'imatges, processament de llenguatge natural i predicció de sèries temporals.

L'aprenentatge en grafs aprofita les relacions i estructures dins del graf per aprendre i fer prediccions, ja que utilitzant l'estructura inherent de les dades en graf, els algorismes d'aprenentatge en grafs poden descobrir coneixements i patrons profunds que no són evidents amb enfocaments tradicionals d'aprenentatge automàtic.

2.2.1 Embeddings de nodes

Els *embeddings* són una tècnica de processament del llenguatge natural que converteix el llenguatge humà en vectors matemàtics. Aquests vectors són una representació del significat subjacent de les paraules, cosa que permet a les computadores processar el llenguatge de manera més efectiva. Els *embeddings* permeten que les paraules siguin tractades com a dades i manipulades matemàticament. Aquesta tècnica s'utilitza àmpliament en la intel·ligència artificial per a tasques com la classificació de text i la traducció automàtica.

Són molt útils per poder fer *forecasting* de quals dels usuaris que es comuniquen entre ells es connectaran. Els usuaris es representen com a nodes, mentre que les relacions entre ells es mostren com a arestes.

El procés de creació d'*embeddings* comença amb la construcció d'un corpus, que és una col·lecció de textos. A partir d'aquest corpus, es crea un model de llenguatge que aprèn a predir paraules segons el seu context. Un cop entrenat el model, s'utilitzen les seves capes internes per generar els vectors d'*embeddings* de les paraules. Els vectors generats pels *embeddings* tenen diverses propietats útils que els fan especialment efectius en aplicacions de processament del llenguatge natural. En primer lloc, els vectors són densos, la qual cosa significa que cadascuna de les seves dimensions conté informació rellevant. En segon lloc, els vectors són similars per a paraules que s'utilitzen en contextos similars, fet que permet utilitzar els *embeddings* per determinar la similitud semàntica entre paraules.

Hi ha diferents tècniques per crear *embeddings*, com per exemple el *word2vec* i el *node2vec*. El primer és una tècnica de processament del llenguatge natural que converteix les paraules en vectors numèrics de manera que les paraules que apareixen en contextos similars tinguin representacions vectorials similars. Es basa en dos models principals: CBOW (Continuous Bag of Words), que prediu una paraula donat el seu context, i Skip-gram, que prediu el context donat una paraula. Això permet capturar relacions semàntiques entre les paraules, fent-lo molt útil en tasques com la similitud de paraules i la classificació de textos. El segon és una extensió de Word2Vec aplicada a grafs. Transforma nodes d'un graf en vectors numèrics d'una manera que preserva les relacions estructurals entre els nodes. Ho fa generant "camins aleatoris" (random walks) a través del graf per crear seqüències de nodes, de manera similar a com Word2Vec treballa amb seqüències de paraules. Aquestes seqüències s'utilitzen després per entrenar un model que genera *embeddings* que representen cada node, fent-los útils per a tasques com la predicció d'enllaços o la classificació de nodes dins del graf.

2.2.2 Randomwalk

Abreviat com RW i anomenat *camí aleatori* en català, el *RandomWalk* és un procés matemàtic que descriu una seqüència de passos aleatoris en algun espai, com ara una línia, un pla, o un graf. En aquest procés, un punt o partícula es mou d'una posició a una altra seguint regles probabilístiques, sense un camí predefinit. Els *random walks* són utilitzats en diverses disciplines, com la física, la biologia, l'economia i la informàtica, per modelar fenòmens que impliquen moviments o decisions aleatòries, com la difusió de partícules o la generació d'estructures de dades.

Hi ha diferents tipus de *RandomWalk* segons si els passos son igualment probables en qualsevol direcció (*RandomWalk* simple) o si tenen diferents probabilitats (*RandomWalk* biaixat), també pot ser que els moviments estiguin restringits per certes condicions (*RandomWalk constrained*), si aquesta restricció és concretament que la caminada no pot passar per una posició que ja ha estat visitada anteriorment, llavors s'anomena *self-avoiding walk*, si, en canvi, es restringeix a una xarxa discreta s'anomena *lattice RandomWalk*, finalment el *brownian motion* modela el moviment de partícules en un fluid. Aquest tipus de procés també es fa servir en simulacions estadístiques i en algorismes de cerca i navegació per Internet. A més, els random walks tenen propietats estadístiques importants, com la convergència cap a una distribució normal en certs casos, la qual cosa els fa valuosos en l'anàlisi i modelatge de fenòmens més complexos.

Els RandomWalks poden ser en 1D, 2D, 3D. Comunament són més utilitzats els 2D, ja que són àmpliament utilitzats en l'estudi de la mobilitat. És un patró de moviment sense memòria on la velocitat actual és independent de la passada, però també pot generar moviments poc realistes, com parades sobtades i girs bruscos.

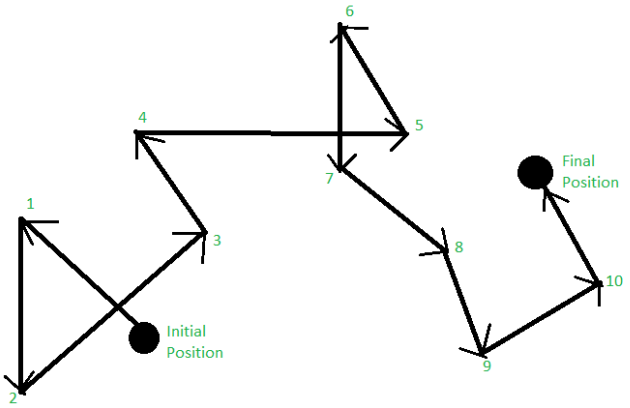


Figure 2.3: Exemple de RandomWalk

A la figura 2.3 es pot observar un exemple de *random walk* en un pla bidimensional. Aquesta imatge mostra una seqüència de passos aleatoris que es mouen en diverses direccions en un pla. La posició inicial i la posició final es marquen clarament, i les fletxes indiquen la direcció i l'ordre dels passos. Comença en una posició inicial marcada i es mou en una sèrie de 10 passos en diverses direccions, cadascun indicat per fletxes numerades. Les fletxes representen la direcció i l'ordre de cada moviment, mentre que la posició final és clarament indicada. Aquesta visualització mostra com un punt es desplaça aleatòriament sense seguir un camí predeterminat, un concepte central en l'estudi de les caminades aleatòries.

2.2.3 DeepWalk

El *DeepWalk* és un algorisme que aprèn representacions latents de nodes en un graf utilitzant caminades aleatòries truncades i posteriorment utilitza l'algorisme Word2Vec per aprendre les representacions dels nodes. Inspirat en tècniques de modelatge de llenguatge, tracta aquestes caminades aleatòries com si fossin "frases" en un llenguatge, permetent capturar la informació estructural del graf. *DeepWalk* genera aquestes representacions vectorials per als nodes, que poden ser utilitzades en tasques com la classificació de nodes, especialment en contextos de xarxes socials. L'algorisme és escalable i eficient, sent capaç de manejar grafs de gran escala i obtenir millors resultats que altres mètodes en situacions amb dades etiquetades escasses.

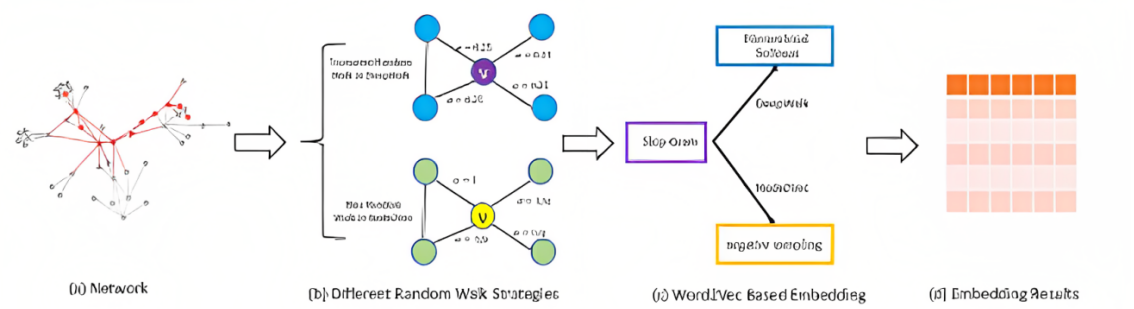


Figure 2.4: Exemple de DeepWalk

A la figura 2.4, es pot observar un exemple de *DeepWalk*. Aquesta imatge il·lustra de manera esquemàtica el flux de treball de l'algorisme, des de la generació de caminades aleatòries fins a la creació dels embeddings utilitzant Word2Vec. La imatge comença mostrant una xarxa o graf, on els nodes estan connectats per arestes a l'apartat (a). Aquest graf és la base sobre la qual es fan les *random walks*. Posteriorment, a l'apartat (b) es mostren diferents estratègies de caminada aleatòria que es poden emprar a partir dels nodes del graf. Aquestes caminades aleatòries exploren les connexions locals del graf, movent-se d'un node a un altre en funció de les connexions disponibles. Això ajuda a capturar les relacions locals entre els nodes.A l'apartat (c) els *random walks* generats s'utilitzen com a entrada per al model Word2Vec. Aquest model, popular en processament de llenguatge natural, es reentrena aquí per aprendre representacions vectorials (embeddings) dels nodes basant-se en les seqüències generades pels *random walks*. Finalment, a l'apartat (d) els resultats es presenten en forma de matriu d'embeddings. Cada fila en aquesta matriu representa el vector d'embeddings d'un node, que encapsula la seva posició relativa i les relacions dins del graf. Aquests embeddings es poden utilitzar posteriorment per a tasques com la classificació de nodes, la predicció d'enllaços o la categorització de grafs.

De forma general es podria resumir en que primerament, l'algorisme *DeepWalk* genera un conjunt de *random walks* que comencen des de cada node del graf. Cada *random walk* és simplement una seqüència de nodes que comença en un node particular i es mou cap a un dels seus veïns en cada pas. Aquestes *random walks* capturen l'estructura local del graf i proporcionen una manera d'aprendre representacions dels nodes que reflecteixen les seves relacions amb els nodes propers.

A continuació, l'algorisme *Word2Vec* s'utilitza per aprendre representacions dels nodes basant-se en les caminades aleatòries generades. Tot i que el *Word2Vec* és un algorisme popular per aprendre representacions de paraules en el processament del llenguatge natural, també es pot aplicar per aprendre representacions dels nodes en un graf. La idea bàsica darrere de *Word2Vec*, de forma resumida, és aprendre una xarxa neuronal que prediu la probabilitat d'una paraula donat el seu context (és a dir, les paraules que apareixen abans i després d'ella).

2.2.4 GNN

Les xarxes neuronals gràfiques (GNNs), o en anglès *Graph Neural Network*, són un tipus especial de xarxes neuronals dissenyades per treballar amb dades estructurades en forma de graf. Aquestes xarxes estan influenciades per les xarxes neuronals convolucionals (CNN), o en anglès *Convolutional Neural Network* i les tècniques d'incrustació de grafs, cosa que permet aplicar-les a la predicció de nodes, arestes i altres tasques basades en grafs. Les GNNs són especialment útils en contextos en què les dades tenen una estructura complexa que no es pot tractar de manera efectiva amb xarxes neuronals convencionals.

De la mateixa manera que les CNN són utilitzades per a la classificació d'imatges, aplicant convolucions sobre una quadrícula de píxels per predir una classe, les GNNs s'apliquen a l'estructura dels grafs per fer prediccions similars. Aquestes xarxes també s'assemblen a les xarxes neuronals recurrents, les quals s'utilitzen per a la classificació de textos, però en aquest cas, les GNNs treballen amb estructures gràfiques on cada paraula pot ser vista com un node dins d'una frase. Això permet a les GNNs tractar dades que no segueixen una estructura lineal o bidimensional, com és el cas dels grafs.

Les GNNs es van introduir perquè les xarxes neuronals convolucionals no obtenien resultats òptims en situacions en què el graf tenia una grandària arbitrària i una estructura complexa. En lloc d'això, les GNNs permeten la propagació d'informació entre nodes seguint les arestes del graf, capturant així les relacions i la topologia local del graf. Això els permet aprendre representacions vectorials que encapsulen informació rellevant sobre la posició i les relacions dels nodes dins del graf, la qual cosa és molt valuós en una àmplia varietat d'aplicacions, des de l'anàlisi de xarxes socials fins a la biologia computacional.

Hi ha diferents tipus de GNN modificats per cada necessitat, alguns dels més comuns son els següents:

- **Graph Convolutional Networks (GCNs):** Aquestes xarxes s'inspiren en les xarxes neuronals convolucionals (CNNs) i són probablement les GNNs més conegudes. Les GCNs aprenen les característiques d'un node inspeccionant els seus nodes veïns i agregant la informació per passar-la a través de capes de la xarxa. Les GCNs es divideixen en dos tipus principals: les xarxes convolucionals espacials, que defineixen convolucions basades en la topologia del graf, i les xarxes convolucionals espectrals, que utilitzen la descomposició espectral del Laplacià del graf per propagar la informació entre nodes.
- **Graph Attention Networks (GATs):** Aquestes xarxes utilitzen un mecanisme d'atenció per ponderar la importància dels nodes veïns en l'actualització de les representacions dels nodes. Això permet que les GATs se centrin més en les parts del graf que són més rellevants per a una tasca específica, millorant així l'eficàcia del model en problemes amb dependències complexes.
- **Message Passing Neural Networks (MPNNs):** En aquestes xarxes, la informació es transmet com a "missatges" entre els nodes d'un graf. Cada node envia i rep missatges dels seus veïns, i després actualitza la seva representació en funció d'aquesta informació agregada. Les MPNNs són molt flexibles i es poden adaptar a diferents tipus de problemes, però poden tenir problemes d'escalabilitat a causa de la necessitat de processar grans quantitats de missatges.
- **Graph Auto-Encoder Networks:** Aquestes xarxes utilitzen un enfocament d'*autoencoders* per aprendre representacions dels grafs. Un encoder aprèn a codificar la informació del graf en un espai latent, mentre que un *decoder* intenta reconstruir el graf original des d'aquest espai latent. Això és particularment útil per a tasques com la predicció d'enllaços, on és important capturar la relació entre nodes de manera efectiva.
- **GraphSAGE:** Aquesta tècnica és popular per al seu ús en grafs de gran escala. GraphSAGE utilitza un mètode de mostreig per seleccionar un subconjunt de veïns per a cada node en cada iteració, la qual cosa redueix els costos computacionals i permet treballar amb grafs molt grans.

Graph Convolutional Networks (GCNs)

Com s'ha explicat abans de forma supèrflua una *Graph Convolutional Network* (GCN) és un tipus especial de xarxa neuronal dissenyada per treballar amb dades estructurades en forma de graf i és la que s'ha utilitzat en aquest projecte.

Com les xarxes neuronals convolucionals (CNNs) aquestes xarxes estenen el concepte de convolucions, utilitzades tradicionalment per a imatges, al domini dels grafs. En lloc de realitzar convolucions sobre una quadrícula regular de píxels, com es faria amb imatges, les GCNs operen sobre la topologia d'un graf, que consisteix en nodes, que són les entitats i arestes que són les relacions entre aquestes.

El funcionament d'una GCN es basa en l'agregació de la informació dels nodes veïns per actualitzar la representació d'un node en particular, el que vol dir que cada node en el graf no només es representa per les seves pròpies característiques, sinó també per la informació dels seus veïns immediats. Aquest procés es repeteix en diverses capes, on cada capa de la GCN pot considerar informació de veïns més llunyans, capturant així totes les relacions dins del graf, tant les locals com les globals.

A més, les GCNs utilitzen funcions d'activació no lineals, com la ReLU, després de cada pas d'agregació. Això introdueix no-linearitat en el model, permetent que les GCNs aprenguin representacions complexes dels nodes.

Per tant, les aplicacions de les GCNs són àmplies i variades. Són especialment útils en casos on les dades tenen una estructura complexa, ja que, com s'ha explicat abans no es poden modelar de manera efectiva amb altres tipus de xarxes neuronals, com les CNNs o les xarxes neuronals recurrents. Per exemple, en xarxes socials, les GCNs poden ajudar a predir noves connexions entre usuaris, o en biologia computacional, poden ajudar a identificar funcions de proteïnes basades en la seva posició en una xarxa de proteïnes, utilitzant les representacions complexes dels nodes. Aquesta capacitat de modelar la complexitat inherent als grafs fa que les GCNs siguin molt útils en el *machine learning*.

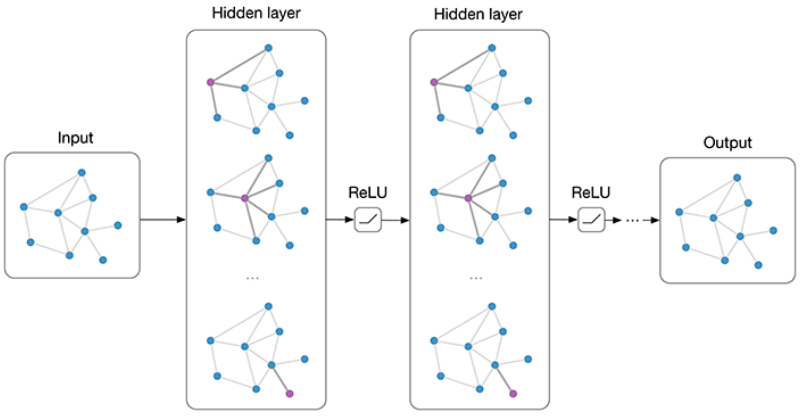


Figure 2.5: Representació d'una GCN

A la figura 2.5 es pot observar el funcionament d'una *Graph Convolutional Network* (GCN), mostrant el procés de propagació d'informació a través de les capes de la xarxa. La imatge representa el flux de dades a través d'una GCN, mostrant com la informació dels nodes es combina i processa per capturar les relacions estructurals del graf, permetent així a la xarxa fer prediccions o aprendre noves característiques sobre el graf.

Primerament, a l'esquerra es veu la representació inicial d'un graf com a *input*, és a dir, com a entrada on els nodes estan connectats per arestes. L'entrada de la xarxa conté la informació bàsica dels nodes i les seves connexions.

A continuació, hi ha un apartat anomenat *hidden layers*, el que en català vol dir capes ocultes. Se'n poden observar dues, cadascuna representant una etapa de processament dins de la GCN. En cada capa oculta, la informació dels nodes s'actualitza en funció de la informació dels seus veïns, agregant dades que reflecteixen les relacions locals dins del graf. Aquest procés permet a la xarxa capturar patrons i dependències complexes. Després de cada capa oculta, la informació processada passa a través d'una funció d'activació ReLU (Rectified Linear Unit), que introdueix no linearitat en el model. Això ajuda la xarxa a aprendre representacions més riques i complexes dels nodes.

Finalment, a la dreta, es veu l'*output*, o en català, la sortida de la xarxa, que és una nova representació del graf, amb la informació actualitzada després de passar per les capes ocultes i les funcions d'activació. Aquesta representació final es pot utilitzar per a diverses tasques, com la classificació de nodes, la predicció d'enllaços, o la categorització del graf complet.

GraphSAGE

Un altre concepte a recalcar és el *GraphSAGE*, el que vol dir *Graph Sample and AggregatE*, i és un algorisme avançat, escalable i versàtil dissenyat per a l'aprenentatge automàtic en grafs. A diferència dels mètodes tradicionals que generen una representació única per a cada node del graf, *GraphSAGE* es focalitza en l'aprenentatge d'una funció d'agregació generalitzada que es pot aplicar tant als nodes existents com als nodes no vistos durant l'entrenament. Això permet generar embeddings de nodes que són capaços de capturar la informació estructural i les característiques locals del graf, proporcionant una representació compacta i útil per a diverses tasques.

El funcionament de GraphSAGE es basa en dues etapes principals: el mostreig i l'agregació. En primer lloc, per a cada node, se selecciona un subconjunt de veïns mitjançant un procés de mostreig aleatori, la qual cosa permet escalar l'algorisme per treballar amb grafs molt grans. A continuació, GraphSAGE aplica una funció d'agregació sobre les representacions dels veïns seleccionats, combinant aquesta informació per actualitzar la representació del node d'interès. Aquest procés es pot repetir a través de diverses capes per capturar informació de veïns més llunyans en el graf.

Una de les característiques més destacades de GraphSAGE és la seva capacitat per generalitzar a nodes nous que no estaven presents durant l'entrenament, una propietat coneguda com a aprenentatge inductiu. Això és especialment útil en aplicacions on els grafs canvien amb el temps, com en xarxes socials, sistemes de recomanació o grafs d'informació dinàmics. Aquesta capacitat d'aprenentatge inductiu, combinada amb la flexibilitat per utilitzar diferents tècniques d'agregació, fa de GraphSAGE una eina poderosa per a l'anàlisi de grafs.

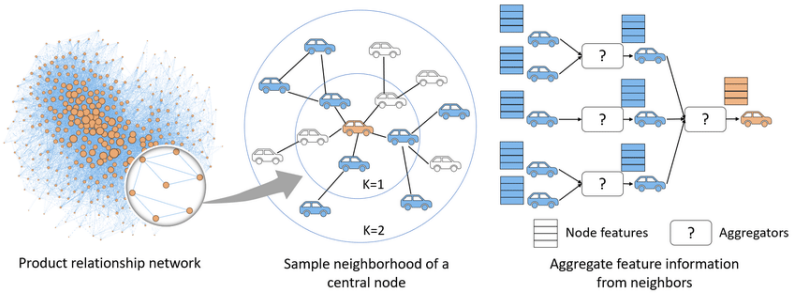


Figure 2.6: Exemple de GraphSAGE

A la figura 2.6, es pot observar un exemple del procés de *GraphSAGE*. Primerament, a la part anomenada *Product relationship network*, a l'esquerra, es mostra un graf que representa les relacions entre diferents productes. Els nodes representen els productes, i les arestes representen les relacions entre aquests productes.

A continuació, hi ha *Sample Neighborhood of a Central Node*, el que en català vol dir Mostreig del Veïnat d'un Node Central. La part central de la imatge mostra com *GraphSAGE* selecciona un conjunt de veïns d'un node central (en aquest cas, un cotxe). Aquest procés de mostreig es realitza en diferents "capes" o nivells (K=1, K=2), on es consideren els veïns immediats (K=1) i els veïns dels veïns (K=2).

Finalment a *Aggregate Feature Information from Neighbors*, o en català, Agregació de la Informació de les Característiques dels Veïns, mostra com s'agreguen les característiques dels veïns per generar la representació del node central. Aquí es veu com les característiques dels veïns es processen a través d'agregadors, que poden ser funcions com la mitjana, una xarxa neuronal, o altres mètodes d'agregació, per combinar la informació i actualitzar la representació del node central.

Difèrencies entre els GCN i el GraphSAGE

La diferencia més gran entre el *GraphSAGE* i els GCN, prèviament explicats, és que tot i que ambdós algorismes són utilitzats per generar representacions de nodes en grafs en GraphSAGE, es fa un mostreig aleatori dels veïns dels nodes per a cada nivell (com es veu en l'esquema central). Això permet escalar l'algorisme a grafs molt grans, ja que només es consideren una quantitat limitada de veïns. I, en canvi, les GCN normalment consideren tots els veïns d'un node per a l'agregació. Això pot ser ineficient en grafs molt densos o grans, perquè el nombre de veïns pot ser molt elevat.

Altres diferències entre els dos és que el *GraphSAGE* utilitza funcions d'agregació explícites que es poden ajustar, com la mitjana, la màxima, o combinacions amb xarxes neuronals. Això fa que sigui més flexible i adaptatiu a diferents tipus de grafs. I els GCN utilitzen una operació de convolució sobre el graf que és més rígida i que pot ser menys personalitzable que l'agregació de GraphSAGE. La convolució en GCN és una combinació lineal de les característiques dels nodes veïns, seguida d'una no-linealitat. A més el GraphSAGE té un aprenentatge més inductiu mentre que els GCN aprenen de manera transductiva, és a dir, aprenen representacions per a nodes específics del graf d'entrenament i no generalitzen fàcilment a nodes nous.

En resum, *GraphSAGE* és més escalable i flexible gràcies a la seva capacitat de mostreig i agregació de veïns, així com la seva habilitat per generalitzar a nodes nous. Els GCN són més directes i utilitzen la convolució del graf per a la propagació de característiques, però pot ser menys eficient en grafs grans i menys adaptable a noves dades.

Chapter 3

Dades

Les dades proporcionades han estat obtingudes de l'Hospital Clínic de Barcelona i posteriorment es van poder afegir dades obtingudes de Nàpols, Itàlia. Els malalts són pacients amb esclerosis múltiple diagnosticada amb diferents nivells de la malaltia i els controls són persones a les quals també se'ls va fer la prova que estaven completament sanes. Aquestes proves han estat fetes tant en homes com en dones de diferents edats.

3.1 Cohort i MRI

En aquest apartat, es descriu la composició de la cohort utilitzada en l'estudi, així com les dades generals dels pacients i els controls.

L'estudi va incloure un total de 270 participants, dividits en dos grups: pacients i controls. D'aquests participants 165 eren de les dades de l'Hospital Clínic, dels quals 147 eren pacients i 18 controls. Aquest biaix tan gran va ser el motiu pel qual després es van afegir les dades provenents de Nàpols, els quals eren 105 participants i per poder arreglar el biaix tots eren controls. Amb els dos grups de dades conjunts de pacients, que tenen edats de des de setze anys fins a setanta, es van acabar tenint 147 pacients i 123 controls.

Les dades es van recollir utilitzant la tècnica de ressonància magnètica (MRI), per obtenir informació detallada sobre les estructures cerebrals durant els anys 2016 i 2017.

Les dades de l'Hospital Clínic de Barcelona estaven dividides en tres carpetes, anomenades FA, GM i RS, essent en ordre l'anisotropia fraccional, la substància grisa i l'estat de repòs. En cada carpeta hi ha un total de 165 matrius d'adjacència, corresponent al nombre de pacients que s'han estudiat. També hi ha un document anomenat *demographics.csv* el qual conté una id que correspon als 165 pacients i la columna important es la de mstype, que és amb la que es treballa durant gran part. Aquesta columna conté números del 0 al 3, on 0 és un pacient control (és a dir, no malalt), i de l'1 al 3 són pacients malalts però augmentant la gravetat, essent 1 el que menys i 3 el que sí. Per últim, hi ha un arxiu anomenat *nodes.csv*, el qual conté dues columnes: ID i nom de les 76 regions del cervell que s'estudien.

Les dades provinents de Nàpols contenen també les tres carpetes FA, en aquest cas anomenada *DTI_networks*, GM que aquí es diu *GM_networks* i RS, aquí anomenada *rsfmri_networks*. En aquest cas hi ha un total de 105 matrius d'adjacència corresponents als 105 pacients que hi ha. Hi ha quatre arxius a més que són *GM_MD_NAPLES.csv* el qual conté dades relacionades amb les mètriques de difusió de la substància grisa dels subjectes, *GM_FA_NAPLES.csv*, el qual conté dades relacionades amb la fracció d'anisotropia de la substància grisa dels subjectes, *naples2barcelona_multilayer.xlsx*, el qual conté dades els pacients, com l'edat i el sexe i per últim *NODES_NAPLES.csv*, el qual conté informació sobre els nodes utilitzats en les anàlisis de grafs com ara noms de regions cerebrals.

D'aquestes dades es van acabar barrejant ambdós FAs, GMs i RSs en un sol per fer més fàcil el treball. El mateix es va fer amb *demographics.csv*, al qual se li van afegir les columnes de més i es van classificar els pacients nous de Nàpols tots com a control.

3.2 Processament de les dades

3.2.1 Creació dels grafs

En aquest apartat s'explica com es van processar les dades explicades anteriorment per poder fer l'anàlisi corresponent.

Primerament es van carregar tots els arxius de les carpetes, tant de l'Hospital Clínic com de Nàpols i es van ajuntar com s'ha explicat prèviament. Un cop fet l'agrupament corresponent, hi havia les dades de FA, GM i RS conjuntes i l'arxiu *demographics.csv* amb les dades obtingudes dels dos llocs.

A continuació, es va fer una anàlisi descriptiva que s'explicarà a l'apartat de Metodologia, una de les seccions a l'anàlisi descriptiva va ser fer histogrames de l'anisotropia fraccional, la substància grisa i l'estat de repòs d'on es van extreure uns *thresholds*, és a dir, uns llindars per poder eliminar arestes per a cada graf. Posteriorment, es van crear els 270 grafs per a cada secció, cadascún de 76 nodes.

3.2.2 NetworkX

Per fer aquests grafs s'ha utilitzat la llibreria *NetworkX*, la qual és una llibreria de Python àmpliament utilitzada per crear, manipular i analitzar grafs o xarxes complexes. NetworkX ofereix una plataforma flexible per a la construcció i anàlisi de grafs, cosa que la converteix en una eina essencial per a qualsevol que treballi amb dades relacionals.

Una de les característiques principals de NetworkX és la seva capacitat per crear diversos tipus de grafs, incloent-hi grafs no dirigits, dirigits, ponderats, bipartits i multigrafs, que són grafs amb múltiples arestes entre els mateixos parells de nodes. A més, NetworkX permet afegir atributs addicionals als nodes i arestes, com pesos, etiquetes o colors, permetent un modelatge molt detallat i adaptat a les necessitats específiques de l'usuari.

Pel que fa a la manipulació de grafs, NetworkX proporciona funcions per afegir, eliminar i modificar nodes i arestes de manera senzilla i eficient. També permet l'extracció de subgrafs basats en criteris específics, facilitant l'anàlisi de parts específiques de la xarxa. Aquesta capacitat de manipulació és clau per a la construcció de grafs complexos i per adaptar-los a les necessitats d'anàlisi.

NetworkX també és molt potent en l'anàlisi de grafs, oferint una àmplia gamma d'algorismes i mesures estructurals. Per exemple, es poden calcular mesures de centralitat, com PageRank o betweenness, així com la clústerització, la distància més curta entre nodes i el diàmetre del graf. Les quals s'han utilitzat per aconseguir dades sobre el graf sobre les quals s'han fet les regressions.

Finalment, NetworkX és compatible amb altres llibreries i formats de dades, com Pandas per treballar amb DataFrames o formats de fitxers com GraphML i GML per importar i exportar grafs. Aquesta compatibilitat la fa molt útil per integrar-se en fluxos de treball més amplis en ciència de dades o investigació.

Chapter 4

Metodologia

En aquest capítol s'explica de forma més detallada el procés que s'ha dut a terme en el programa, mentre que els resultats es trobaran en el capítol següent.

Primerament, en mètodes d'anàlisi hi ha representacions gràfiques de les dades per poder entendre millor amb el que es treballa. Es pot observar una matriu de control-pacient, un estudi de la connectivitat i els histogrames de les dades. A continuació, un cop s'han creat els grafs utilitzant la llibreria *NetoworkX*, com s'ha explicat al capítol anterior, s'han extret les mètriques dels grafs sobre els quals es faran els models.

A tècniques de preprocessament, s'explica que s'ha utilitzat per acabar de retocar les dades abans de fer els models, entre ells quins *feature scaling* s'han utilitzat per desbiaixar les dades i per què s'han utilitzat altres comandes. A l'entrenament dels models s'expliquen els diferents models.

Finalment, s'expliquen com s'han aplicat les *Graph Neural Networks* i el *GraphSAGE*.

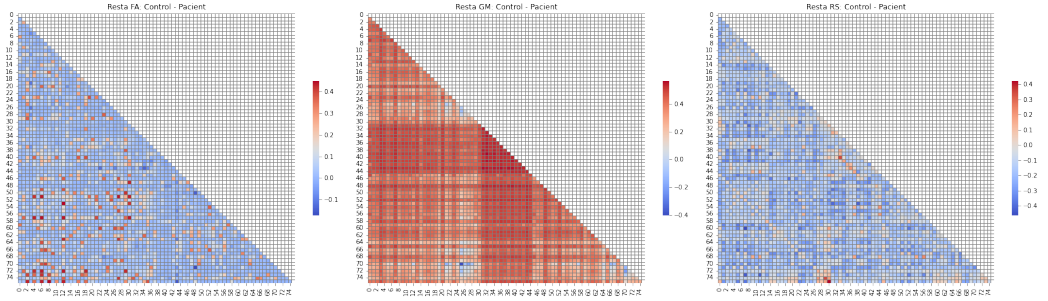
4.1 Mètodes d'anàlisi

Primerament, s'ha produït la separació mitjançant el *mstype* de l'arxiu de *demographcis.csv* classificant com a diferents variables els controls i els pacients i s'han produit diferents formes de poder observar les dades de forma més gràfica i comparativa entre elles com ara matrius o mapes de calor de diferències de connectivitat funcional cerebral entre dos grups: "Control" i "Pacient", de connectivitat en general, hisògrames, etc. Totes les diferents matrius de connectivitat que s'estudien estan tallades per la meitat ja que les matrius són simètriques.

4.1.1 Representació gràfica de les dades i anàlisi descriptiu

Matrius de diferències de connectivitat funcional cerebral entre els controls i els pacients

La figura 4.1 mostra tres martius de diferències de connectivitat funcional cerebral entre dos grups: "Control" i "Pacient". Aquestes matrius representen les diferències en connectivitat cerebral entre controls sans i pacients amb esclerosi múltiple (EM), representades com a matrius de diferències per a diverses mesures com FA (Anisotropia Fraccional), GM (Matèria Grisa) i RS (connectivitat funcional en estat de repòs). Els eixos representen diferents regions cerebrals i cada cel·la de la matriu indica la diferència de connectivitat entre dues regions. Si el color de la cel·la tendeix més a blau indica una reducció de connectivitat en els pacients comparat amb el grup de control i en canvi si el color de la cel·la tendeix més a vermell és al contrari. Aquest tipus d'anàlisi ajuda a identificar com l'esclerosi múltiple pot afectar la connectivitat del cervell en comparació amb persones sanes.



(a) FA **(b)** GM **(c)** RS

Figure 4.1: Diferències de connectivitat control-pacient

En la matriu de diferències de FA (Anisotropia Fraccional), es pot observar una disminució en diverses connexions, suggerint una pèrdua de la integritat de la matèria blanca en els pacients. Això és coherent amb els danys associats amb l'esclerosis mútilple, on la desmielinització és una característica predominant. Els pocs increments en FA podrien indicar regions on hi ha una compensació o canvis estructurals adaptatius, suggerint que certes àrees del cervell poden estar intentant adaptar-se als danys estructurals.

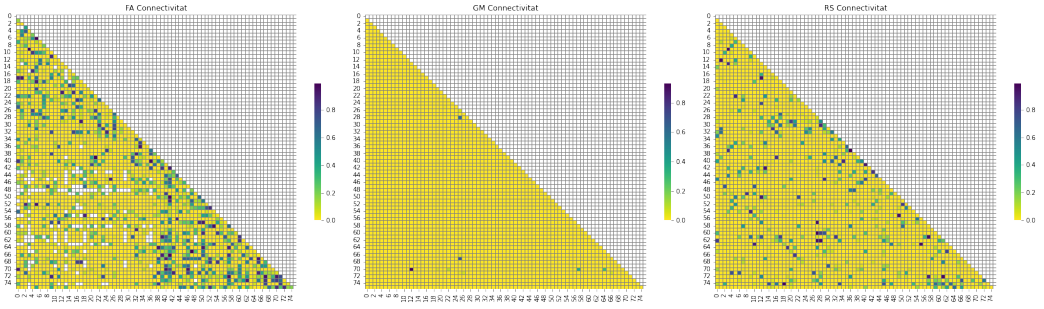
Pel que fa a GM (Matèria Grisa), les regions amb disminució de substància grisa podrien correspondre a atrofia o pèrdua de volum de substància grisa, un fenomen que sovint s'observa en pacients amb EM. La reducció de la substància grisa està associada amb el deteriorament cognitiu i altres símptomes clínics de l'EM, com ara problemes de memòria i funció executiva. Això reflecteix l'impacte que l'esclerosis múltiple té en les àrees del cervell implicades en funcions cognitives crítiques.

En la matriu de RS (connectivitat funcional en estat de repòs), la reducció de connectivitat funcional observada en els pacients pot reflectir una desconnexió funcional entre regions cerebrals, cosa que podria estar relacionada amb la simptomatologia clínica de l'esclerosis múltiple, com la fatiga o la dificultat cognitiva. Les zones amb un augment de connectivitat podrien indicar, de nou, mecanismes de compensació o reorganització funcional, on certes àrees del cervell intenten adaptar-se als dèficits funcionals.

Com a conclusió general, els gràfics de diferències mostren canvis significatius en la connectivitat cerebral entre controls i pacients amb esclerosis múltiple. Aquests resultats suggereixen que la malaltia afecta tant la connectivitat estructural com funcional, amb una disminució generalitzada en la integritat de les fibres de substància blanca, atrofia de substància grisa, i alteracions en la connectivitat funcional. Aquestes troballes són coherents amb el coneixement actual sobre com l'esclerosis múltiple impacta el cervell, oferint una visió més detallada dels canvis neurològics associats amb la malaltia.

Estudi de la connectivitat

En la figura 4.2, es poden observar tres matrius de connectivitat generals, els eixos representen diferents regions cerebrals (nodes) dins d'una xarxa de connectivitat cerebral. Cada cel·la de la matriu indica la força de la connectivitat dels diferents tipus (FA, GM i RS) entre dues regions específiques del cervell. Les cel·les que tenen un color més groguenc indiquen un valor alt i per tant suggereix una alta integritat de la connectivitat entre els nodes, d'altra banda les cel·les que tenen un color més porpra indiquen un valor més baix i per tant una integritat més baixa. Els colors verd i turquesa són més intermedis.



(a) FA **(b)** GM **(c)** RS

Figure 4.2: Matrius de connectivitat

En la matriu de FA (Anisotropia Fraccional), es poden observar diverses connexions amb valors de connectivitat elevats, indicats per colors groguencs. Això suggereix una alta integritat de les fibres de matèria blanca entre aquestes regions cerebrals. Tanmateix, també hi ha zones amb valors més baixos, representades en colors verdosos o porpra, que podrien indicar una disminució en la integritat de la matèria blanca. Aquesta variabilitat en la connectivitat de FA pot estar relacionada amb les diferències individuals en la integritat estructural de la matèria blanca.

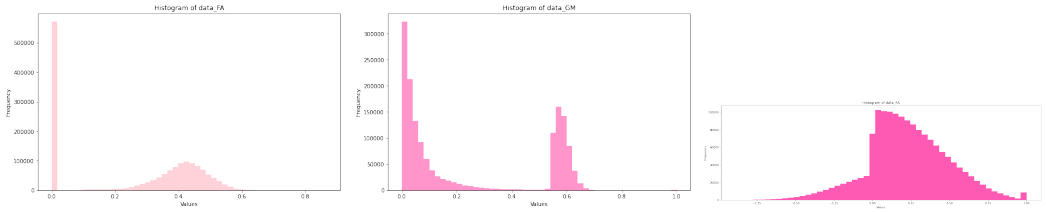
La matriu de GM (Matèria Grisa), presenta una distribució més homogènia de valors de connectivitat, predominantment en tons groguencs, el que indica una alta integritat i una estructura cerebral relativament estable en aquestes regions. No obstant això, hi ha algunes cel·les amb colors més foscos, el que suggereix possibles àrees amb menor densitat o volum de matèria grisa. Aquestes diferències poden estar associades amb processos neurodegeneratius o amb variacions individuals en la composició de la matèria grisa generats per l'esclerosis múltiple.

En la matriu de RS (connectivitat funcional en estat de repòs), s'observa una certa variabilitat en els patrons de connectivitat entre diferents regions del cervell. Les zones amb colors groguencs indiquen una connectivitat funcional forta, mentre que les regions amb colors més foscos, com el porpra, podrien reflectir una desconnexió funcional o una menor interacció entre aquestes àrees en estat de repòs. Aquesta variabilitat en la connectivitat RS pot estar relacionada amb la capacitat del cervell per mantenir comunicació funcional entre regions clau, fins i tot quan no s'està realitzant cap tasca específica tot i que es pateixi la malaltia.

Com a conclusió general d'aquestes matrius de connectivitat, es pot observar que la connectivitat cerebral varia entre les diferents regions i tipus de dades analitzades (FA, GM i RS). Les regions amb alta integritat estructural o funcional són evidents, així com les àrees amb menor connectivitat que podrien estar associades amb alteracions o deficiències neurològiques generades per l'esclerosis múltiple.

Histogrames

Les figures mostren els histogrames dels valors generals obtinguts en les dades de FA (Anisotropia Fraccional), GM (Matèria Grisa) i RS (connectivitat funcional en estat de repòs). Aquests histogrames s'han utilitzat per establir un umbral (*threshold*) amb l'objectiu de filtrar i eliminar les arestes menys significatives en els grafs de connectivitat cerebral. Aquest *threshold* s'ha escollit amb la funció `calculate_threshold`, i s'ha comprovat mitjançant els histogrames si tenia sentit.



(a) FA **(b)** GM **(c)** RS

Figure 4.3: Histogrames

L'histograma de FA (Anisotropia Fraccional), mostra una distribució de valors amb una gran concentració prop de zero, indicant que moltes de les connexions tenen un valor d'anisotropia fraccional molt baix. Aquest pic a zero pot suggerir la presència de connexions que no són rellevants o que tenen poca integritat estructural. A mesura que els valors augmenten, la freqüència disminueix, creant una distribució asimètrica amb una cua a la dreta. Aquesta informació s'ha utilitzat per determinar un umbral que permeti eliminar les arestes amb valors baixos, preservant així les connexions més fortes i potencialment més significatives en el graf. La funció `calculate_threshold` ha situat l'umbral dels grafs entre 0.3 i 0.4, el qual es correlaciona completament amb el pic que hi ha a l'histograma.

L'histograma de GM (Matèria Grisa) , presenta una distribució bimodal amb dos pics, un a prop de zero i un altre al voltant de 0.6. Aquestes dades suggereixen que hi ha dues poblacions de valors: una amb densitats de matèria grisa molt baixes, que poden correspondre a connexions menys importants o soroll, i una altra amb valors més alts, que probablement representen connexions significatives. A partir d'aquesta distribució, s'ha establert un umbral per a GM que permet filtrar les connexions amb valors baixos, assegurant que només es mantinguin les connexions més rellevants en l'anàlisi dels grafs. La funció `calculate_threshold` ha situat l'umbral dels grafs entre 0.6 i 0.7, el qual es correlaciona completament amb el pic que hi ha a l'histograma.

En l'histograma de RS (connectivitat funcional en estat de repòs), s'observa una distribució més simètrica amb un pic al voltant de 0, però que inclou una gamma de valors negatius i positius. La major part dels valors estan en el rang positiu baix, però també hi ha presència de valors en els extrems negatius i positius, indicant variabilitat en la connectivitat funcional en estat de repòs. Aquest patró ha estat utilitzat per determinar un umbral que elimina les connexions més dèbils o inestables, centrant-se en les connexions que tenen un paper més significatiu en la xarxa funcional cerebral. A diferència de la resta d'histogràmes aquest no es entre 0 i 1, sino entre -1 i 1. De nou s'ha utilitzat la funció `calculate_threshold` tot i que en aquest cas es on més ha variat el graf escollit tot i això la majoria han estat al voltant del 0.4, de forma sorprenent, ja que el pic està una mica abans. Això probablement sigui degut a que el rang de x es més gran que en els altres dos casos.

Aquests histogrames han estat crucials per establir umbrals que permetin filtrar les arestes dels grafs de connectivitat cerebral, eliminant les connexions menys rellevants i mantenint aquelles que podrien ser més significatives en l'anàlisi. Això ajuda a simplificar i clarificar la xarxa, centrant l'atenció en les connexions que tenen un impacte més gran en la interpretació de les dades i en la comprensió de les alteracions neurològiques. Gràcies a aquests histogrames s'han pogut reduir el nombre d'arestes de cada graf de 2800 a un umbral de desde 400 fins a 700.

4.1.2 Mètriques dels grafs

Un cop creats els diferents grafs utilitzant la llibreria *networkx*, es poden contar 270 grafs (un per cada persona) cadascun de 76 nodes (regions del cervell), de cadascun dels tres tipus (FA, GM i RS). D'aqueststst grafs i nodes s'extreuen diferents mètriques per poder-les utilitzar en diferents models, de nou amb l'ajut de la llibreria *NetworkX*.

Dels nodes:

- **Degree (Grau):** El grau d'un node és el nombre de connexions o arestes que té amb altres nodes. En el context d'una xarxa cerebral, indica quantes altres regions cerebrals estan connectades amb una determinada regió. Serveix per identificar nodes importants que poden actuar com a centres de comunicació o punts de connexió en la xarxa. Un grau alt indica una alta interconnexió.
- **Strength (Força):** La força d'un node és la suma dels pesos de les arestes connectades a aquest node. La força mesura la intensitat total de les connexions d'un node, per tant, permet identificar quins nodes tenen una connectivitat més forta, la qual cosa pot indicar regions cerebrals amb una alta activitat o influència.
- **Eigenvector Centrality (Centralitat de Vector Propi):** Mesura la influència d'un node en una xarxa, valorant no només el nombre de connexions sinó també la importància dels nodes connectats. És útil per identificar nodes crítics o regions clau en xarxes on es poden destacar regions amb gran influència per la seva connexió amb altres regions influents.
- **Betweeness Centrality (Centralitat d'Intermediació):** Mesura quant un node actua com a pont o intermediari en el camí més curt entre altres dos nodes de la xarxa. Un node amb alta centralitat d'intermediació participa en molts dels camins més curts entre parelles de nodes. Identifica nodes que són crítics per a la comunicació o el flux d'informació en una xarxa. En xarxes cerebrals, pot ajudar a trobar regions clau per a la transmissió de senyals, on la seva disfunció podria interrompre la comunicació entre diverses parts del cervell.
- **Closeness Centrality (Centralitat de Proximitat):** Mesura com de prop està un node de tots els altres nodes de la xarxa, basant-se en la distància mitjana dels camins més curts. Un node amb alta centralitat de proximitat pot arribar a altres nodes amb menys passos. Ajuda a identificar nodes que poden accedir ràpidament a tota la resta de la xarxa. En el context del cervell, una alta centralitat de proximitat pot indicar regions que poden influir ràpidament sobre altres regions.

Dels grafs:

- **Average Degree (Grau Mitjà):** És la mitjana del nombre de connexions per node en la xarxa. Es calcula sumant els graus de tots els nodes i dividint per el nombre total de nodes. Indica el nivell general de connectivitat en la xarxa. Un grau mitjà alt suggereix que, de mitjana, cada regió cerebral està connectada amb moltes altres.
- **Average Clustering (Clúster Mitjà):** Mesura el grau en què els nodes d'una xarxa tendeixen a agrupar-se. Es calcula com la mitjana dels coeficients de clustering de tots els nodes, on el coeficient de clustering d'un node és la proporció de connexions existents entre els seus veïns comparat amb el nombre màxim possible de connexions entre aquests. Permet identificar la presència de subgrups o comunitats dins la xarxa. Un alt coeficient de clustering indica que els nodes tendeixen a formar clústers densos, la qual cosa pot ser important en l'organització funcional del cervell.
- **Density (Densitat):** És la proporció del nombre real de connexions en la xarxa respecte al nombre màxim possible de connexions. Es calcula com el nombre total d'arestes dividit pel nombre màxim possible d'arestes. Mesura la cohesió global de la xarxa. Una alta densitat indica que una gran proporció de nodes estan connectats entre si, cosa que pot reflectir una comunicació eficient dins de la xarxa.

4.2 Tècniques de preprocessament

4.2.1 Feature Scaling

StandardScaler

Les característiques numèriques han estat normalitzades utilitzant StandardScaler per garantir que totes tinguessin la mateixa escala, fet que millora el rendiment dels models.

El StandardScaler és una eina disponible a la llibreria scikit-learn de Python que s'utilitza per estandarditzar les característiques d'un conjunt de dades. Estandarditzar significa rescalar les dades perquè tinguin una mitjana de 0 i una desviació estàndard de 1. Això es fa per assegurar que cada característica tingui la mateixa escala, cosa que pot millorar el rendiment de molts algorismes de machine learning que són sensibles a la magnitud de les característiques.

El StandardScaler es basa en la següent fórmula per transformar acada característica X_i:

$$X_i' = \frac{X_i - \text{mean}(X_i)}{\text{std}(X_i)}$$

On:

- X_i és el valor original de la característica.
- $mean(X_i)$ és la mitjana dels valors de la característica X_i.
- $std(X_i)$ és la desviació estàndard dels valors de la característica X_i.

Després d'aplicar el StandardScaler, els valors transformats tindran una distribució amb mitjana 0 i desviació estàndard 1, el que és molt útil per aplicar models com la SVM o la regressió logística.

K-Fold Cross-Validation

Inicialment, abans d'obtenir les dades provinents de Nàpols, hi havia un gran desbalanceig en pacients i controls, ja que hi havia molts pacients, i molt pocs controls. Per aquest motiu s'havia fet *K-Fold Cross-Validation*, la qual és una tècnica de validació utilitzada en l'aprenentatge automàtic per avaluar el rendiment d'un model de manera més robusta.

El *K-Fold Cross-Validation* consisteix a dividir el conjunt de dades en K subconjunts o *folds*. El model s'entrena K vegades, cada vegada utilitzant un dels *folds* com a conjunt de validació i els altres K-1 *folds* com a conjunt d'entrenament. Al final, es calcula la mitjana de les mètriques de rendiment obtingudes en cadascuna de les K iteracions, proporcionant una estimació més fiable de la capacitat predictiva del model. Aquesta tècnica ajuda a evitar el sobreajustament i aprofita millor les dades disponibles. Un cop es van afegir les dades provinents de Nàpols ja no va ser necessari perquè no hi havia desbalanceig.

SMOTE

Tot i haver descartat utilitzar el *K-fold Cross-Validation* un cop afegides les noves dades hi havia alguns models que demanaven encara menys desbalanceig de les classes motiu pel qual es va utilitzar de forma ràpida l'SMOTE.

L'SMOTE, que en anglès son les sigles de *Synthetic Minority Over-sampling Technique*, és una tècnica utilitzada en l'aprenentatge automàtic per tractar el problema de classes desbalancejades. Quan una classe està subrepresentada en un conjunt de dades, SMOTE genera noves instàncies sintètiques d'aquesta classe minoritària mitjançant la interpolació entre exemples existents. Això ajuda a equilibrar les classes, millorant el rendiment dels models en predir la classe minoritària i reduint el biaix cap a la classe majoritària.

4.2.2 Shuffle

L'última tècnica de preprocessament que s'ha utilitzat abans d'entrenar els models ha estat la comanda `shuffle`. Aquesta comanda ha estat necessària perquè les dades de Nàpols, on tots els voluntaris eren controls, estaven situades a continuació de les dades provinents de l'Hospital Clínic. Si no es barrejaven, aquesta organització podria haver induït biaixos en els models, fent-los menys generalitzables i més susceptibles a sobreajustar-se als patrons específics de les dades de control.

Amb l'ús de shuffle, s'ha aconseguit barrejar de manera aleatòria els elements d'una llista o conjunt de dades, alterant l'ordre original dels elements. Això proporciona una seqüència nova que és una permutació aleatòria de l'original, barrejant les dades abans de dividir-les en entrenament i prova. D'aquesta manera, es garanteix una distribució més homogènia i representativa de les dades durant l'entrenament, minimitzant el risc de biaixos i millorant el rendiment del model.

Aquesta barreja és especialment important per assegurar que el model no aprengui patrons espuris associats a la ubicació original dels dades, sinó que es concentri en les característiques rellevants per a la tasca de classificació.

4.3 Entrenament dels Models

Un cop ja estava tot preparat, amb els datasets desbiaixats i les dades barrejades correctament, s'ha fet l'entrenament dels models.

Els models s'han entrenat amb les dades d'entrada i cada mètrica dels nodes i grafs per separat, la y sempre ha estat el mateix, que es l'*mstype* de l'arxiu *demographics.csv* que es segons el que volem classificar (si son pacients o controls). Els resultats es podran veure al capítol següent.

S'ha aplicat el *train_test_split* de *Scikit-learn* per dividir les dades correctament en els dos subgrups: un conjunt d'entrenament i un conjunt de prova. Aquest procés és fonamental en l'entrenament de models de machine learning, ja que permet avaluar el rendiment del model en dades que no s'han utilitzat per entrenar-lo, proporcionant així una indicació de com generalitzarà el model a dades noves. S'han destinat el 80% de les dades al conjunt d'entrenament i el 20% restant al de prova. També s'ha aplicat `random_state=42` per garantir que cada cop que s'executi el codi hi hagi reproduïbilitat dels resultats, és a dir, que sempre siguin els mateixos.

SVM

L'SVM (Support Vector Machine) és un algorisme d'aprenentatge automàtic utilitzat principalment per a tasques de classificació. El seu objectiu és trobar un hiperplà òptim que separi les dades en dues classes diferents, maximitzant la distància (marge) entre les dades més properes de cada classe i l'hiperplà. SVM pot treballar en espais lineals i no lineals, utilitzant "nuclis" per transformar les dades i fer-les separables. És conegut per la seva efectivitat en conjunts de dades d'alta dimensionalitat i per ser robust contra el sobreajustament.

Logstic Regression

La regressió logística és un mètode d'aprenentatge automàtic utilitzat per a la classificació binària. Modela la probabilitat que una observació pertanyi a una de les dues classes possibles, utilitzant una funció sigmoide que transforma els valors d'entrada en una probabilitat entre 0 i 1. A partir d'un llindar (per exemple, 0,5), es decideix a quina classe pertany l'observació. Tot i el seu nom, és un mètode de classificació, no de regressió, i és especialment útil per predir esdeveniments binaris com sí/no, veritat/fals, o present/absent.

4.3.1 Embeddings, RandomWalk i DeepWalk

4.4 gnns

4.4.1 graphsage

Chapter 5

Resultats

En aquest capítol s'explicaran els resultats obtinguts de tots els diferents mètodes que s'han plantejat al capítol de Metodologia.

5.1 Entrenament dels Models

Primerament, abans de donar els resultats obtinguts a l'entremanet dels models es definiran els continguts que s'utilitzen.

5.1.1 Accuracy

L'*accuracy*, anomenada en català precisió, és una mesura utilitzada per avaluar el rendiment d'un model de classificació. Representa la proporció de prediccions correctes que el model ha realitzat sobre el total de prediccions fetes. És una de les mètriques més senzilles i comunes per avaluar models de classificació.

Un accuracy alta propera a 100 indica que el model és bó classificant correctament, tot i això té certes limitacions, especialment quan es treballa amb conjunts de dades desbalancejats, on una classe és molt més freqüent que les altres. En aquests casos, un model que sempre prediu la classe majoritària pot tenir un accuracy alt, però no serà útil en la pràctica perquè no identificarà correctament les classes menys freqüents.

L'accuracy es calcula amb la següent fórmula:

$$\text{Accuracy} = \frac{\text{Número de prediccions correctes}}{\text{Total de prediccions fetes}}$$

O bé:

$$\text{Accuracy} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

On:

- VP (Veritables Positius): Nombre de casos positius correctament classificats com a positius.
- VN (Veritables Negatius): Nombre de casos negatius correctament classificats com a negatius.
- FP (Falsos Positius): Nombre de casos negatius incorrectament classificats com a positius.
- FN (Falsos Negatius): Nombre de casos positius incorrectament classificats com a negatius.

5.1.2 Precision

La *Precision*, o en català precisió, és una mètrica utilitzada en l'avaluació de models de classificació. Representa la proporció de prediccions positives correctes respecte al total de prediccions que el model ha fet com a positives. En altres paraules, la precision mesura quantes de les prediccions fetes pel model que van ser etiquetades com a positives realment ho eren.

Si la precision és alta, significa que quan el model prediu una classe positiva, té una alta probabilitat d'haver encertat (és a dir, d'haver classificat correctament). La diferència amb l'accuracy és que només contempla les classes positives mentre que l'accuracy contempla les positives i negatives.

Per tant, la precisió es calcula amb la següent fórmula:

$$\text{Precisió} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

On:

- VP (Veritables Positius): Nombre de casos positius correctament classificats com a positius.
- FP (Falsos Positius): Nombre de casos negatius incorrectament classificats com a positius.

5.1.3 Recall

El Recall, en català sensibilitat o veritable taxa positiva, és una mètrica utilitzada per avaluar models de classificació. Representa la proporció de casos positius reals que han estat correctament identificats pel model. És una mesura de la capacitat del model per detectar tots els casos positius dins d'un conjunt de dades.

Un *recall* alt indica que el model està identificant correctament la majoria dels casos positius (pocs falsos negatius). Això és especialment important en situacions on és crucial detectar tots els casos positius, com en la detecció de malalties o fraus. En canvi si el recall és baix indica que el model està passant per alt molts casos positius, el que significa que hi ha un alt nombre de falsos negatius.

La fórmula pel *recall* és:

$$\text{Accuracy} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

On:

- VP (Veritables Positius): Nombre de casos positius correctament classificats com a positius.
- FN (Falsos Negatius): Nombre de casos positius incorrectament classificats com a negatius.

5.1.4 F1 Score

L'F1 Score és una mesura utilitzada en l'avaluació de models de classificació, especialment útil quan hi ha un desbalanceig en les classes. L'F1 Score combina la precisió (*precision*) i el *recall* en una única mètrica, donant una idea del rendiment del model en termes de falsos positius i falsos negatius.

L'F1 Score oscil·la entre 0 i 1, on 1 indica el millor rendiment possible del model (alta precisió i recall). Pel que un valor d'F1 Score baix indica que el model té problemes per identificar correctament les classes positives, o bé que està fent moltes prediccions incorrectes com a positives.

L'F1 Score es calcula amb la següent fórmula:

$$\text{F1 Score} = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

5.1.5 Anàlisi General

Primerament, s'ha dut a terme un anàlisi general, on s'ha posat com a coordenada X les dades de de les carpetes combinades de l'Hospital Clínic i les provinents de Nàpols. Com a coordenada y s'ha posat el *mstype* de *demographics.csv*. S'han utilitzat els dos models explicats previament. La classe 0 es refereix als controls mentre que la classe 1 es refereix als pacients que tenen esclerosis múltiple.

SVM

S'han analitzat els datasets combinats de FA, GM i RS amb el *mstype*. S'ha analitzat l'accuracy, la precisió, el F1 score i el recall. El *report* de classificació que s'ha obtingut ha estat el següent:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
FA	0.8519	0.85	0.85	0.85	0.85	0.85	0.85
GM	0.8642	0.89	0.84	0.83	0.90	0.86	0.87
RS	0.926	0.97	0.89	0.88	0.97	0.92	0.93

Primerament, per a FA, el model mostra un rendiment consistent en ambdues classes amb valors d'accuracy, precision, recall i F1-Score tots iguals al 85_. Això indica que el model és equilibrat en la seva capacitat per predir tant la classe 0 com la classe 1, però amb un rendiment global que és més aviat moderat

Seguidament per a GM el model mostra un rendiment lleugerament millor que FA amb una accuracy del 86.42%. La classe 0, que es la dels controls, té una millor precision (0.89), però un recall més baix (0.83), mentre que la classe 1, la dels malalts, té una precisió lleugerament inferior (0.84) però un millor recall (0.90). Això suggereix que GM és més equilibrat en la detecció de la classe 1, però encara presenta una petita variabilitat en la classe 0.

Per últim per a RS el model mostra el millor rendiment global, amb una accuracy del 92.59%. Té una precision i un F1-Score molt alts per a ambdues classes. Especialment, el model és excel·lent en la detecció de la classe 1 (recall de 0.97), cosa que significa que identifica gairebé tots els casos positius correctament. Aquest model és clarament superior als altres dos en termes d'equilibri entre precision i recall.

Logstic Regression

S'ha fet el mateix estudi que anteriorment però en aquest cas utilitzant la regressió logística. El report de classificació ha estat el següent:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
FA	0.8642	0.92	0.82	0.80	0.93	0.86	0.87
GM	0.9136	0.95	0.88	0.88	0.95	0.91	0.92
RS	0.9383	1.0	0.89	0.88	1.0	0.94	0.94

Primerament, per a FA el model mostra un bon equilibri entre les dues classes, amb una millor precisió per a la classe 0 però millor sensibilitat per a la classe 1. L'accuracy global és bona (86.42%), però el model tendeix a identificar millor la classe 1.

Seguidament per a GM el model presenta una millora respecte al primer, amb una accuracy de 91.36%. Les mètriques són equilibrades i molt bones per a ambdues classes, indicant que el model és molt eficient en la detecció i classificació de les dues classes.

Per últim per a RS, es de nou on s'obtenen millors resultats, amb una accuracy de 93.83%. Té una precisió perfecta per a la classe 0 (1.00), encara que la sensibilitat és lleugerament més baixa. Per a la classe 1, la sensibilitat és màxima (1.00), la qual cosa significa que identifica correctament tots els casos de la classe 1. Aquest model és el més fiable dels tres en termes de rendiment general.

5.1.6 Anàlisi de nodes de FA

Les mètriques dels nodes que s'han utilitzat han estat el *degree*, l'*strength*, la *closeness centrality*, la *betwenness centrality* i la *eigenvector centrality*.

SVM

Utilitzant l'SVM els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Degree	0.8765	0.90	0.86	0.85	0.90	0.88	0.88
Strength	0.8395	0.91	0.79	0.76	0.93	0.83	0.85
Closeness Centr.	0.8025	0.82	0.79	0.78	0.82	0.80	0.80
Betwenness Centr.	0.8642	0.88	0.85	0.85	0.88	0.86	0.86
Eigenvector Centr.	0.8395	0.89	0.80	0.78	0.90	0.83	0.85

Com a conclusió, el model basat en degree va obtenir la millor *accuracy* general amb un 0.8765, mostrant un bon equilibri entre precisió i recall per ambdues classes. D'altra banda, el model basat en strength va aconseguir una alta precisió per a la classe 0 (0.91), però el seu recall per a la classe 1 va ser lleugerament inferior (0.85), amb una *accuracy* general de 0.8395. El model basat en *closeness centrality* va ser el més baix en rendiment, amb una *accuracy* de 0.8025 i una precisió relativament baixa per a la classe 0 (0.82). El model basat en *betweenness centrality* va obtenir una accuracy de 0.8642, destacant per la seva consistència en precisió i recall per ambdues classes. Finalment, el model basat en e*eigenvector centrality* va mostrar un rendiment moderat, amb una *accuracy* de 0.8395, mantenint una precisió i recall relativament equilibrats per a totes dues classes. En conjunt, aquests resultats indiquen que les diferents mesures de centralitat dels nodes poden influir significativament en el rendiment dels models SVM per a la classificació de malalts i no malalts, amb *degree* i *betweenness centrality* mostrant un rendiment globalment més fort en aquesta tasca.

Logstic Regression

Utilitzant una regressió logística els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Degree	0.8395	0.83	0.85	0.85	0.82	0.84	0.84
Strength	0.8272	0.83	0.82	0.83	0.82	0.83	0.82
Closeness Centr.	0.8395	0.85	0.83	0.83	0.85	0.84	0.84
Betwenness Centr.	0.8888	0.90	0.88	0.88	0.90	0.89	0.89
Eigenvector Centr.	0.8518	0.89	0.82	0.80	0.90	0.85	0.86

Com a conclusió, el model basat en *betweenness centrality* destaca com el més efectiu, amb una *accuracy* de 0.8888 i un equilibri excel·lent entre precisió, *recall* i F1-Score per a ambdues classes, especialment per a la detecció de malalts (classe 1) amb un F1-Score de 0.89. D'altra banda, els models basats en *degree* i *closeness centrality* tenen una *accuracy* de 0.8395, mantenint un rendiment acceptable, però amb limitacions en la precisió i F1-Score, especialment en la classe 1. El model basat en *eigenvector centrality* ofereix una *accuracy* de 0.8518, destacant-se lleugerament per sobre dels anteriors en algunes mètriques, però sense arribar al nivell de *betweenness centrality*. Finalment, el model basat en *strength* és el que presenta el rendiment més baix, amb una *accuracy* de 0.8272, i mostra dificultats notables tant en la precisió com en el *recall* per a ambdues classes, especialment en la detecció de la classe 1. En resum, *betweenness centrality* es revela com la mesura més adequada per a la classificació de malalts i no malalts mitjançant regressió logística en aquest conjunt de dades, mentre que *strength* resulta ser la menys efectiva.

5.1.7 Anàlisi de nodes de GM

En aquest cas s'han utilizat les dades de la substància grisa.

SVM

Utilitzant l'SVM els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Degree	0.9259	0.95	0.90	0.90	0.95	0.92	0.93
Strength	0.9012	0.95	0.86	0.85	0.95	0.90	0.90
Closeness Centr.	0.8519	0.82	0.89	0.90	0.80	0.86	0.84
Betwenness Centr.	0.8765	0.92	0.84	0.83	0.93	0.87	0.88
Eigenvector Centr.	0.9259	0.97	0.89	0.88	0.97	0.92	0.93

Com a resum per a les dades de la substància grisa es destaca que els models basats en *degree* i *eigenvector centrality* ofereixen els millors resultats, amb una *accuracy* de 0.9259 cadascun. Aquests models mantenen un alt rendiment tant en precisió com en recall, aconseguint un F1-Score de 0.93 per a la classe 1, cosa que els fa especialment efectius per a la detecció de malalts. El model basat en strength també mostra un rendiment excel·lent, amb una *accuracy* de 0.9012 i un equilibri gairebé perfecte entre precisió i recall per a ambdues classes. Per contra, el model basat en *closeness centrality* és el que presenta el rendiment més baix, amb una *accuracy* de 0.8519 i un F1-Score més modest, especialment per a la classe 1. Finalment, el model basat en *betweenness centrality* ofereix un rendiment sòlid amb una *accuracy* de 0.8765, mantenint un bon equilibri entre les mètriques, però sense arribar als nivells dels models basats en *degree*, *strength*, o *eigenvector centrality*. En resum, *degree* i *eigenvector centrality* es confirmen com les millors mesures de centralitat per a la classificació de malalts en aquest conjunt de dades, mentre que closeness centrality és la menys efectiva.

Logstic Regression

Utilitzant una regressió logística els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Degree	0.9259	0.97	0.89	0.88	0.97	0.92	0.93
Strength	0.9136	0.97	0.87	0.85	0.97	0.91	0.92
Closeness Centr.	0.9012	0.92	0.88	0.88	0.93	0.90	0.90
Betwenness Centr.	0.9136	0.97	0.87	0.85	0.97	0.91	0.92
Eigenvector Centr.	0.9383	1.0	0.89	0.88	1.0	0.94	0.94

Com a conclusió es destaca que el model basat en *eigenvector centrality* és el més efectiu, amb una *accuracy* de 0.9383 i una precisió perfecta per a la classe 0 (1.0). Aquest model també aconsegueix un F1-Score molt alt de 0.94 per a la classe 1, mostrant una excel·lent capacitat de classificació tant per a malalts com per a no malalts. Els models basats en *degree, strength, closeness centrality* i *betweenness centrality* també mostren un rendiment molt bo, amb *accuracies* que oscil·len entre 0.9136 i 0.9259, i mantenint un equilibri excel·lent entre precisió, *recall* i F1-Score per a ambdues classes. Particularment, els models basats en *strength* i *betweenness centrality* comparteixen una *accuracy* de 0.9136, destacant per un alt F1-Score de 0.92 per a la classe 1, cosa que els fa molt fiables per a la detecció de casos positius. En resum, *eigenvector centrality* es presenta com la millor mesura de centralitat per a la classificació mitjançant regressió logística en aquest conjunt de dades, oferint el rendiment més consistent i alt en totes les mètriques considerades, mentre que les altres mesures també ofereixen resultats molt competitius.

5.1.8 Anàlisi de nodes de RS

En aquest cas s'han utilitzat les dades en estat de repós.

SVM

Utilitzant l'SVM els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Degree	0.7407	0.71	0.79	0.83	0.65	0.76	0.71
Strength	0.6666	0.63	0.72	0.80	0.53	0.71	0.61
Closeness Centr.	0.8765	0.84	0.92	0.93	0.82	0.88	0.87
Betwenness Centr.	0.6790	0.67	0.68	0.71	0.65	0.69	0.67
Eigenvector Centr.	0.6666	0.67	0.67	0.68	0.65	0.67	0.66

En l'anàlisi dels models SVM aplicats al conjunt de dades RS, s'observa que el model basat en *closeness centrality* és el que ofereix el millor rendiment, amb una *accuracy* de 0.8755 i un F1-Score equilibrat per a ambdues classes, destacant especialment en la detecció de malalts (classe 1) amb un F1-Score de 0.87. Els models basats en degree i betweenness centrality ofereixen un rendiment moderat, amb *accuracy* de 0.7407 i 0.6790, respectivament, mantenint un equilibri acceptable entre precisió i *recall*, però sense arribar al nivell de closeness centrality. D'altra banda, els models basats en *strength* i *eigenvector centrality* tenen el rendiment més baix, amb una *accuracy* de 0.6666 cadascun, i presenten dificultats per identificar correctament la classe 1, tal com es reflecteix en els seus F1-Scores més baixos. En resum, *closeness centrality* és la mesura més efectiva per a la classificació de malalts i no malalts en aquest conjunt de dades, mentre que strength i eigenvector centrality són les menys eficients.

Logstic Regression

Utilitzant una regressió logística els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Degree	0.7777	0.74	0.82	0.85	0.70	0.80	0.86
Strength	0.6914	0.66	0.74	0.80	0.57	0.73	0.65
Closeness Centr.	0.8516	0.84	0.87	0.88	0.82	0.86	0.85
Betwenness Centr.	0.7407	0.75	0.73	0.73	0.75	0.74	0.74
Eigenvector Centr.	0.7407	0.72	0.77	0.80	0.68	0.76	0.72

Com a conclusió, *closeness centrality* és el que mostra el millor rendiment global, amb una *accuracy* de 0.8516 i un bon equilibri entre precisió, *recall* i F1-Score per a ambdues classes, especialment destacant-se en la identificació de malalts (classe 1) amb un F1-Score de 0.86. Per contra, el model basat en strength és el que presenta el rendiment més baix, amb una *accuracy* de 0.6914 i un F1-Score molt baix de 0.65 per a la classe 1, indicant dificultats importants en la classificació correcta dels casos positius.

Els models basats en *degree*, *betweenness centrality*, i *eigenvector centrality* presenten un rendiment intermedi, amb accuracies al voltant de 0.7407 a 0.7777, però amb limitacions evidents en el *recall* i F1-Score, especialment per a la classe 1, cosa que suggereix una menor efectivitat en la detecció de malalts. En resum, mentre que closeness centrality sembla ser la millor opció per a la classificació mitjançant regressió logística en aquest conjunt de dades, les altres mesures de centralitat mostren un rendiment inferior, amb *strength* sent la menys efectiva per a la tasca de classificació de malalts.

5.1.9 Anàlisi de grafs de FA

Les mètriques dels grafs que s'han utilitzat han estat l'*average degree*, l'*average clustering* i la *density*.

SVM

Utilitzant l'SVM els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Av Degree	0.9259	1.0	0.87	0.85	1.0	0.92	0.93
Av Clustering	0.8888	0.94	0.84	0.83	0.95	0.88	0.89
Density	0.9259	1.0	0.87	0.85	1.0	0.92	0.93

Com a conclusió, s'observa que les mesures *average degree* i *density* ofereixen els millors rendiments, amb una *accuracy* general de 0.9259 cadascuna. Aquests models també mostren una precisió perfecta per a la classe 0 (1.0) i un excel·lent equilibri entre recall i F1-Score per a ambdues classes, aconseguint un F1-Score de 0.93 per a la classe 1, la qual cosa indica una alta capacitat de detecció tant de malalts com de no malalts. La mètrica *average clustering* presenta un rendiment lleugerament inferior, amb una *accuracy* de 0.8888 i un F1-Score de 0.89 per a la classe 1, tot i que segueix mantenint una precisió i recall robustos.

Logstic Regression

Utilitzant una regressió logística els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Av Degree	0.9136	0.97	0.85	0.85	0.97	0.91	0.92
Av Clustering	0.8519	0.87	0.83	0.83	0.88	0.85	0.85
Density	0.9136	0.97	0.87	0.85	0.97	0.91	0.92

En resum les mesures *average degree* i *density* es destaquen com les més efectives, amb una *accuracy* general de 0.9136 cadascuna. Aquests models mostren un alt rendiment en totes les mètriques, amb una precisió de 0.97 per a la classe 0 i un F1-Score de 0.92 per a la classe 1, indicant una excel·lent capacitat de classificació tant per a malalts com per a no malalts. D'altra banda, la mètrica *average clustering* presenta un rendiment inferior, amb una *accuracy* de 0.8519 i un F1-Score de 0.85 per a la classe 1, tot i mantenir una bona precisió i *recall*, però sense arribar al nivell de les altres dues mètriques.

5.1.10 Anàlisi de grafs de GM

SVM

Utilitzant l'SVM els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Av Degree	0.9012	1.0	0.83	0.80	1.0	0.89	0.91
Av Clustering	0.8765	1.0	0.80	0.76	1.0	0.86	0.89
Density	0.9012	1.0	0.83	0.80	1.0	0.89	0.91

En resum les mesures *average degree* i *density* mostren rendiments molt similars i alts, amb una *accuracy* general de 0.9012 cadascuna. Ambdues mètriques aconsegueixen una precisió perfecta per a la classe 0 (1.0) i un equilibri consistent entre *recall* i F1-Score per a ambdues classes, amb un F1-Score de 0.91 per a la classe 1. Això indica que aquests models són molt efectius en la classificació tant de malalts com de no malalts. D'altra banda, la mètrica *average clustering* presenta un rendiment lleugerament inferior, amb una *accuracy* de 0.8765 i un F1-Score de 0.89 per a la classe 1, mantenint encara una bona precisió i *recall*, però amb una menor capacitat de detecció comparada amb les altres dues mètriques.

Logstic Regression

Utilitzant una regressió logística els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Av Degree	0.9012	1.0	0.83	0.80	1.0	0.89	0.91
Av Clustering	0.9012	1.0	0.83	0.80	1.0	0.89	0.91
Density	0.9012	1.0	0.83	0.80	1.0	0.89	0.91

En conclusió, les tres mesures avaluades (*average degree*, *average clustering* i *density*) mostren un rendiment idèntic en termes d'accuracy general, amb un valor de 0.9012 cadascuna. Totes les mètriques aconsegueixen una precisió perfecta per a la classe 0 (1.0) i un bon equilibri entre *recall* i F1-Score per a ambdues classes, amb un F1-Score de 0.91 per a la classe 1. Això indica que aquestes mètriques són igualment efectives en la classificació tant de malalts com de no malalts quan s'utilitza regressió logística. Tot i que totes les mètriques ofereixen un rendiment alt i consistent, no hi ha una diferència clara que faci que una d'elles sobresurti significativament sobre les altres en aquest conjunt específic d'anàlisi.

5.1.11 Anàlisi de grafs de RS

SVM

Utilitzant l'SVM els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Av Degree	0.8765	0.88	0.88	0.88	0.88	0.88	0.88
Av Clustering	0.7531	0.82	0.71	0.66	0.85	0.73	0.77
Density	0.8765	0.88	0.88	0.88	0.88	0.88	0.88

En resum cal destacar que les mesures *average degree* i *density* ofereixen els millors rendiments, amb una *accuracy* general de 0.8765 cadascuna. Ambdues mètriques mostren un equilibri consistent entre precisió, *recall* i F1-Score per a ambdues classes, aconseguint un F1-Score de 0.88 per a la classe 1, indicant una bona capacitat per classificar tant malalts com no malalts. D'altra banda, la mètrica *average clustering*, de nou, presenta un rendiment inferior, amb una *accuracy* de 0.7531 i un F1-Score de 0.77 per a la classe 1, reflectint una menor efectivitat en la detecció de casos positius en comparació amb les altres dues mètriques.

Logstic Regression

Utilitzant una regressió logística els resultats son els següents:

Mètrica	**Accuracy**	**Precisió**		**Recall**		**F1 score**	
Classe	General	Classe 0	Classe 1	Classe 0	Classe 1	Classe 0	Classe 1
Av Degree	0.8642	0.86	0.87	0.88	0.85	0.87	0.86
Av Clustering	0.7654	0.79	0.74	0.73	0.80	0.76	0.77
Density	0.8642	0.86	0.87	0.88	0.85	0.87	0.86

En conclusió, s'observa que les mesures *average degree* i *density* ofereixen un rendiment superior, amb una *accuracy* general de 0.8612 cadascuna. Ambdues mètriques mostren un bon equilibri entre precisió, *recall* i F1-Score per a ambdues classes, amb un F1-Score de 0.86 per a la classe 1, indicant una capacitat efectiva per classificar tant malalts com no malalts. En canvi, la mètrica *average clustering* presenta un rendiment inferior, amb una *accuracy* de 0.7654 i un F1-Score de 0.77 per a la classe 1, reflectint una menor efectivitat en la detecció de casos positius comparada amb les altres dues mètriques. Per tant, *average degree* i *density* es consoliden com les mètriques més eficients per a la classificació amb regressió logística, mentre que *average clustering* mostra un rendiment notablement inferior, especialment en la detecció de malalts.

5.1.12 embeddings randomwalk i deepwalk

5.2 gnns

5.2.1 graphsage

Chapter 6

Conclusions

Chapter 7

References

[1] Labonne, M. (2023). *Hands-on graph neural networks using Python: practical techniques and architectures for building powerful graph and deep learning apps with Pytorch*. Packt Publishing. ISBN: 978-1804617526. Disponible a: https://github.com/PacktPublishing/Hands-On-Graph-Neural-Networks-Using-Python.

[2] Casas-Roma, J., Martinez-Heras, E., Solé-Ribalta, A., Solana, E., Lopez-Soley, E., Vivó, F., Diaz-Hurtado, M., Alba-Arbalat, S., Sepulveda, M., Blanco, Y., Saiz, A., Borge-Holthoefer, J., Llufriu, S., & Prados, F. (2022). Applying multilayer analysis to morphological, structural, and functional brain networks to identify relevant dysfunction patterns. *Network Neuroscience*. Disponible a: https://direct.mit.edu/netn/article/6/3/916/111665/Applying-multilayer-analysis-to-morphological.

[3] Solana, E., Martinez-Heras, E., Casas-Roma, J., Calvet, L., Lopez-Soley, E., Sepulveda, M., Sola-Valls, N., Montejo, C., Blanco, Y., Pulido-Valdeolivas, I., Andorra, M., Saiz, A., Prados, F., & Llufriu, S. (2019). Modified connectivity of vulnerable brain nodes in multiple sclerosis, their impact on cognition and their discriminative value. *Scientific Reports*. Disponible a: https://www.nature.com/articles/s41598-019-56806-z.

[4] Lozano-Bagén, T., Martinez-Heras, E., Solana, E., Garrido-Romero, S., Llufriu, S., Prados, F., & Casas-Roma, J. (2023). Characterization of Brain Networks through the lens of Persistent Homology. *Pending publication or journal placeholder*.

[5] Scikit-learn. *f1_score - Scikit-learn*. Disponible a: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html.

[6] Scikit-learn. *precision_score - Scikit-learn*. Disponible a: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html.

[7] Scikit-learn. *recall_score - Scikit-learn*. Disponible a: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html.

[8] Scikit-learn. *accuracy_score - Scikit-learn*. Disponible a: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html.

[9] Datacamp. *Comprehensive Introduction to Graph Neural Networks (GNNs) Tutorial*. Disponible a: https://www.datacamp.com/tutorial/comprehensive-introduction-graph-neural-networks-gnns-tutorial.

[10] ResearchGate. *Illustration of sampling and aggregation in GraphSAGE method*. Disponible a: https://www.researchgate.net/figure/Illustration-of-sampling-and-aggregation-in-GraphSAGE-method-A-sample-of-neighboring_fig1_351575091.

[11] Towards Data Science. *Graph Convolutional Networks: Introduction to GNNs*. Disponible a: https://towardsdatascience.com/graph-convolutional-networks-introduction-to-gnns-24b3f60d6c95.

[12] Geekflare. *Graph Neural Networks*. Disponible a: https://geekflare.com/graph-neural-networks/.

[13] Espindola, G. *Què són els embeddings i com s'utilitzen en la intel·ligència artificial amb Python*. Disponible a: https://gustavo-espindola.medium.com/qu%C3%A9-son-los-embeddings-y-c%C3%B3mo-se-utilizan-en-la-inteligencia-artificial-con-python-45b751ed86a5.

[14] Abhyuday, T. *DeepWalk and Node2Vec: Graph Embeddings*. Disponible a: https://medium.com/@tejpal.abhyuday/deep-walk-and-node2vec-graph-embeddings-faf02d369442.

[15] GeeksforGeeks. *Random Walk*. Disponible a: https://www.geeksforgeeks.org/random-walk/.

[16] Towards Data Science. *Logistic Regression Explained*. Disponible a: https://towardsdatascience.com/logistic-regression-explained-9ee73cede081.

[17] Viquipèdia. *Màquina de vector de suport*. Disponible a: https://ca.wikipedia.org/wiki/M%C3%A0quina_de_vector_de_suport.

[18] Viquipèdia. *Regressió logística*. Disponible a: https://ca.wikipedia.org/wiki/Regressi%C3%B3_log%C3%ADstica.

[19] MathWorks. *Support Vector Machine*. Disponible a: https://es.mathworks.com/discovery/support-vector-machine.html.

[20] National Multiple Sclerosis Society. *Definition of MS*. Disponible a: https://www.nationalmssociety.org/What-is-MS/Definition-of-MS.

[21] Mayo Clinic. *Multiple Sclerosis: Symptoms & Causes*. Disponible a: https://www.mayoclinic.org/diseases-conditions/multiple-sclerosis/symptoms-causes/syc-20350269.

[22] MSIF. *About MS*. Disponible a: https://www.msif.org/about-ms/.

[23] Scikit-learn. *OneToOneFeatureMixin*. Disponible a: https://scikit-learn.org/stable/modules/generated/sklearn.base.OneToOneFeatureMixin.html.

[24] Scikit-learn. *Support Vector Machines*. Disponible a: https://scikit-learn.org/stable/modules/svm.html.

[25] Scikit-learn. *Logistic Regression*. Disponible a: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression.

[26] NetworkX. *NetworkX: Overview*. Disponible a: https://networkx.org/.

[27] NetworkX. *Graph.degree*. Disponible a: https://networkx.org/documentation/stable/reference/classes/generated/networkx.Graph.degree.html.

[28] NetworkX. *Average Clustering*. Disponible a: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.average_clustering.html.

[29] NetworkX. *Density Function*. Disponible a: https://networkx.org/documentation/stable/reference/generated/networkx.classes.function.density.html.

[30] NetworkX. *Betweenness Centrality in Bipartite Graphs*. Disponible a: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.bipartite.centrality.betweenness_centrality.html#networkx.algorithms.bipartite.centrality.betweenness_centrality.

[31] NetworkX. *Degree Centrality in Bipartite Graphs*. Disponible a: `https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.bipartite.centrality.degree_centrality.html#networkx.algorithms.bipartite.centrality.degree_centrality`.

[32] NetworkX. *Closeness Centrality in Bipartite Graphs*. Disponible a: `https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.bipartite.centrality.closeness_centrality.html#networkx.algorithms.bipartite.centrality.closeness_centrality`.

Appendix A

CALDRIA POSAR EL CODI?

Aqui podeu posar qualsevol cosa que no s'ho mirarà ningú