



This is the **published version** of the bachelor thesis:

Bruix Fernández, Martí; Casas Roma, Jordi, dir. Anàlisi de xarxes per caracteritzar malalties neurodegeneratives. 2024. (Grau en Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/298917>

under the terms of the  license

Anàlisi de Xarxes per Caracteritzar Malalties Neurodegeneratives

Martí Bruix Fernández

Resum—A partir d'imatges de ressonància magnètica del cervell, es pot obtenir informació sobre la connectivitat entre les seves regions, la qual pot ajudar a la identificació de persones que pateixen Esclerosi Múltiple. En aquest treball, s'estudien diverses estratègies per resoldre aquest problema, tant combinant informació estructural, morfològica i funcional del cervell per crear una xarxa multilayer com treballant amb xarxes single-layer que continguin un sol tipus d'informació. A partir de les xarxes, s'ha fet el càlcul de diferents mètriques de graf i s'han utilitzat els valors obtinguts per entrenar models d'aprenentatge automàtic que classifiquin els subjectes entre persones amb Esclerosi Múltiple i voluntaris sans, arribant a aconseguir un accuracy de fins a 0,965 en el cas del millor model generat, que es tracta d'un Support Vector Machine. A part, també s'han generat models de regressió per predir el valor de la mètrica EDSS, que indica l'estat de discapacitat de la persona.

Paraules clau—Esclerosi Múltiple, MRI, xarxa multilayer, xarxa single-layer, EDSS, aprenentatge automàtic, Support Vector Machine, Random Forest, Multi-layer Perceptron.

Abstract— Using magnetic resonance imaging of the brain, information about the connectivity between its regions can be obtained, which can help in identifying people with Multiple Sclerosis. In this study, various strategies to solve this problem are explored, both by combining structural, morphological and functional information of the brain to create a multilayer network and by working with single-layer networks that contain a single type of information. From these networks, various graph metrics were calculated, and the obtained values were used to train machine learning models that classify subjects between people with Multiple Sclerosis or healthy volunteers, achieving an accuracy of up to 0.965 with the best generated model, which is a Support Vector Machine. Additionally, regression models were also generated to predict the value of the EDSS metric, which indicates the disability status of the person.

Index Terms—Multiple Sclerosis, MRI, multilayer network, single-layer network, EDSS, machine learning, Support Vector Machine, Random Forest, Multi-layer Perceptron.



1 INTRODUCCIÓ - CONTEXT DEL TREBALL

L'Esclerosi Múltiple és una malaltia inflamatòria i neurodegenerativa crònica del sistema nerviós central que condueix a una discapacitat física i cognitiva. L'ús de mètodes d'aprenentatge automàtic pot ajudar a la detecció de persones que pateixen aquesta malaltia, per això mateix, s'han realitzat molts estudis durant els últims anys explorant diverses estratègies per avançar en aquest camp [1].

Una de les eines que més s'utilitzen a l'hora d'estudiar aquesta malaltia són les imatges obtingudes per ressonància magnètica (MRI), ja que permeten analitzar l'estructura del cervell i obtenir bastanta informació rellevant. Per exemple, se'n poden extreure xarxes de connectivitat que indiquin la connexió que hi ha entre cada una de les 76 regions definides del cervell humà. El valor de la connexió entre dues regions concretes pot tenir diferents significats, de forma que existeixen diversos tipus de xarxes de connectivitat, segons el que representi una connexió. Tres dels

exemples que es poden obtenir a partir de les MRIs són les xarxes de connectivitat estructural o FA (*Fractional Anisotropy*), xarxes d'associació morfològica o GM (*Gray Matter*) i xarxes de connectivitat funcional o RS (*Resting State*).

En diversos estudis s'ha fet servir la informació proporcionada per aquestes xarxes per tal de detectar l'Esclerosi Múltiple, però normalment només s'ha fet analitzant un tipus de xarxa individualment. Per tant, una de les estratègies que s'han proposat recentment és generar una xarxa multilayer [2], que és una estructura de dades formada per diverses capes, de forma que a cada capa es guardin les dades de cada una de les tres xarxes mencionades, podent combinar d'aquesta forma diferents tipus d'informació útil. A partir d'aquesta estructura multilayer, es poden aplicar tècniques d'anàlisi de xarxes per tal de calcular diverses mètriques de graf i identificar si amb els valors obtinguts és possible diferenciar entre persones amb Esclerosi Múltiple (*people with Multiple Sclerosis*, PwMS) i voluntaris sans (*healthy volunteers*, HV).

- E-mail de contacte: marti.bruix@autonoma.cat
- Menció realitzada: Computació
- Treball tutoritzat per: Jordi Casas Roma (Ciències de la Computació)
- Curs 2023/24

2 OBJECTIUS

Per dur a terme aquest treball, es disposa d'un dataset que conté informació sobre diferents individus, tant de tipus PwMS com de HV. Per cada individu, es tenen tres matrius que representen els tres tipus de xarxes, és a dir, les xarxes de connectivitat estructural (FA), morfològica (GM) i funcional (RS), amb els valors de les connexions ja obtinguts a partir de les MRIs.

Amb aquestes dades podem explorar l'estratègia de la creació d'una xarxa multilayer i comparar-la amb la de l'ús de xarxes single-layer, és a dir xarxes d'una sola capa utilitzant únicament un tipus d'informació concreta. També podem, a partir de mètriques de graf obtingudes de les xarxes, intentar generar models d'aprenentatge automàtic de classificació per distingir entre PwMS i HV.

A part, per cada individu també disposem dels valors de la mètrica EDSS (*Expanded Disability Status Scale*), la qual indica amb una escala del 0 al 10 quin és l'estat de discapacitat de la persona, sent pitjor com més alt és el valor. Per tant, també tenim l'oportunitat d'explorar si a partir de les xarxes de connectivitat cerebral és viable predir el valor de l'EDSS i podem intentar generar algun model de regressió que solucioni aquest problema.

Tenint en compte tot això, aquest és el llistat concret d'objectius que s'han identificat de cara a aquest treball:

1. Realitzar un anàlisi estadístic de les dades inicials per fer-nos una idea de quines regions o connexions del cervell ens poden ser més útils de cara a aquest estudi.
2. Estudiar el funcionament de les estructures multilayer i implementar alguna solució en aquest sentit a partir dels tres tipus de xarxes de connectivitat cerebral.
3. Obtenir diverses mètriques de graf a partir de les xarxes de connectivitat cerebral i identificar quines són les més rellevants.
4. Entrenar diversos models, fent servir diferents tècniques d'aprenentatge automàtic i conjunts de dades, analitzar els resultats obtinguts i fer una comparació entre ells per tal de quedar-nos amb el que solucioni el problema de millor forma.

3 PLANIFICACIÓ

3.1 Metodologia

La metodologia en la qual ens basarem per desenvolupar aquest treball és la CRISP-DM [3]. Es tracta del mètode més utilitzat a dia d'avui en projectes de *data science*, i indica quines haurien de ser generalment les fases que conformen un projecte d'aquest tipus:

1. Comprensió del problema
2. Comprensió de les dades
3. Preparació de les dades
4. Modelatge
5. Avaluació
6. Desplegament

Tot i així, és una metodologia bastant flexible, la qual es pot adaptar al context de cada projecte, ja sigui fent salts intermedis entre fases o realitzant diverses iteracions del procés fins a complir els objectius inicials. La figura 1 mostra un esquema amb les fases i el cicle de treball habitual

de la metodologia CRISP-DM.

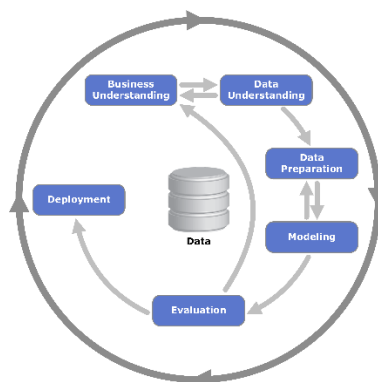


Fig. 1. Esquema de la metodologia CRISP-DM

A més a més, per portar un control de les tasques que es van realitzant i el que queda per fer s'utilitzarà també la metodologia Kanban [4]. Aquesta consisteix en tenir un tauler on s'hi poden veure les tasques del projecte col·locades en columnes segons el seu estat, com per exemple "A Fer", "En Progrés" o "Finalitzat".

A part d'això, també es disposarà d'un repositori de GitHub amb el codi del projecte, per facilitar-ne la compartició i transparència. S'hi pot accedir a través del següent enllaç: https://github.com/martibruix/Codi_TFG

3.2 Fases del treball

A continuació, es mostra la taula 1, que conté les fases en les quals s'ha estructurat el desenvolupament del projecte i la data d'inici i de finalització planejada per cada una.

A la figura A1 de l'apèndix, per altra banda, es pot veure el diagrama de Gantt amb la planificació temporal de totes aquestes fases.

TAULA 1
FASES DEL PROJECTE

	Fase	Inici	Final
1	Revisió de l'estat de l'art	11/3	24/3
2	Exploració inicial de les dades	25/3	31/3
3	Preparació de les dades	1/4	14/4
4	Obtenció de les mètriques de graf	15/4	28/4
5	Implementació de models	29/4	19/5
6	Anàlisi de resultats	20/5	26/5
7	Implementació de millores	27/5	5/6
8	Finalització del projecte	6/6	16/6

4 ESTAT DE L'ART

En aquest apartat, es farà una revisió d'estudis amb un context similar al d'aquest treball. S'exploraran els dos tipus d'enfocaments que podem considerar: classificació per determinar si es tracta de PwMS o HV i predicció del valor de l'EDSS.

4.1 Classificació entre PwMS i HV

Començarem analitzant els estudis que s'han basat en l'ús de les xarxes de connectivitat cerebral per solucionar el problema. En el primer d'aquests [5], l'estratègia va ser

treballar a partir de la informació sobre la connectivitat estructural, és a dir amb les xarxes FA. A partir de les xarxes, es van analitzar diferents mètriques de graf per cada node i es van agafar les que millor diferenciaven entre PwMS i HV, que va resultar que eren local efficiency i node strength. Es van utilitzar els valors d'aquestes dues mètriques per entrenar un Support Vector Machine, aconseguint un accuracy de 77,15% utilitzant només local efficiency, d'un 74,84% utilitzant només node strength i d'un 76,88% fent servir les dues mètriques.

Més endavant, en un altre estudi [2], es va provar de combinar la informació sobre la connectivitat morfològica (xarxes GM), la connectivitat estructural (xarxes FA) i la connectivitat funcional (xarxes RS). A partir d'aquestes tres xarxes es va dissenyar una xarxa multilayer. Aquest tipus d'estructura està formada per diverses capes, i conté dos tipus diferents de connexions: *intralayer links*, que representen la relació entre nodes d'una mateixa capa, i *interlayer links*, que relacionen nodes de diferents capes. En el cas d'aquest estudi, la xarxa multilayer estava formada per 2 capes, que eren les xarxes GM i RS, mentre que la xarxa FA es va utilitzar per definir els *interlayer links*. A la figura 2 podem veure tant l'estructura de la xarxa multilayer com de la matriu d'adjacència que la representa. Posteriorment, es va realitzar un anàlisi estadístic sobre cinc mètriques de graf: strength, degree, betweenness centrality, closeness centrality i local efficiency. L'anàlisi va concloure que la mètrica que millor diferenciava entre grups era local efficiency, però també servien el closeness centrality, degree i strength. Tot i així, a l'estudi no s'arriben a entrenar models d'aprenentatge automàtic a partir d'aquestes mètriques.

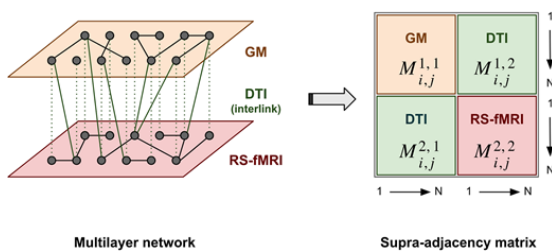


Fig. 2. Estructura d'una xarxa multilayer (Font: [2])

No és l'únic estudi que ha intentat combinar diferents dades de connectivitat cerebral. En un altre cas [6], es va provar de fer la classificació a partir de les dades de connectivitat estructural (FA) i funcional (RS). Es van generar diferents models de Support Vector Machine, un a partir de la matriu de connectivitat FA, un altre a partir de la matriu RS i un altre amb una combinació normalitzada de les dues matrius. En cada cas, es van agafar com a característiques únicament els valors de les matrius més rellevants per dividir els grups, per evitar una dimensionalitat massa alta. El model que pitjor va funcionar va ser el que utilitzava la matriu FA, obtenint un accuracy de 72,5%, mentre que amb la matriu RS es va obtenir un accuracy de 85,7% i el millor model va ser el que combinava les dues matrius, obtenint un accuracy de 87,8%. En aquest estudi també es va provar de classificar entre PwMS amb un EDSS ≤ 1,5 i PwMS amb un EDSS > 1,5, però obtenint uns resultats

bastant pitjors.

Una altra estratègia que s'ha estudiat és treballar directament a partir de les imatges MRI, mitjançant eines de Deep Learning [7]. Els principals mètodes que s'han utilitzat en aquest sentit són diferents tipus de Convolutional Neural Networks, com per exemple 2-D CNN, 3-D CNN, U-Net o Generative Adversarial Networks (GAN). Un dels avantatges de l'ús de tècniques de Deep Learning és que tots els passos relacionats amb la obtenció i la selecció de característiques no són necessaris, ja que això es fa de forma automàtica a través de les capes profundes dels models. Els resultats obtinguts amb aquest enfocament són millors que els que utilitzen tècniques de Machine Learning més tradicionals, però com a desavantatge, aquestes arquitectures solen necessitar una quantitat de dades més gran, cosa que passa amb pocs datasets actualment.

4.2 Predicció de l'EDSS

En quant a estudis que tinguin com a objectiu entrenar models per predir el valor de l'EDSS, son menys habituals que els de classificació. Tot seguit se'n mostren dos de rellevants.

En el primer cas [8], va ser un estudi fet a partir de les dades que contenen les notes clíniques dels pacients, com per exemple informació sobre la pèrdua de mobilitat, de vista, ús de medicació... Es van entrenar dos models de XGBoost com a regressió acompanyats d'un node de decisió binaria. Com a resultats, es va obtenir un Spearman R value de 0,75 i un Pearson R value de 0,74 al comparar els EDSS obtinguts amb els valors reals. També es va avaluar el resultat de forma binària, establint la classe positiva com a EDSS ≥ 6 i la classe negativa com a EDSS < 6, obtenint un Positive Predictive Value de 0,85, Negative Predictive Value de 0,85 i una AUC de 0,91.

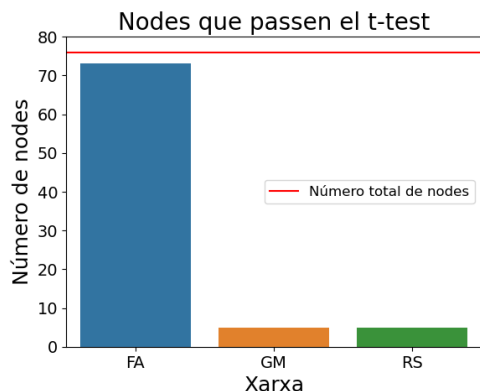
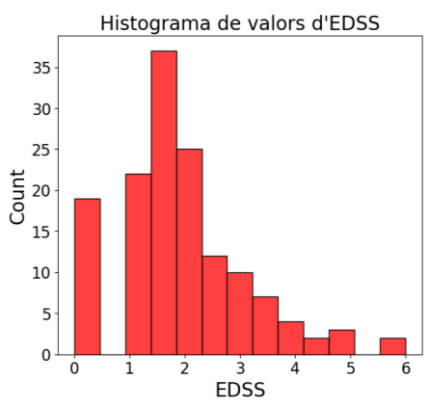
En el segon exemple [9], es van provar diferents models i estratègies per fer la predicció de l'EDSS. Les dades de les quals es disposava eren l'edat i el gènere dels subjectes i FLAIR MRIs. El primer model que es va generar va ser utilitzant una Convolutional Neural Network (CNN) sobre les dades. Una altra estratègia va ser aplicar una reducció de la dimensionalitat per entrenar dos models més: un amb un regressor Random Forest i l'altre amb l'algoritme UMAP (Uniform Manifold Approximation and Projection). Finalment, es va generar un últim predictor utilitzant una mitjana ponderada dels outputs de tots els altres models, aconseguint un MSE de 2,21 sobre el conjunt de validació, i obtenint després un MSE de 3 provant-ho sobre un nou conjunt de test.

5 ANÀLISI DE LES DADES

5.1 Descripció de les dades

Per realitzar aquest estudi s'han fet servir dades de pacients de l'Hospital Clínic de Barcelona. Aquestes han passat pel comitè d'ètica i tots els pacients han signat un consentiment permetent el seu ús. Disposem d'una cohort de 143 persones, de les quals 18 són HV i 125 són PwMS. Els valors de la mètrica EDSS es troben en el rang [0, 6]. A la figura 3 podem veure la distribució dels valors de l'EDSS, mentre que a la taula 2 es descriuen les característiques

dels subjectes de l'estudi.



TAULA 2
DADES DEMOGRÀFIQUES DELS SUBJECTES

	HV (n=18)	PwMS (n=125)
Edat	36,62 ± 9,6	45,66 ± 9,48
Homes, n (%)	3 (16,67%)	35 (28%)
Dones, n (%)	15 (83,33%)	90 (72%)
EDSS	-	2,11 ± 1,1

Per l'edat i l'EDSS es mostra la mitjana i la desviació estàndard, mentre que en el cas del gènere es mostra el nombre d'ocurrències i el percentatge sobre el total.

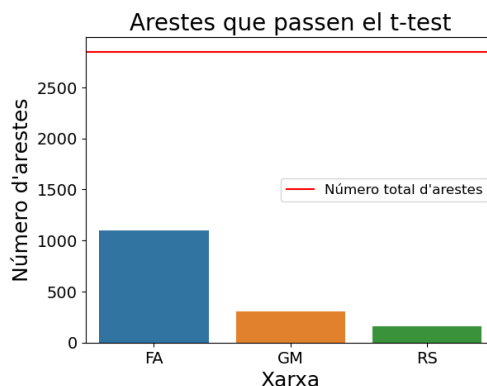
Cadascun dels subjectes té associades tres matrius d'adjacència, les quals corresponen als tres tipus de xarxes cerebrals: les FA, les GM i les RS. Les tres son de mida 76x76, degut a les 76 regions del cervell establertes, i cada posició de la matriu representa el pes de l'aresta que connecta dos nodes concrets. En els tres casos, tots els pesos es troben dins del rang [0, 1], i ja venen amb una correcció aplicada per efectes de l'edat i del gènere. A l'apartat A2 de l'apèndix, concretament a les figures A2.1, A2.2 i A2.3, es poden veure els histogrames amb els pesos de totes les arestes per cadascun dels tres tipus de xarxes.

5.2 Anàlisi estadístic

El primer que s'ha fet és un anàlisi inicial de les dades amb l'objectiu de saber quines ens poden ser més útils i fer-nos una idea de si caldrà aplicar algun tipus de preprocessament.

Primer de tot, s'ha intentat identificar quins nodes són més rellevants de cara a distingir entre HV i PwMS, calculant per cada node el sumatori dels pesos de totes les arestes connectades, és a dir la mètrica strength, i duent a terme un t-test on la hipòtesi nul·la és que les mostres de l'strength de node pels grups HV i PwMS tenen una mitjana idèntica. Concretament, s'ha realitzat el t-test de Welch [10], ja que no podem assegurar que les variàncies de les dues mostres siguin semblants. Per poder passar el test s'ha establert que s'havia d'obtenir un p-value inferior a 0,05. Aquest anàlisi s'ha realitzat a cadascuna de les tres xarxes, obtenint que a les FA 73 dels 76 nodes han passat el t-test, però en canvi a les GM i a les RS tan sols 5 nodes l'han superat. Es pot veure de forma gràfica a la figura 4.

També s'ha analitzat quines arestes són més rellevants per distingir entre HV i PwMS duent a terme l'estratègia del t-test, en aquest cas fent-ne un per cada aresta i on la hipòtesi nul·la és que els pesos de l'aresta pels grups HV i PwMS tenen una mitjana idèntica. Al igual que abans, s'ha realitzat el t-test de Welch. De les 2.850 arestes que hi ha en total, 1.095 han passat el t-test a les xarxes FA, 306 a les xarxes GM i 159 a les xarxes RS. Es pot veure de forma gràfica a la figura 5.



Per últim, s'ha analitzat també la relació entre les xarxes i els valors de l'EDSS. S'ha fet, al igual que abans, a partir de la mètrica strength dels nodes, però en aquest cas, en comptes de realitzar un t-test, s'ha calculat per cada node la correlació entre l'strength i l'EDSS. A les xarxes FA, la mitjana de les correlacions obtingudes a cada node ha sigut de 0,3, a les GM de 0,136 i a les RS de 0,071. Es pot veure de forma gràfica a la figura 6.

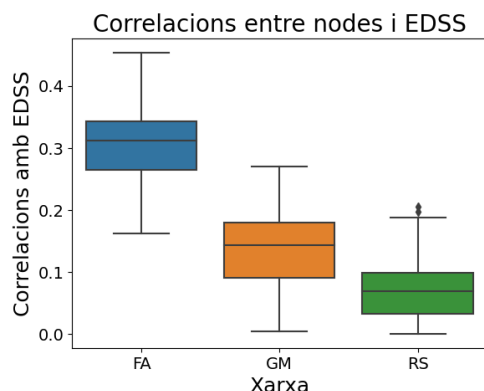


Fig. 6. Boxplot de les correlacions entre cada node i l'EDSS per cada xarxa.

Als tres anàlisis realitzats, les xarxes FA són les que clarament han obtingut uns millors resultats, per tant podem pensar que són les que contenen informació més rellevant de cara a solucionar tant el problema de classificació com el de regressió.

6 PREPARACIÓ DE LES DADES

En aquest apartat presentarem un parell d'estratègies de preprocessament que s'han dut a terme i s'indicaran les diferents combinacions de xarxes amb les que es treballarà.

6.1 Aplicació de llindars

A l'apartat anterior hem pogut veure que ni els nodes ni les arestes de les xarxes GM i RS a priori donen gaire informació rellevant. Si analitzem els histogrames dels pesos de les arestes (els quals es poden veure a l'apartat A3 de l'apèndix) veiem que, sobretot en el cas de les GM, la majoria de valors són bastant propers al 0 i hi ha poca variància entre ells. Una idea per solucionar aquest problema seria aplicar un llindar a aquests valors, de forma que els que quedin per sota d'una frontera determinada siguin transformats a 0, establint així una major diferència entre els pesos. Per tant, s'ha aplicat aquesta estratègia a les xarxes GM i RS.

Per cada xarxa, s'han provat diferents valors com a llindar, i per cada un d'ells s'ha tornat a realitzar el t-test sobre la mètrica strength de cada node, quedant-nos amb el que fes que més nodes superessin el test. En el cas de les xarxes GM, el llindar que millor ha funcionat ha sigut el de 0,04, amb el qual s'ha incrementat el número de nodes que superen el t-test de 5 a 8. A les RS, en canvi, cap dels llindars que s'han provat han millorat els resultats respecte els valors per defecte, que eren de 5 nodes que superaven el t-test. Per tant, a aquestes xarxes no li apliquem cap tipus de llindar.

6.2 Selecció d'arestes rellevants

Un altre factor que pot estar empitjorant l'habilitat de les xarxes per diferenciar entre HV i PwMS és que són altament connexes i és possible que moltes de les arestes existents no ens aportin informació útil. Per tant, una altra estratègia que s'ha provat és eliminar totes les arestes que no ens siguin útils. Per cada un dels tres tipus de xarxes

originals, crearem una nova xarxa que només contingui les arestes rellevants. Com a criteri de selecció, ens quedarem únicament amb les arestes que han superat el t-test que s'ha fet a l'apartat 5.2 i que per tant han provat ser rellevants. D'aquesta forma, les xarxes FA passaran a tenir 1.095 arestes, les xarxes GM 306 arestes i les xarxes RS 159 arestes.

6.3 Creació de les xarxes

A l'hora de seleccionar amb quines dades treballarem, hi ha diferents decisions que podem prendre:

- Podem escollir entre quedar-nos amb les tres xarxes i crear una estructura multilayer, utilitzar-ne només dues o inclús fer servir només una de les tres xarxes. En el cas de crear una estructura multilayer es farà basat en el disseny definit a [2] i que s'ha explicat a l'estat de l'art.
- En el cas de seleccionar les xarxes GM, podem utilitzar-les amb el llindar aplicat o amb els valors originals.
- També podem escollir si agafar les xarxes originals amb totes les arestes o agafar únicament les arestes rellevants.

Per poder realitzar diferents proves i comparar els resultats entre ells, s'ha decidit seleccionar fins a set conjunts diferents de les dades amb els que treballar, combinant les possibles decisions que es podien prendre:

1. Agafar únicament les xarxes FA, amb els valors originals, degut a que segons els anàlisis duts a terme són les que contenen informació més rellevant.
2. Crear una xarxa multilayer a partir dels tres tipus de xarxes, amb els valors originals i amb totes les arestes seleccionades.
3. Crear una xarxa multilayer a partir dels tres tipus de xarxes, però en aquest cas els valors de les xarxes GM tindran el llindar aplicat.
4. Crear una xarxa multilayer a partir dels tres tipus de xarxes, però en aquest cas mantenint únicament les arestes rellevants a cada una de les capes.
5. Crear una estructura single-layer amb les xarxes FA mantenint només les arestes rellevants.
6. Crear una estructura single-layer amb les xarxes GM mantenint només les arestes rellevants.
7. Crear una estructura single-layer amb les xarxes RS mantenint només les arestes rellevants.

7 OBTENCIÓ DE LES MÈTRiques DE GRAF

Un cop definides les estructures de dades amb les que treballarem, el següent pas consisteix en escollir diferents mètriques de graf, calcular el seu valor per cada una de les xarxes que tenim i identificar quines de les mètriques ens poden ser útils a l'hora d'entrenar models.

7.1 Mètriques de graf

Totes les mètriques que s'han escollit es calculen a nivell de node, de forma que per cada xarxa s'obtingran 76 atributs quan es calculi una mètrica, cada un representant el valor d'un node concret. Per poder operar sobre les xarxes

i fer els càlculs de les mètriques s'ha fet servir la llibreria NetworkX [11], la qual proporciona diversos mètodes útils per instanciar grafs, obtenir propietats d'aquests i calcular mètriques i algorismes amb funcions optimitzades.

A continuació, es llisten quines són les mètriques de graf que s'han escollit, amb una breu explicació i especificant la seva fórmula en cada cas.

Strength: Sumatori dels pesos de totes les arestes connectades a un node. En el cas de les xarxes multilayer, es fa el sumatori amb totes les arestes que hi ha als diferents nivells.

$$s_i = \sum_j M_{i,j}$$

Degree centrality: Número d'arestes connectades a un node respecte al màxim possible. A diferència de la mètrica degree, el rang de possibles valors està entre 0 i 1. En el cas de les xarxes multilayer, es calcula considerant totes les arestes que hi ha als diferents nivells.

$$d_i = \frac{\Gamma(i)}{n-1}$$

On $\Gamma(i)$ és el nombre de veïns del node v_i .

Betweenness centrality: És la fracció de camins més curts entre cada parella de nodes que passen per un node concret.

$$b_i = \sum_{s,t \neq i} \frac{\sigma_{s,t}(i)}{\sigma_{s,t}}$$

On $\sigma_{s,t}$ és el nombre de camins més curts entre s i t i $\sigma_{s,t}(i)$ és el nombre de camins més curts entre s i t que passen pel node v_i .

Closeness centrality: És la inversa de la distància mitjana entre un node i tots els nodes accessibles des d'aquest.

$$c_i = \frac{1}{\sum_j d(i,j)}$$

On $d(i,j)$ és la distància entre els nodes v_i i v_j .

Clustering coefficient: Calculat a partir dels triangles que forma un node amb els seus veïns, considerant els pesos de les arestes.

$$CC_i = \frac{\sum_{v,w} (\hat{w}_{uv} \hat{w}_{uw} \hat{w}_{vw})^{1/3}}{\Gamma(i)(\Gamma(i)-1)}$$

On $\hat{w}_{uv}$ és el pes normalitzat de l'aresta uv segons el pes màxim del graf.

PageRank: És un algorisme dissenyat per classificar nodes dins d'una xarxa segons la seva importància, de forma que dona un valor a cada node amb una fórmula que s'aplica de forma iterativa.

$$PR_i = (1-d) + d \sum_{j=1}^n \frac{PR(j)}{\Gamma(j)}$$

On d és un factor de correcció i $PR(j)$ el valor de PageRank de cada node predecessor de v_i .

Local efficiency: Calcula la inversa de les distàncies entre els veïns d'un node.

$$LE_i = \frac{\sum_{j,k \in G_i} \frac{1}{d(j,k)}}{\Gamma(i)(\Gamma(i)-1)}$$

On G_i és el subgraf que conté tots els veïns del node v_i i $d(j,k)$ és la distància entre j i k .

Les fórmules de les mètriques seleccionades estan pensades per ser calculades en xarxes d'una sola capa, per tant, s'ha hagut d'escollir una estratègia sobre com fer els càlculs a les xarxes multilayer. En el cas de les mètriques basades en distàncies (betweenness centrality, closeness centrality i local efficiency), el que s'ha fet és crear una matriu de 76x76 on es guardin les distàncies mínimes entre cada parella de nodes, considerant tant les connexions de les dues capes com els *interlinks*. Aquesta estratègia està definida a [2]. Per altra banda, en el cas de les mètriques on els pesos representen força (clustering coefficient i PageRank), s'ha definit una matriu on a cada posició es guarda el pes màxim entre cada parella de nodes, ja que en aquests casos un pes més alt significa més connexió entre nodes. Per tant, a les fórmules anteriors s'operaria sobre la matriu de distàncies mínimes o sobre la de pesos màxims, depenent del cas.

7.2 Anàlisi dels valors obtinguts

Per cada una de les 7 estructures de dades que tenim hem calculat totes les mètriques mencionades a l'apartat anterior, obtenint un valor per cada node. Un cop fet això, s'ha analitzat la capacitat que té cada mètrica per distingir entre els grups HV i PwMS. S'ha fet d'una forma semblant a l'anàlisi estadístic inicial, realitzant a cada node un t-test establint un p-value de màxim 0,05 per superar-lo, i quant quants dels nodes superen el test. A la taula 3 es poden veure els resultats obtinguts, on cada columna correspon a una de les 7 estructures de dades i cada fila a una de les mètriques avaluades. Cada casella conté el nombre de nodes que han superat el t-test per la mètrica i la xarxa corresponent.

Podem veure que les mètriques que generalment han obtingut uns millors resultats són l'*strength*, el *degree centrality*, el *closeness centrality* i el *local efficiency*. En quant a les xarxes, amb els valors per defecte s'han obtingut millors resultats amb la FA com a single-layer que amb l'estructura multilayer, i tot i que la tècnica d'aplicar el llindar no ha acabat de funcionar, sí que hi ha hagut una millora significativa amb la selecció d'arestes rellevants. També ha sigut bastant útil la tècnica de la selecció d'arestes per les xarxes single-layer, sobretot en el cas de les FA, però també s'han obtingut resultats decents inclús amb les GM i les RS.

A les figures 7 i 8 podem veure dues gràfiques de barres, de les dues xarxes que millor han funcionat, on cada columna representa el nombre de nodes que han passat el t-test per una mètrica.

TAULA 3
NODES QUE SUPEREN EL T-TEST PER CADA MÈTRICA I XARXA

	1	2	3	4	5	6	7
Strength	73	28	28	75	76	57	47
Degree centrality	47	47	21	75	75	47	70
Betweenness centrality	21	8	8	17	23	5	23
Closeness centrality	76	35	35	74	76	18	60
Clustering coefficient	0	21	21	74	14	45	18
PageRank	26	7	7	46	39	27	33
Local efficiency	76	76	76	54	74	45	27

1 = Single-layer FA

2 = Multilayer estàndard

3 = Multilayer amb llinar

4 = Multilayer amb selecció d'arestes

5 = Single-layer FA amb selecció d'arestes

6 = Single-layer GM amb selecció d'arestes

7 = Single-layer RS amb selecció d'arestes

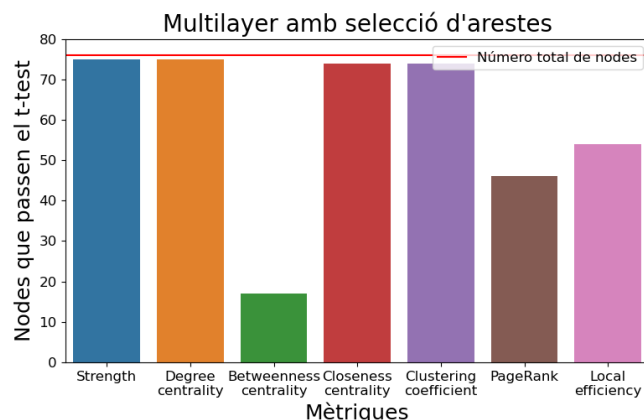


Fig. 7. Gràfica de barres amb el nombre de nodes que passen el t-test per cada mètrica a la xarxa multilayer amb selecció d'arestes.

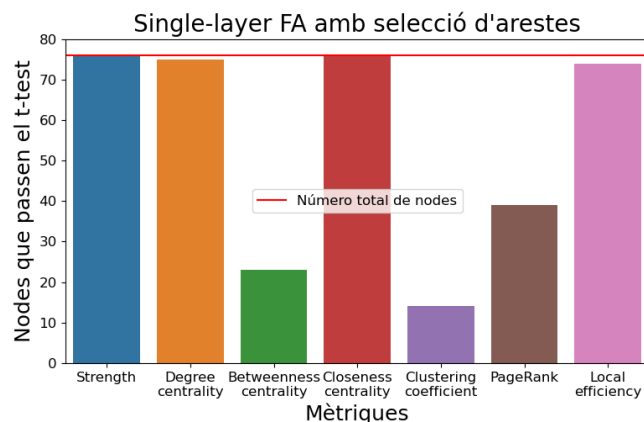


Fig. 8. Gràfica de barres amb el nombre de nodes que passen el t-test per cada mètrica a la xarxa single-layer FA amb selecció d'arestes.

També s'ha avaluat la capacitat de les mètriques

obtingudes per predir el valor de l'EDSS. S'ha seguit el mateix procediment que a l'apartat 5.2, obtenir pel valor de cada node la correlació que té amb el valor de l'EDSS i calcular a cada xarxa la mitjana de les correlacions obtingudes a tots els nodes. Els resultats obtinguts es poden veure a la taula 4. Al igual que abans, cada columna representa una estructura de dades, cada fila una mètrica i cada casella conté la mitjana de les correlacions per la mètrica i xarxa corresponents.

No s'han obtingut correlacions especialment altes en cap dels casos, la majoria de valors ronden el 0,1 o el 0,2 i poques vegades s'ha arribat a superar el 0,3. Les mètriques que millors resultats han obtingut en aquest cas són l'strength i el closeness centrality, seguits d'a prop pel local efficiency, mentre que les xarxes que han tingut millors resultats, al igual que al classificar segons els grups, han sigut la single-layer FA i la multilayer, amb selecció d'arestes en els dos casos.

TAULA 4
CORRELACIONS ENTRE ELS NODES I L'EDSS PER CADA MÈTRICA I XARXA

	1	2	3	4	5	6	7
Strength	0,30	0,19	0,19	0,34	0,34	0,13	0,13
Degree centrality	0,20	0,20	0,11	0,23	0,23	0,00	0,00
Betweenness centrality	0,11	0,09	0,09	0,08	0,10	0,01	0,08
Closeness centrality	0,28	0,17	0,17	0,34	0,32	0,12	0,12
Clustering coefficient	0,14	0,15	0,15	0,34	0,16	0,14	0,05
PageRank	0,13	0,07	0,07	0,14	0,13	0,13	0,10
Local efficiency	0,29	0,23	0,23	0,17	0,30	0,14	0,04

1 = Single-layer FA

2 = Multilayer estàndard

3 = Multilayer amb llinar

4 = Multilayer amb selecció d'arestes

5 = Single-layer FA amb selecció d'arestes

6 = Single-layer GM amb selecció d'arestes

7 = Single-layer RS amb selecció d'arestes

8 IMPLEMENTACIÓ DE MODELS

En aquest apartat es mostraran els diferents models d'aprenentatge automàtic que s'han implementat a partir de les mètriques de graf obtingudes, tant per fer la classificació entre HV i PwMS com per fer la predicció del valor de l'EDSS, s'exposaran els resultats de totes les proves i s'intentarà identificar el millor model obtingut.

S'han provat 3 mètodes diferents per resoldre el problema de classificació: SVM (Support Vector Machine), Random Forest i Multi-layer Perceptron. A part, per fer la predicció del valor de l'EDSS s'han provat SVM com a regressors. Per cada un dels mètodes mencionats hem generat 7 models, un per cada una de les xarxes amb les que estem treballant, per poder fer així una comparació entre

elles. Cada model que s'ha creat, però, no ha sigut passant-li tots els valors de les 7 mètriques com a atributs, sinó que s'ha seleccionat en cada cas només les mètriques que més relació tenien amb la variable objectiu, segons el que s'ha vist a l'apartat 7.2. Pels models de classificació, les mètriques seleccionades de cada xarxa són les següents:

Single-layer FA

- Strength, closeness centrality, local efficiency

Multilayer estàndard

- Degree centrality, closeness centrality, local efficiency

Multilayer amb lllindar

- Strength, closeness centrality, local efficiency

Multilayer amb selecció d'arestes

- Strength, degree centrality, closeness centrality, clustering coefficient

Single-layer FA amb selecció d'arestes

- Strength, degree centrality, closeness centrality, local efficiency

Single-layer GM amb selecció d'arestes

- Strength, degree centrality, clustering coefficient, local efficiency

Single-layer RS amb selecció d'arestes

- Strength, degree centrality, closeness centrality

Pels models de regressió, en canvi, el llistat de mètriques seleccionades és el següent:

Single-layer FA

- Strength, closeness centrality, local efficiency

Multilayer estàndard

- Strength, degree centrality, local efficiency

Multilayer amb lllindar

- Strength, closeness centrality, local efficiency

Multilayer amb selecció d'arestes

- Strength, closeness centrality, clustering coefficient

Single-layer FA amb selecció d'arestes

- Strength, closeness centrality, local efficiency

Single-layer GM amb selecció d'arestes

- Strength, closeness centrality, clustering coefficient

Single-layer RS amb selecció d'arestes

- Strength, closeness centrality, PageRank

També cal mencionar que per entrenar i testear els models s'ha fet servir cross validation, dividint en cada cas el dataset en 5 plecs. D'aquesta forma, ens assegurem que els resultats obtinguts tinguin la màxima fiabilitat possible, sobretot tenint en compte la baixa dimensionalitat del nostre dataset, que només conté 143 mostres. Per últim, s'ha realitzat en cada cas una cerca d'hiperparàmetres en grid per trobar la combinació que més millori els resultats del model. En el cas dels models de classificació, s'ha escollit com a mètrica a maximitzar la ROC AUC, ja que tenim molt poques mostres de la classe negativa i aquesta mètrica proporciona una millor avaluació del rendiment del model en la classificació de la classe minoritària. Pel model de regressió, la mètrica a maximitzar escollida ha sigut el MSE (Mean Squared Error).

8.1 Support Vector Machine (Classificació)

El primer mètode de classificació que s'ha provat ha sigut el SVM. Aquest és un tipus d'algoritme que sol funcionar millor quan els atributs estan en un rang similar, cosa que no passa amb totes les mètriques de graf que tenim. Per això mateix, s'han escalat els valors de totes les mètriques utilitzades en el rang [0,1]. Un altre problema trobat és el desequilibri de classes, ja que tenim 125 subjectes de la classe PwMS contra tan sols 18 de la classe HV. Per gestionar-ho, s'han generat els 5 plecs del dataset amb els quals fer cross validation de forma que hi hagi la mateixa proporció de mostres de cada classe a cada plec generat [12]. També s'ha indicat al model que ajusti els pesos de cada classe de forma inversament proporcional a la seva freqüència. A la taula 5, es mostren els resultats que ha obtingut cada un dels models generats a partir de les mètriques ROC AUC i accuracy.

TAULA 5

RESULTATS DELS MODELS DE SUPPORT VECTOR MACHINE PER CLASSIFICACIÓ

Dades	ROC AUC	Accuracy
FA	0,815	0,693
ML estàndard	0,898	0,888
ML amb lllindar	0,849	0,839
ML selecció d'arestes	0,997	0,965
FA selecció d'arestes	0,967	0,923
GM selecció d'arestes	0,895	0,560
RS selecció d'arestes	0,944	0,951

8.2 Random Forest

El següent mètode que s'ha implementat ha sigut el Random Forest. En aquest cas, es tracta d'un algoritme al qual no li afecta el treballar amb atributs en rangs diferents, per tant s'han mantingut els valors originals de les mètriques de graf. Al igual que abans, s'han generat els plecs pel cross validation amb la mateixa proporció de cada classe i s'han ajustat els pesos de cada classe segons la seva freqüència, per tal de gestionar el desequilibri de classes. Els resultats obtinguts pels models de Random Forest es mostren a la taula 6.

TAULA 6

RESULTATS DELS MODELS DE RANDOM FOREST

Dades	ROC AUC	Accuracy
FA	0,735	0,826
ML estàndard	0,851	0,874
ML amb lllindar	0,800	0,874
ML selecció d'arestes	0,955	0,902
FA selecció d'arestes	0,892	0,895
GM selecció d'arestes	0,893	0,875
RS selecció d'arestes	0,921	0,930

8.3 Multi-layer Perceptron

L'últim mètode de classificació amb el que s'ha treballat ha sigut el Multi-layer Perceptron. Aquí, al igual que en el cas del SVM, el model funciona millor quan els atributs es troben en un rang similar, per tant, s'han escalat totes les

mètriques de graf al rang [0,1]. Per gestionar el desequilibri de classes també s'han generat els plecels pel cross validation amb la mateixa proporció de classes en cada cas. Els resultats que s'han obtingut amb els models de MLP es mostren a la taula 7.

TAULA 7
RESULTATS DELS MODELS DE MULTI-LAYER PERCEPTRON

Dades	ROC AUC	Accuracy
FA	0,692	0,126
ML estàndard	0,857	0,860
ML amb llindar	0,837	0,853
ML selecció d'arestes	0,992	0,944
FA selecció d'arestes	0,924	0,874
GM selecció d'arestes	0,797	0,819
RS selecció d'arestes	0,951	0,960

8.4 Support Vector Machine (Regressió)

Per generar models de regressió que resolguin el problema de la predicció del valor de l'EDSS s'ha decidit fer servir SVM, ja que han demostrat pel problema de classificació que funcionen bé amb les dades que tenim. Com en el cas de classificació, s'han escalat les mètriques de graf al rang [0,1]. Els resultats obtinguts es poden veure a la taula 8. En aquest cas la mètrica que es mostra és l'MSE.

TAULA 8
RESULTATS DELS MODELS DE SUPPORT VECTOR MACHINE PER REGRESSIÓ

Dades	MSE
FA	1,327
ML estàndard	1,243
ML amb llindar	1,377
ML selecció d'arestes	1,116
FA selecció d'arestes	1,245
GM selecció d'arestes	1,388
RS selecció d'arestes	1,298

8.5 Anàlisi dels resultats

Comparant els resultats obtinguts pels models de classificació, podem veure com generalment els models que millor han funcionat han sigut els SVM. En el cas del Multi-layer Perceptron, això sí, també s'ha obtingut algun model amb un molt bon rendiment. Si ens fixem en les xarxes, amb la que millors models s'han generat ha sigut la multilayer amb selecció d'arestes, seguit també per la single-layer FA amb selecció d'arestes, cosa que podíem esperar. En canvi, més sorprenents han sigut els grans resultats obtinguts a partir de la single-layer RS amb selecció d'arestes, ja que segons els anàlisis previs la informació proporcionada per aquestes xarxes no semblava que fos tan rellevant. Si fem servir la mètrica ROC AUC com a criteri, el millor model que hem obtingut ha estat el SVM de la multilayer amb selecció d'arestes, amb un valor de 0,997, seguit pel Multi-layer Perceptron de la multilayer amb selecció d'arestes, amb 0,992, i pel SVM de la single-layer FA amb selecció d'arestes, amb 0,967. Podem veure que la estratègia de seleccionar només les arestes més rellevants ha

resultat molt efectiva a l'hora de classificar els subjectes amb la màxima precisió possible. També veiem que l'ús de xarxes multilayer combinant diferents tipus d'informació ha sigut més efectiu per aquesta tasca que l'ús de xarxes single-layer.

A continuació, es mostren dues figures per visualitzar el rendiment del millor model de classificació. A la figura 9 podem veure una matriu de confusió a la qual es comparen les classes predites pel model amb les classes reals, mentre que a la figura 10 es pot veure la corba ROC de la classe PwMS:

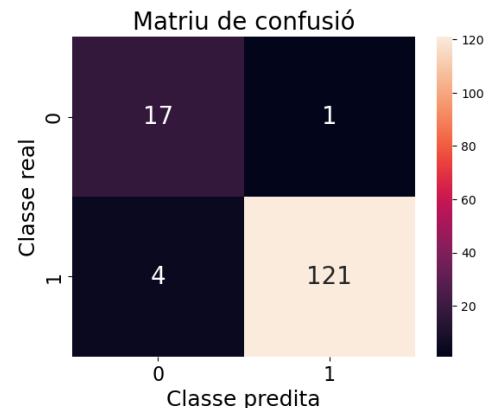


Fig. 9. Matriu de confusió amb les classificacions fetes pel model de SVM de la multilayer amb selecció d'arestes (classe 1 = PwMS; classe 0 = HV).

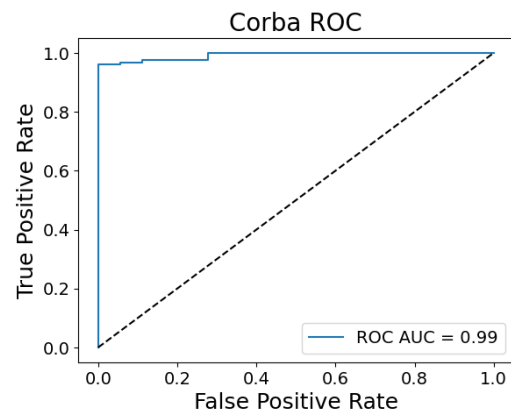


Fig. 10. Corba ROC de la classe PwMS del model de SVM de la multilayer amb selecció d'arestes.

En quant als models de regressió generats, la xarxa que millors resultats ha obtingut, segons la mètrica MSE, ha sigut la multilayer amb selecció d'arestes, amb 1,116, seguit per la multilayer estàndard, amb 1,243, i per la single-layer FA amb selecció d'arestes, amb 1,245. En aquest cas, la tècnica de la selecció d'arestes no ha fet tant efecte com als models de classificació, el que més ha influït és tenir una estructura multilayer en comptes d'una single-layer.

A la figura 11 podem veure una gràfica de dispersió on cada punt representa un subjecte concret. Els subjectes estan distribuïts al llarg de l'eix horitzontal segons el seu valor real de l'EDSS i al llarg de l'eix vertical segons el valor de l'EDSS que ha predit el millor model de regressió obtingut. S'observa una diferència entre els valors predits pels

subjectes que tenen EDSS=0, que són els de la classe HV, i la resta. En canvi, es veu com a l'hora de diferenciar entre els subjectes de la classe PwMS segons el valor de l'EDSS, el model té més dificultats.

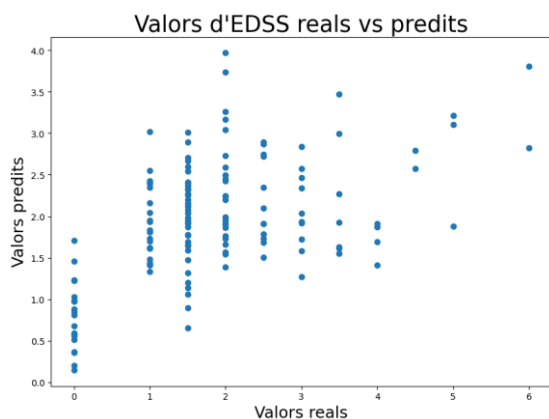


Fig. 11. Gràfica de dispersió comparant els valors reals d'EDSS amb els valors predits pel model de SVM de la multilayer amb selecció d'arestes.

9 CONCLUSIONS

En aquest treball s'ha pogut aprofundir en la feina que s'havia estat fent últimament d'explorar la capacitat de les dades sobre la connectivitat cerebral per identificar persones amb Esclerosi Múltiple. S'han realitzat noves aportacions, com per exemple combinar diferents estratègies de creació de les xarxes, tant multilayer com single-layer, tenir en compte només unes arestes concretes a l'hora de crear-les per eliminar així la informació no rellevant, provar algunes mètriques de graf noves o implementar nous models d'aprenentatge automàtic a partir de les mètriques obtingudes. Alguns d'aquests models implementats han aconseguit distingir HV de PwMS amb bastant èxit, de forma que se'n poden treure dues conclusions: La primera, que l'ús d'estructures multilayer dona més informació que l'ús d'un sol tipus de xarxa individualment i ajuda a millorar el rendiment dels models generats. La segona, que seleccionant únicament les arestes més rellevants i eliminant d'aquesta forma el soroll provocat per la resta de dades es pot millorar molt el resultat de la classificació.

Tot i així, els resultats obtinguts s'han de considerar en el context de que el nombre de subjectes de l'estudi no és gaire alt, per tant, seria interessant poder confirmar en un futur el funcionament d'aquestes estratègies afegint noves mostres al procés, especialment de més voluntaris sans. Aquest treball també podria tenir algunes línies de continuació en el futur per seguir explorant, per exemple, encara es podrien provar noves mètriques de graf, i si resultessin rellevants, afegir-les com a atributs a la creació de models. Una altra millora a provar seria definir un nou mètode per calcular les mètriques de graf per les xarxes multilayer que no sigui amb la matriu de distàncies mínimes o la de pesos màxims, i comprovar si es milloren els resultats.

AGRAÏMENTS

M'agradaria donar les gràcies al meu tutor, el Jordi Casas, per haver-me guiat durant la realització del treball per tal que el pogués completar amb èxit.

BIBLIOGRAFIA

- [1] Vázquez-Marrufo, M., Sarrias-Arrabal, E., García-Torres, M., Martín-Clemente, R., & Izquierdo, G. (2022). A systematic review of the application of machine-learning algorithms in multiple sclerosis. *Neurologia* (English Edition).
- [2] Casas-Roma, J., Martínez-Heras, E., Solé-Ribalta, A., Solana, E., López-Soley, E., Vivó, F., ... & Prados, F. (2022). Applying multi-layer analysis to morphological, structural, and functional brain networks to identify relevant dysfunction patterns. *Network Neuroscience*, 6(3), 916-933.
- [3] Hotz, N. (2023, January 19). What is CRISP DM? - Data Science Process Alliance. Data Science Process Alliance. <https://www.datascience-pm.com/crisp-dm-2/>
- [4] Martins, J. (2024, January 19). ¿Qué es la metodología Kanban y cómo funciona? [2024] • Asana. Asana. <https://asana.com/es/resources/what-is-kanban>
- [5] Solana, E., Martínez-Heras, E., Casas-Roma, J., Calvet, L., López-Soley, E., Sepulveda, M., ... & Llufríu, S. (2019). Modified connectivity of vulnerable brain nodes in multiple sclerosis, their impact on cognition and their discriminative value. *Scientific reports*, 9(1), 20172.
- [6] Zurita, M., Montalba, C., Labbé, T., Cruz, J. P., da Rocha, J. D., Tejos, C., ... & Uribe, S. (2018). Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data. *NeuroImage: Clinical*, 20, 724-730.
- [7] Shoeibi, A., Khodatars, M., Jafari, M., Moridian, P., Rezaei, M., Alizadehsani, R., ... & Acharya, U. R. (2021). Applications of deep learning techniques for automated multiple sclerosis detection using magnetic resonance imaging: A review. *Computers in Biology and Medicine*, 136, 104697.
- [8] Alves P, Green E, Leavy M, et al. Validation of a machine learning approach to estimate expanded disability status scale scores for multiple sclerosis. *Multiple Sclerosis Journal - Experimental, Translational and Clinical*. 2022;8(2). doi:10.1177/20552173221108635
- [9] Roca, P., Attye, A., Colas, L., Tucholka, A., Rubini, P., Cackowski, S., ... & OFSEP Investigators. (2020). Artificial intelligence to predict clinical disability in patients with multiple sclerosis using FLAIR MRI. *Diagnostic and Interventional Imaging*, 101(12), 795-802.
- [10] Wikipedia contributors. (2024, February 10). Student's t-test. Wikipedia. https://en.wikipedia.org/wiki/Student%27s_t-test#Independent_two-sample_t-test
- [11] Software for Complex Networks — NetworkX 3.3 documentation. (n.d.). <https://networkx.org/documentation/stable/>
- [12] StratifiedKFold. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

APÈNDIX

A1. DIAGRAMA DE GANTT

A continuació, es mostra el diagrama de Gantt que conté les fases definides durant la planificació del projecte.

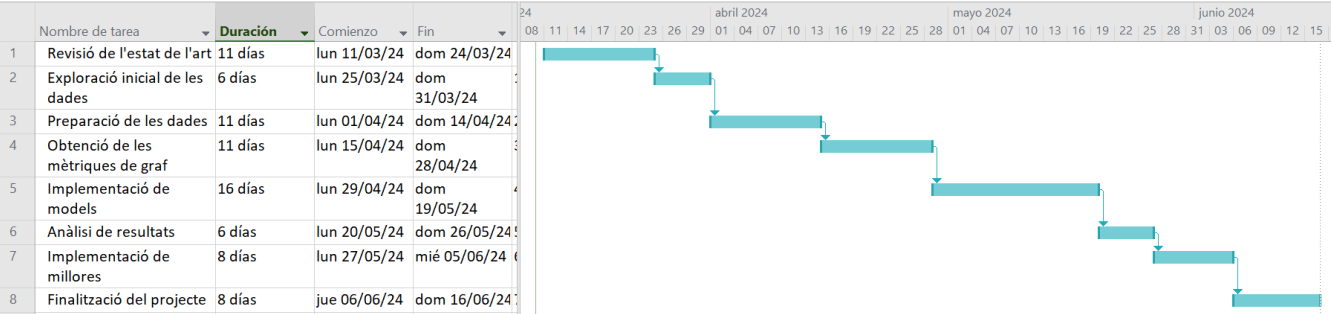


Fig. A1: Diagrama de Gantt amb les fases del treball

A2. HISTOGRAMES DE LES XARXES

En aquest apartat podem observar tres figures, corresponents als histogrames dels pesos de totes les arestes per cadascun dels tres tipus de xarxes, considerant les matrius de tots els subjectes.

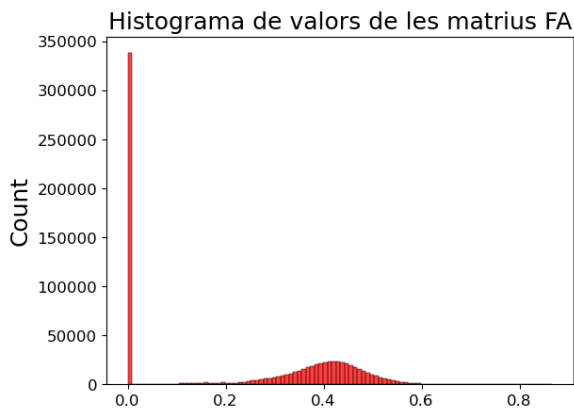


Fig. A2.1: Histograma de valors de les matrius FA

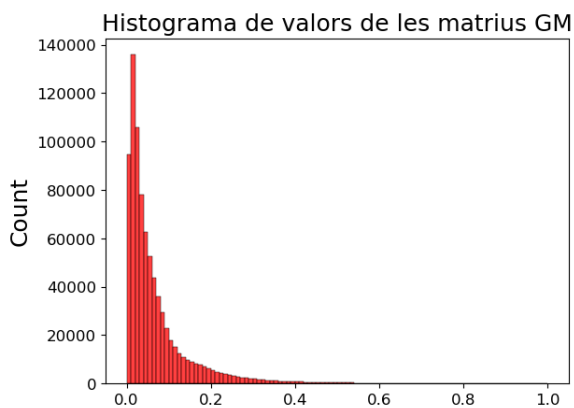


Fig. A2.2: Histograma de valors de les matrius GM

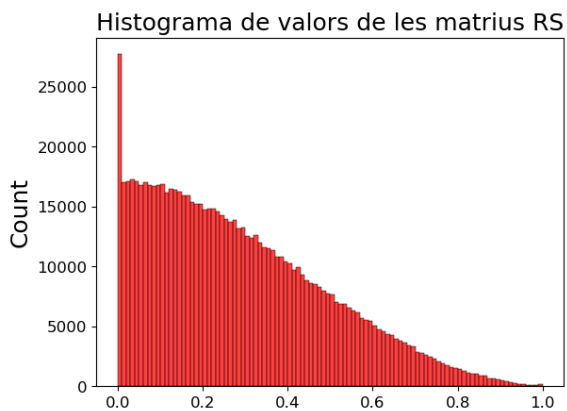


Fig. A2.3: Histograma de valors de les matrius RS