Predicció de Matrícula Basada en Aprenentatge Automàtic

Anass Anhari

Enginyeria de Sistemes TIC
Universitat Politècnica de Catalunya
http://epsem.upc.edu

20 de juliol de 2023

Introducció

- A la Universitat Politècnica de Catalunya (UPC) una tasca administrativa important és la planificació del següent curs acadèmic.
- La previsió d'estudiants que es matricularan a cada assignatura és important.
- El nombre de grups, determina la necessitat de professorat per a cada assignatura i la necessitat d'espais suposant un impacte pressupostari important.

Introducció

- Actualment, aquest procés es fa en base a l'experiència personal i esforç dels gestors (en mans de poques persones).
- El procediment actual es destaquen els següents problemes:
  - És un procediment sensible a la disponibilitat de les persones que tenen el coneixement. Transferir l'experiència i el «bon saber» a tercers és difícil.
  - El procediment s'ha d'executar en uns terminis estrets de temps.
  - Són freqüents les prediccions poc acurades. Si subestimen la matrícula la capacitat de les assignatures es satura. Si se sobreestima la matrícula, es malbaraten recursos cars.

Objectius

L'objectiu d'aquest treball és aplicar l'ús de tècniques de aprenentatge automàtic (Machine Learning) per automatizar la predicció de la matrícula universitària i determinar-ne la viabilitat. Per assolir aquest objectiu es plantegen els següents objectius específics:

- Entendre en detall els algoritmes d'aprenentatge automàtic que s'aplicaran en aquest treball.
- Obtindre la màxima precisió possible en la predicció de matriculacions per a cada assignatura.
- Construir una sistema d'entrenament i processat de dades escalable, eficient, i fàcil de mantenir.
- Avaluar i comparar diferents models i tècniques d'aprenentatge automàtic per identificar el més adequat per aquesta aplicació.

Context

- Les dades disponibles per a aquest treball són els expedients dels estudiants de TIC, anonimitzats per garantir la confidencialitat.
- Aquest conjunt de dades proporciona un històric acadèmic dels estudiants en les diferents assignatures del pla d'estudis.

estudianth	assignatura	acrònim	curs	quad	nota
19h3j	330212	MBE	2010	1	7.4
19h3j	330213	F	2010	1	6.2
19h3j	330214	I	2010	1	10.0
19h3j	330215	ISD	2010	1	9.0
	...				
2837k	330212	PCTR	2020	2	9.2
2837k	330213	SO	2020	2	7.9
2837k	330214	XC	2020	2	9
2837k	330215	PDS	2020	2	7

Machine Learning

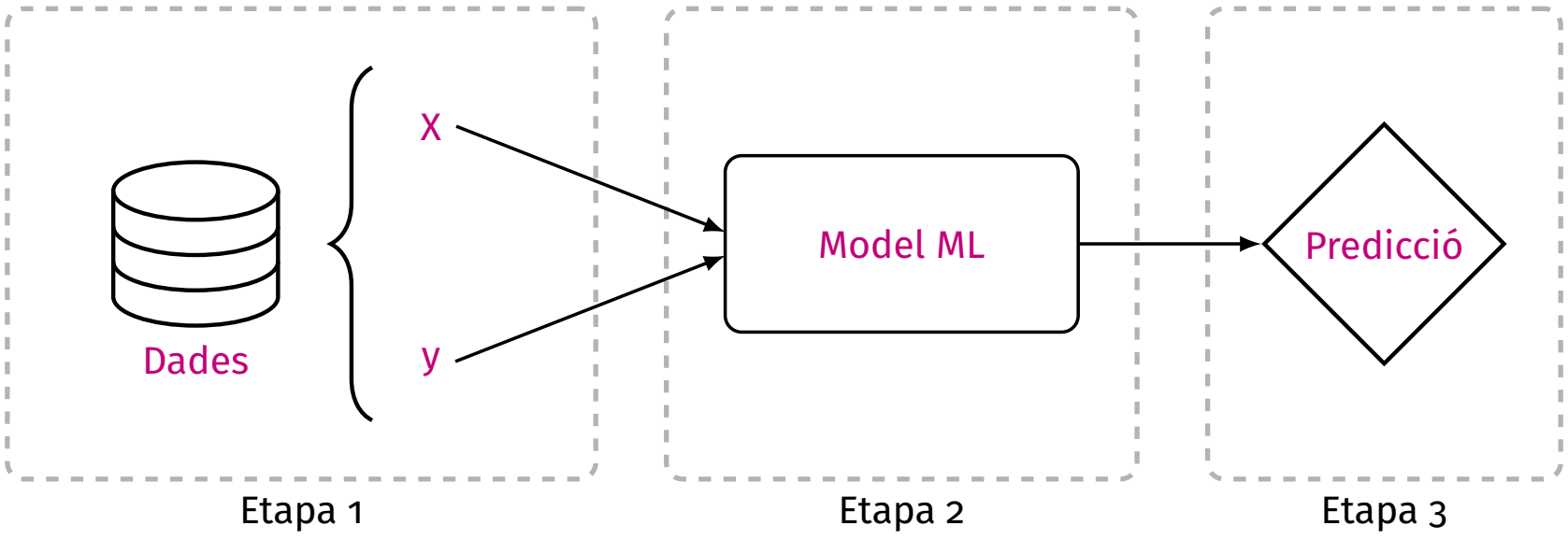
Machine Learning (aprenentatge automàtic), és la ciència de la programació que presenta i defineix una col·lecció d'algorismes perquè un computador pugui aprendre de les dades.

- *Label*. L'atribut o el conjunt d'atributs que representa cada mostra.
- *Feature*. La solució o el conjunt de solucions per a cada mostra.

Sample	*Feature 1*	...	*Feature n*	*Label 1*	...	*Label n*
Sample 1		...			...	
⋮						
Sample n		...			...	

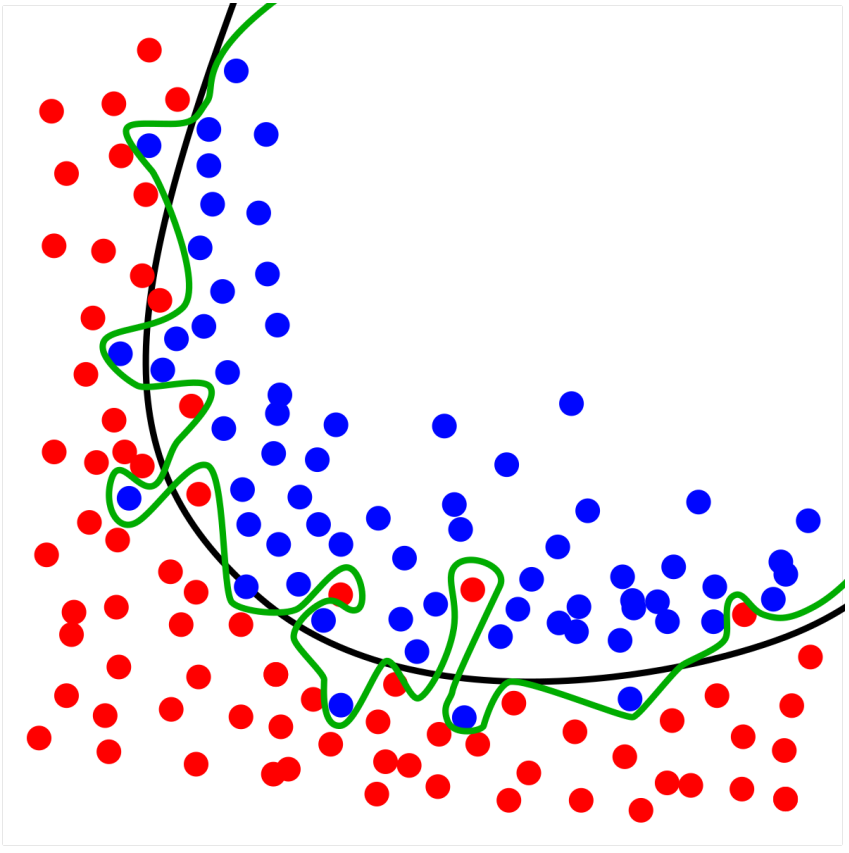
Taula: Estructura habitual d'un *dataset* (conjunt de dades)

- *Problemes de classificació*. En aquests casos, les mostres es poden classificar en dues o més classes. L'algorisme del model aprèn de les dades i dels atributs de cada mostra per determinar la solució.
- *Problemes de regressió*. En aquest tipus de problemes, l'objectiu és predir sortides contínues, és a dir, un valor o un conjunt de valors numèrics.



Machine Learning: Overfitting

L'overfitting és un problema comú en la construcció de models d'aprenentatge automàtic. Es produeix quan el model s'ajusta massa als detalls i al soroll de les dades d'entrenament, perdent així la seva capacitat de generalització i predir amb precisió dades noves o de test.



Estructura de les dades

En l'àmbit de l'aprenentatge automàtic, les dades són un factor crucial. En la majoria de casos, les dades no sempre són consistents. Sovint, les dades pateixen d'anomalies, errors o inconsistències que afecten la qualitat dels resultats:

- **Acrònims repetits o que falten**.
- **Convalidacions**.
- **Estudiants que deixen la carrera**.
- **Notes buides**:
  - Nota 0?
  - Assignatura no presentada?
  - Assignatura convalidada?
  - L'estudiant ha abandonat la carrera?

Estructura de les dades

Dins de les dades dels expedients acadèmics, es distingeixen dues categories principals:

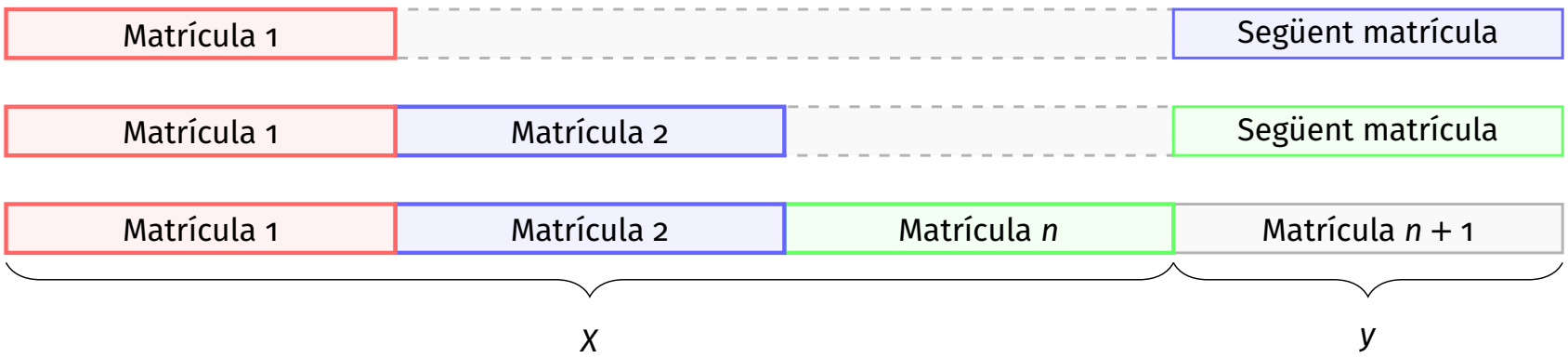
- **Les dades de «l'històric acadèmic»**.
- **Les dades de la «motxilla de l'estudiant»**.

Les dades dels expedients acadèmics són proporcionades en un fitxer *CSV* i conté els següents registres:

- codi_expedient
- curs
- quad
- codi_upc_ud
- nota_num_def
- nota_des_def
- beca
- any_naix
- via_acces
- ordre_assignacio
- nota_acces

Transformació de les dades (1)

$$e_i = \Big\|_{i=1}^{N_{mat}} \left(\Big\|_{x=1}^{N_{assig}} \Big(mat(x, e_i) \,\Big\|\, nota(x, e_i, 0) \Big) \,\Big\|\, \left(\Big\|_{x=1}^{N_{assig}} mat(x, e_i + 1) \right) \right)_{[1368x900]}$$



expid	curs	quad	1:MBE.n	1:MBE.m	1:F.n	1:F.m	1:I.n	1:I.m	...
19h3j	2010	1	7.4	True	6.2	True	10.0	True	
19h3j	2010	2	7.4	True	6.2	True	10.0	True	...
	⋮								
2837k	2020	2	9.2	True	7.9	True	9	True	...
2837k	2021	1	9.2	True	7.9	True	9	True	...

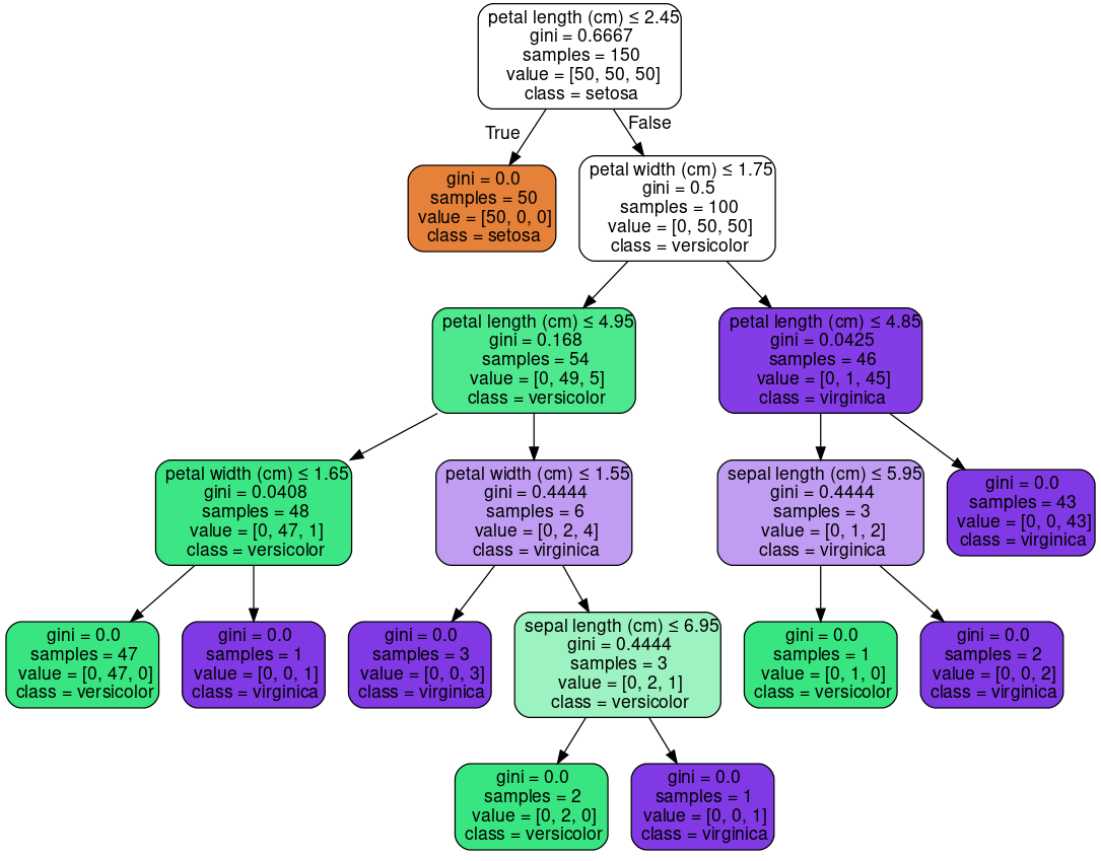
Transformació de les dades (2)

$$e_i = \Big\|_{i=1}^{n-1} \left(\Big\|_{x=1}^{N_{assig}} \left(a'_r \,\middle\|\, mat(x, e_i) \,\middle\|\, nota(x, e_i, 0) \right) \middle\| \left(\Big\|_{x=1}^{N_{assig}} mat(x, e_i + 1) \right) \right)_{[1368,135]}$$

expid	MBE.da	MBE.n	MBE.m	F.da	F.n	F.m	...
est_1	1	4	True	1	True	10.0	
est_1	1	4	False	1	False	10.0	
est_1	2	7.5	True	1	False	10.0	
$\vdots$							

Arbres de decisió

- Els arbres de decisió són un tipus d'algorismes per classificació i regressió.
- Els arbres de decisió són particularment atractius en l'àmbit de la intel·ligència artificial explicable (*Explainable AI*).



Algorisme CART (Classification And Regression Tree)

- Vectors d'entrenament: $x_i \in R^n$ (*features*)
- Vectors de classificació: $y \in R^l$ (*labels*)
- Q_m és el conjunt dades al node m amb n_m mostres.
- Per cada candidat es particiona $\theta = (f, t_m)$ amb un *feature* f i un llindar (*threshold*) t_m.

$$Q_m^{esquerra}(\theta) = (x, y)|x_f \leq t_m$$
$$Q_m^{dreta}(\theta) = Q_m \setminus Q_m^{esquerra}(\theta)$$

El candidat de la partició del node m és seleccionat en base a una funció de pèrdua $H()$ (o impuritat):

$$G(Q_m, \theta) = \frac{n_m^{esquerra}}{n_m} H(Q_m^{esquerra}(\theta)) + \frac{n_m^{dreta}}{n_m} H(Q_m^{dreta}(\theta))$$

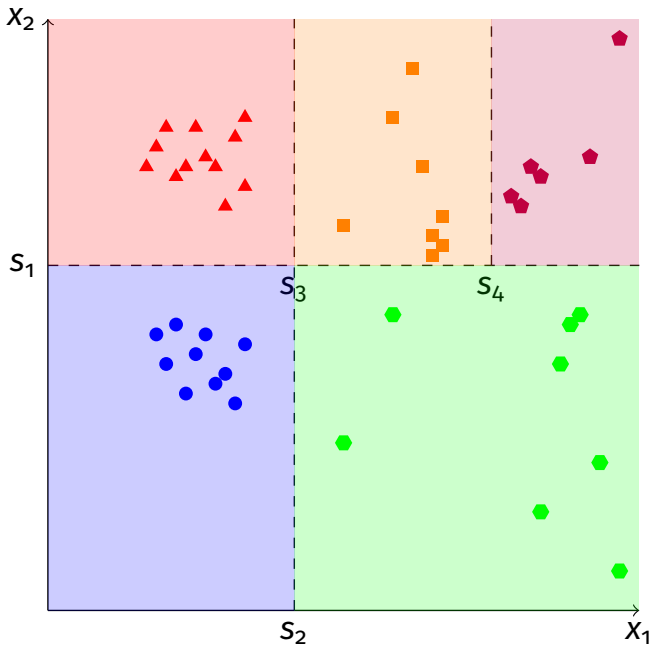
Algorisme CART (Classification And Regression Tree)

Existeixen vàries funcions d'impuritat $H()$ en problemes de classificació, en el nostre cas, aplicarem *Gini Index*:

$$H(Q_m) = \sum_{k=1} p_{mk}(1 - p_{mk})$$

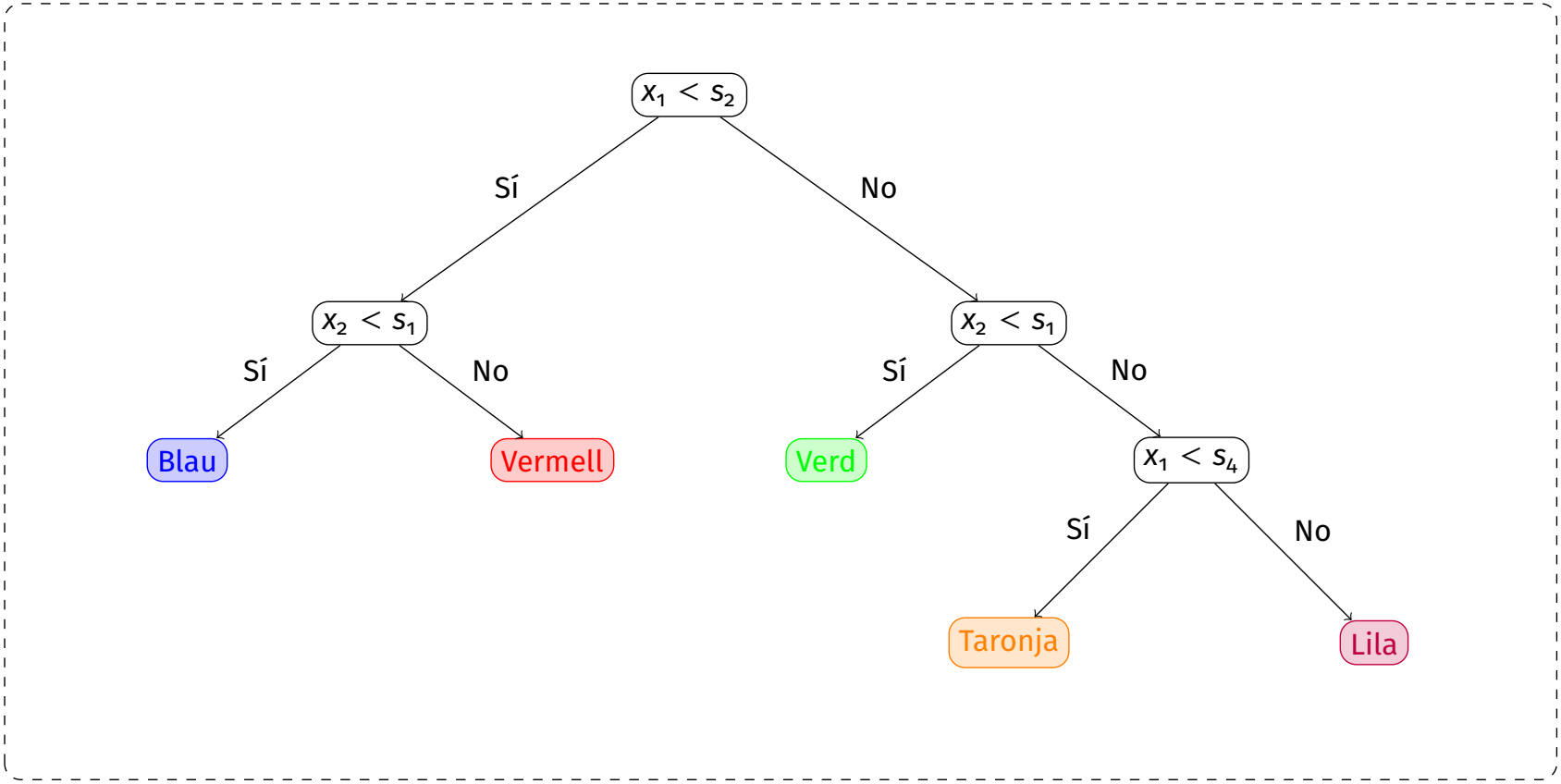
on, p_{mk} defineix la proporció de classes k que sobserven a la partició al node m.

Algorisme CART (Classification And Regression Tree)




```
if x1 < s2:
  if x2 < s1: Classe: Blau
  else:       Classe: Vermell
else:
  if x2 < s1: Classe: Verd
  else:
    if x1 < s4: Classe: Taronja
    else:       Classe: Lila
```

Algorisme CART (Classification And Regression Tree)



Esporgat d'arbres *(Pruning)*

Per minimitzar l'overfitting, s'utilitza principalment Minimal Cost-Complexity Pruning (MCCP), és un algorisme parametritzat per $\alpha \geq 0$, aquest, definirà la mesura de complexitat $R_\alpha(T)$ del arbre (T):

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

- $\tilde{T}$: El nombre de nodes terminals
- T: Un subarbre
- $R(T)$: La taxa total de classificació errònia. En Scikit-learn, es calcula com la mitjana poderada total de la impuritat dels nodes terminals. Normalment, la impuritat d'un node és major a la suma de la impuritat dels seus nodes terminals $R(T_t) < R(t)$
- T_t: La branca T_t defineix l'arbre on el node t és l'arrel principal.

MCCP trobarà un subarbre de T que minimitzi $R_\alpha(T)$:

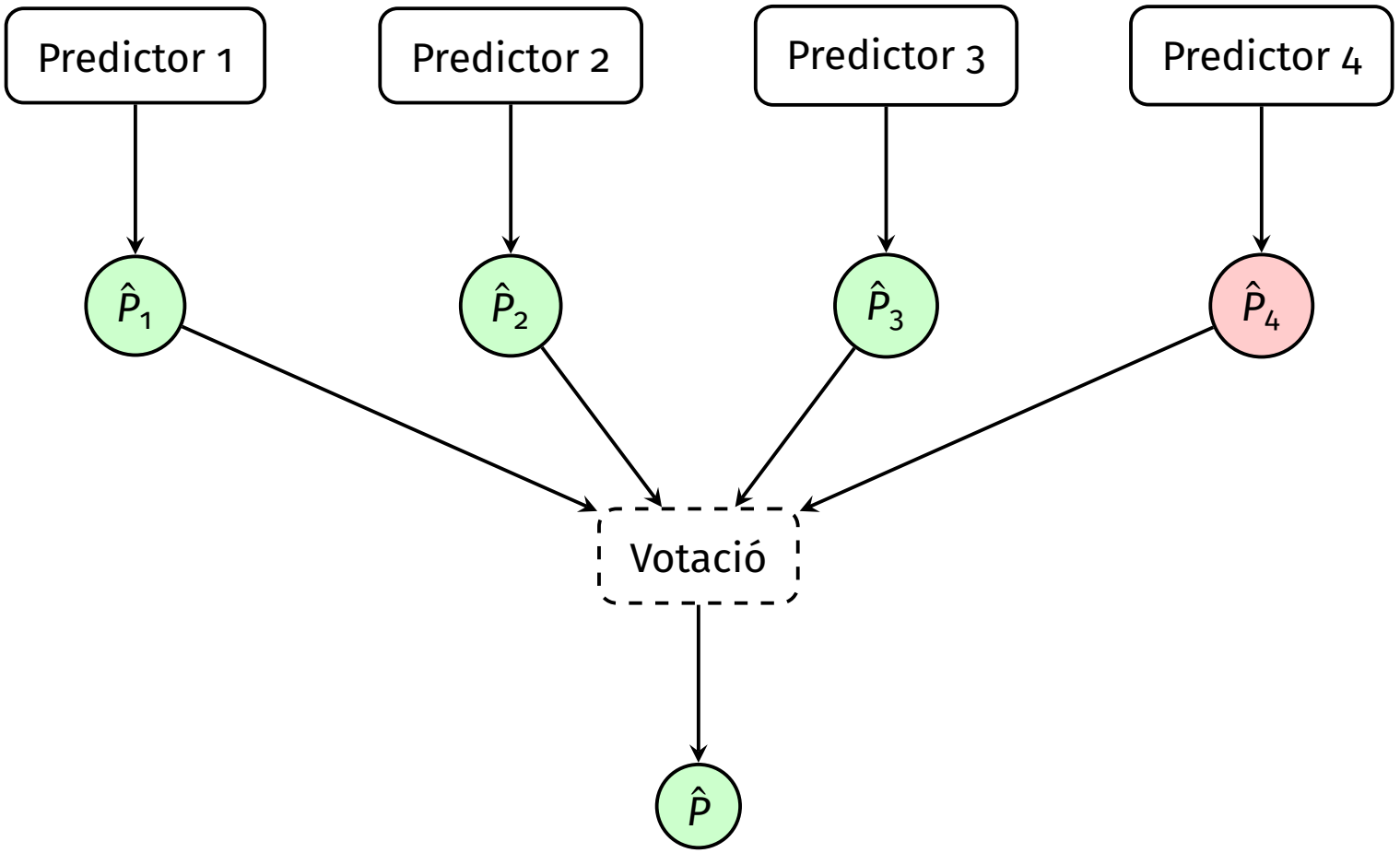
$$\alpha_{eff}(t) = \frac{R(t) - R(T_t)}{|\tilde{T}| - 1}$$

Esporgat d'arbres *(Pruning)*

- Finalment, es retalla el node no terminal amb el valor més petit de α_{eff}.
- El procès finalitzarà quan α_{eff} sigui major al paràmetre de complexitat α.

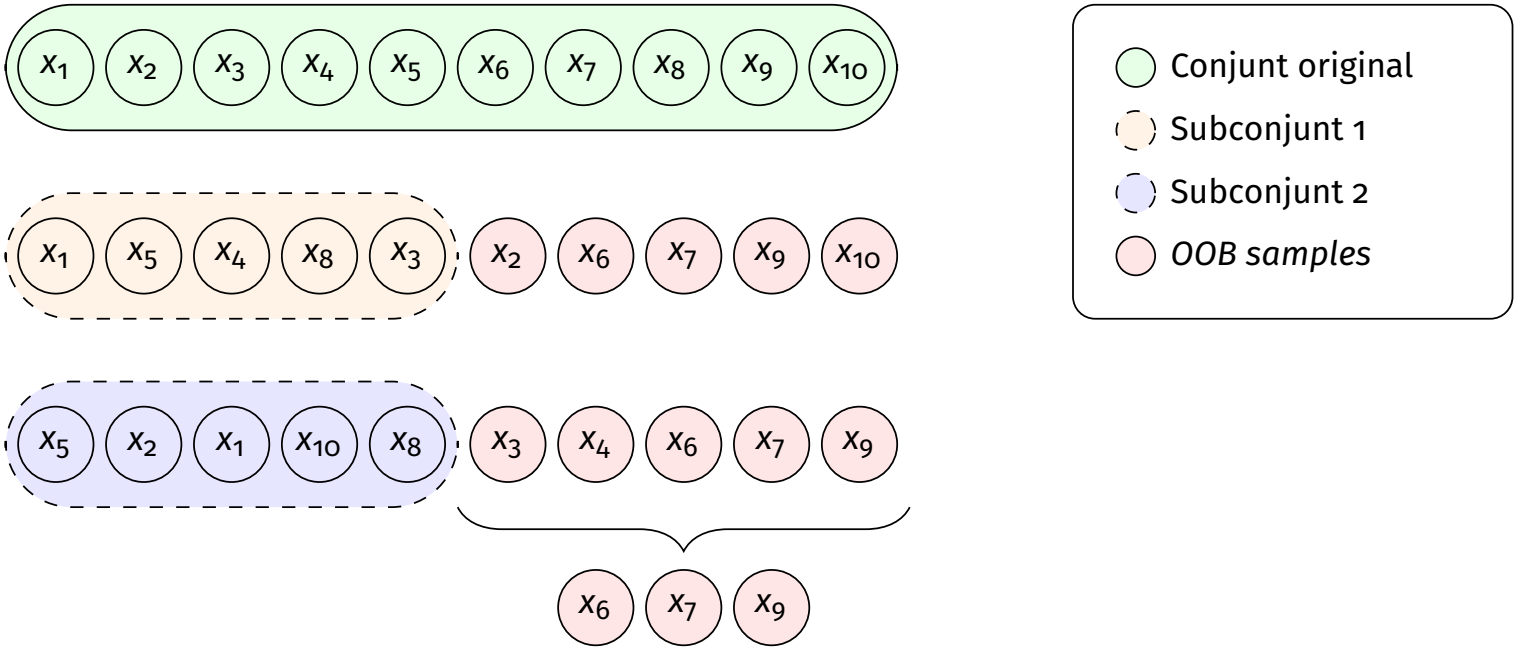
Boscos d'arbres aleatoris

- Existeix una altra metodologia d'aprenentatge coneguda com a Ensemble Methods (mètodes d'aprenentatge en conjunt) que es basa en el principi de la «sabiduria de la multitud» (wisdom of the crowd).



Bootstraping

- Una tècnica comuna és l'aplicació del *Bootstraping* també anomenat *Bagging*. Aquest mètode obté una selecció aleatòria de les mostres (*samples*) de les dades originals.
- Les mostres no seleccionades (*OOB samples*) poden utilitzar-se per avaluar el rendiment del model sense necessitat de disposar d'un conjunt de validació separat.



Bootstraping

Feature 1	*Feature 2*	*Label*
0.5	0	A
1.5	1	B
2.0	2	A
1.0	3	B
2.5	4	A

Bootstraping

Model	*Feature 1*	*Feature 2*	*Label*
Arbre 1	0.5	0	A
	2.0	2	A
	1.5	1	B
Arbre 2	1.5	1	B
	1.0	3	B
	2.0	2	A
Arbre 3	0.5	0	A
	2.5	4	A
	1.0	3	B

Feature 1	*Feature 2*	*Label*	*Predicció*
1.0	3	B	A
2.5	4	A	A

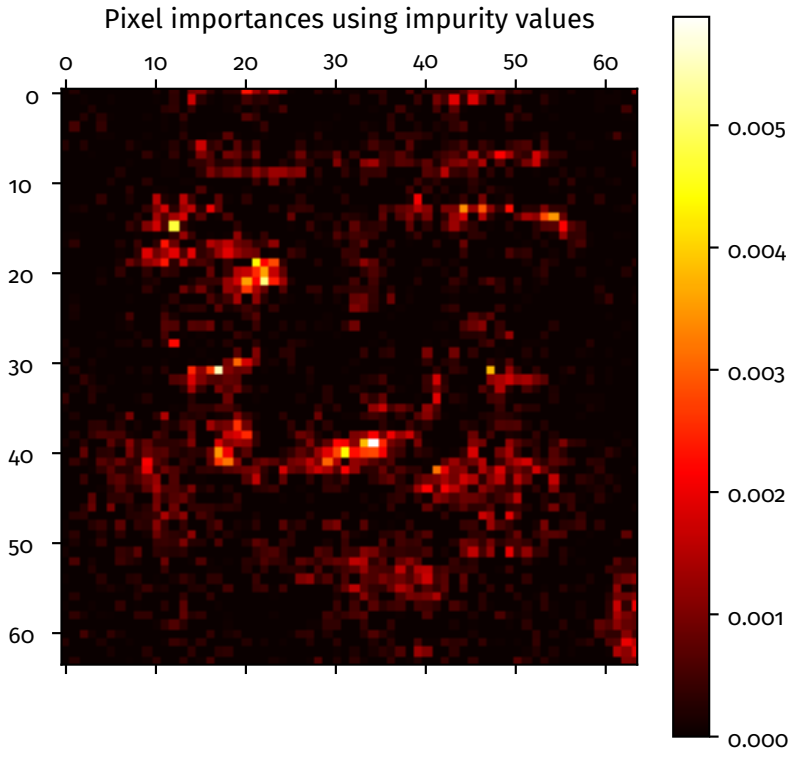
Taula: Predicció basada en les *OOB samples* (Arbre 1)

$$\text{OOB Error}_{\text{Arbre 1}} = \frac{\text{Classificacions errònies}}{\text{Mostres totals}} = \frac{1}{2} = 0.5$$

$$\text{OOB Error (Bosc)} = \frac{\sum_1^n \text{Error}_{\text{Arbre } n}}{\text{Quantitat d'arbres}} = 2$$

Feature Importance

- Un concepte rellevant en els boscos d'arbres aleatoris és el feature importance, que permet determinar quines *features* són més importants i tenen un major impacte en la predicció del model.
- Es diu que una *feature* és important quan el valor de la impuritat en cada partició es veu reduïda significativament.



Mètriques per a models de classificació

	Positiu (predicció)	Negatiu (predicció)
Positiu (actual)	Positiu Real (TP)	Fals Negatiu (FN)
Negatiu (actual)	Fals Positiu (FP)	Negatiu Real (TN)

Les mètriques més comunes son les següents:

- Exactitud (accuracy). És la proporció de prediccions correctes obtingudes pel model sobre tot el conjunt de prediccions.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Cobertura (recall). Mesura la capacitat del model per identificar tots els casos positius. Un valor de recall elevat significa que el model té bona detecció de tots els casos rellevants.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

Mètriques per a models de classificació

- Precisió (precision). Mesura la capacitat del model per identificar correctament els casos positius sense classificar incorrectament casos negatius com a positius.

$$precision = \frac{TP}{TP + FP} \tag{3}$$

- F1-score. És la mitjana harmònica de la precisió i el recall.

$$f1_{score} = 2\,\frac{precision \cdot recall}{precision + recall} \tag{4}$$

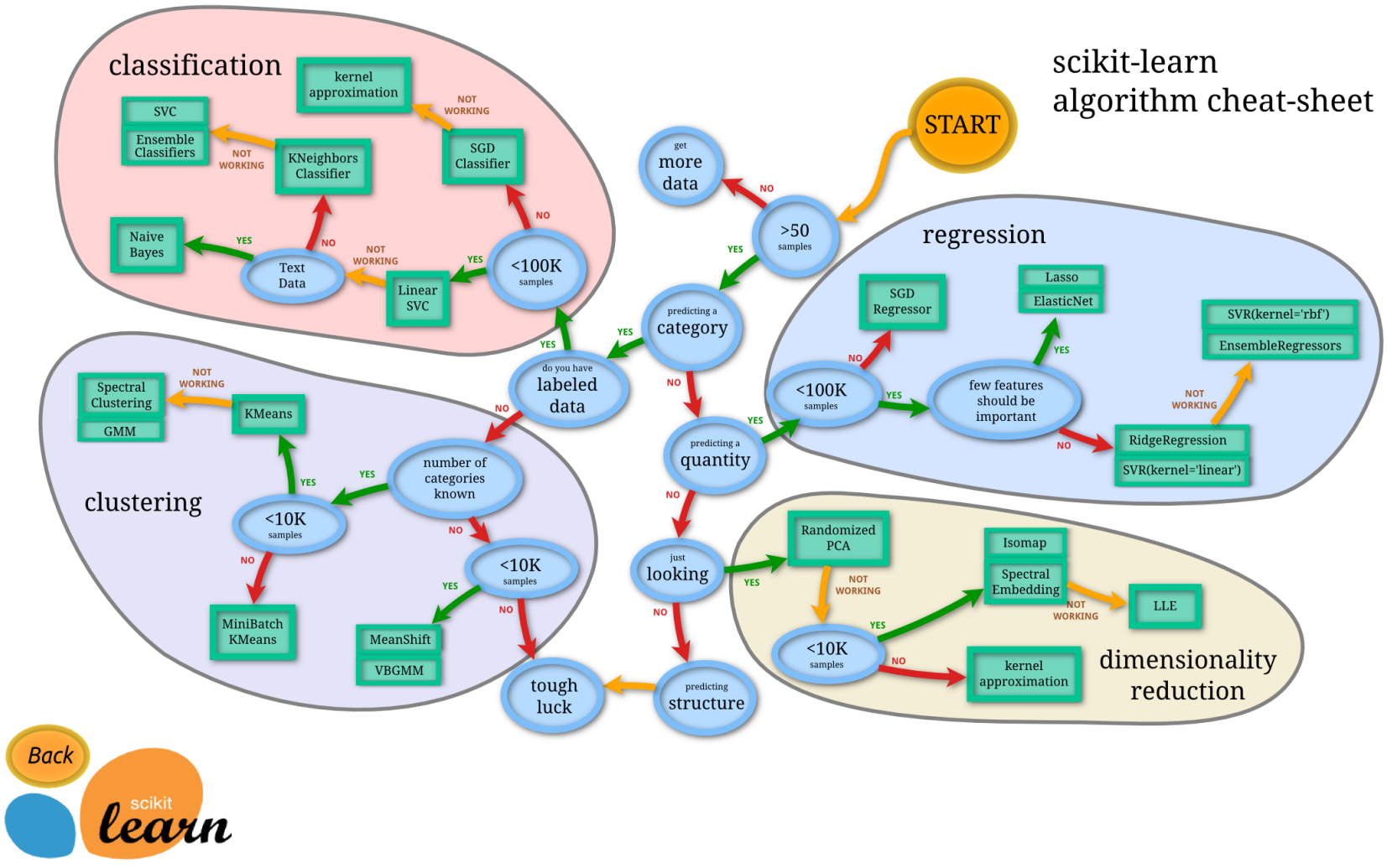
- Segons l'aplicació interessarà maximitzar una mètrica, per exemple:
  - En l'àmbit de la salut és extremadament important evitar falsos negatius (casos positius que són classificats incorrectament com a negatius)
  - En la detecció d'alarmes, és millor que es doni una falsa alarma (fals positiu) que no pas una alarma real no s'arribi a detectar (fals negatiu).

Arquitectura software

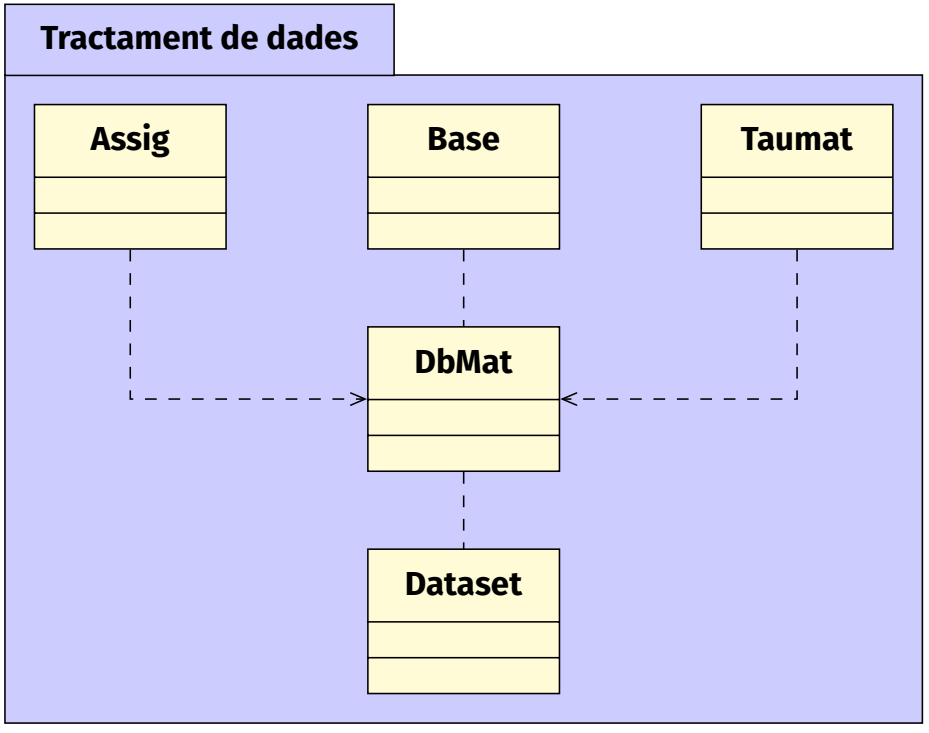
En aquest treball s'ha decidit treball amb les següents eines:

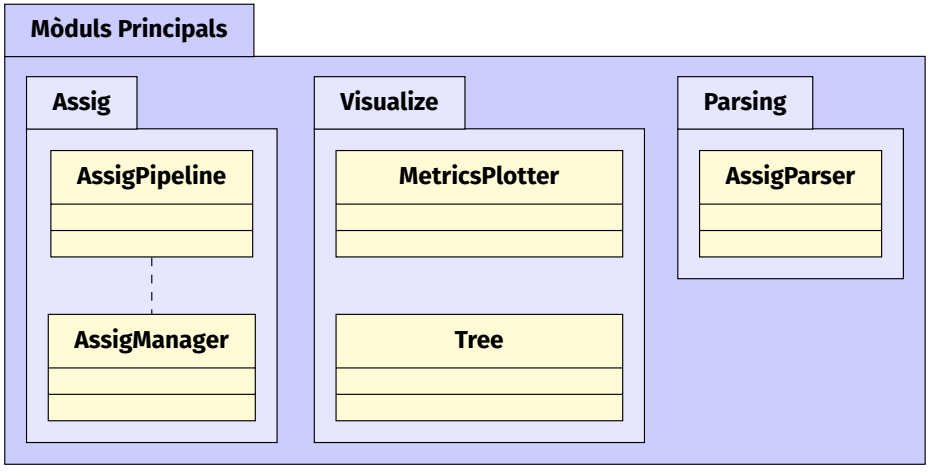
- Python. Un llenguatge de programació molt popular en l'àmbit de la ciència de dades i l'aprenentatge automàtic.
- NumPy. Una extensió de Python que ofereix operacions matemàtiques avançades i manipulació de dades (vectors i matrius) de manera eficient.
- Pandas. És eina extremadament potent, flexible i fàcil d'utilitzar per a l'anàlisi i la manipulació de dades. Proporciona estructures de dades com per exemple DataFrames, que faciliten la gestió i la transformació de conjunts de dades.
- Matplotlib. Una biblioteca especialitzada en la generació de gràfics estàtics o animats a partir de les dades. Amb Matplotlib, és possible crear una gran varietat de visualitzacions, com ara gràfics de línies, barres, dispersió, histograma, entre d'altres.

Arquitectura software



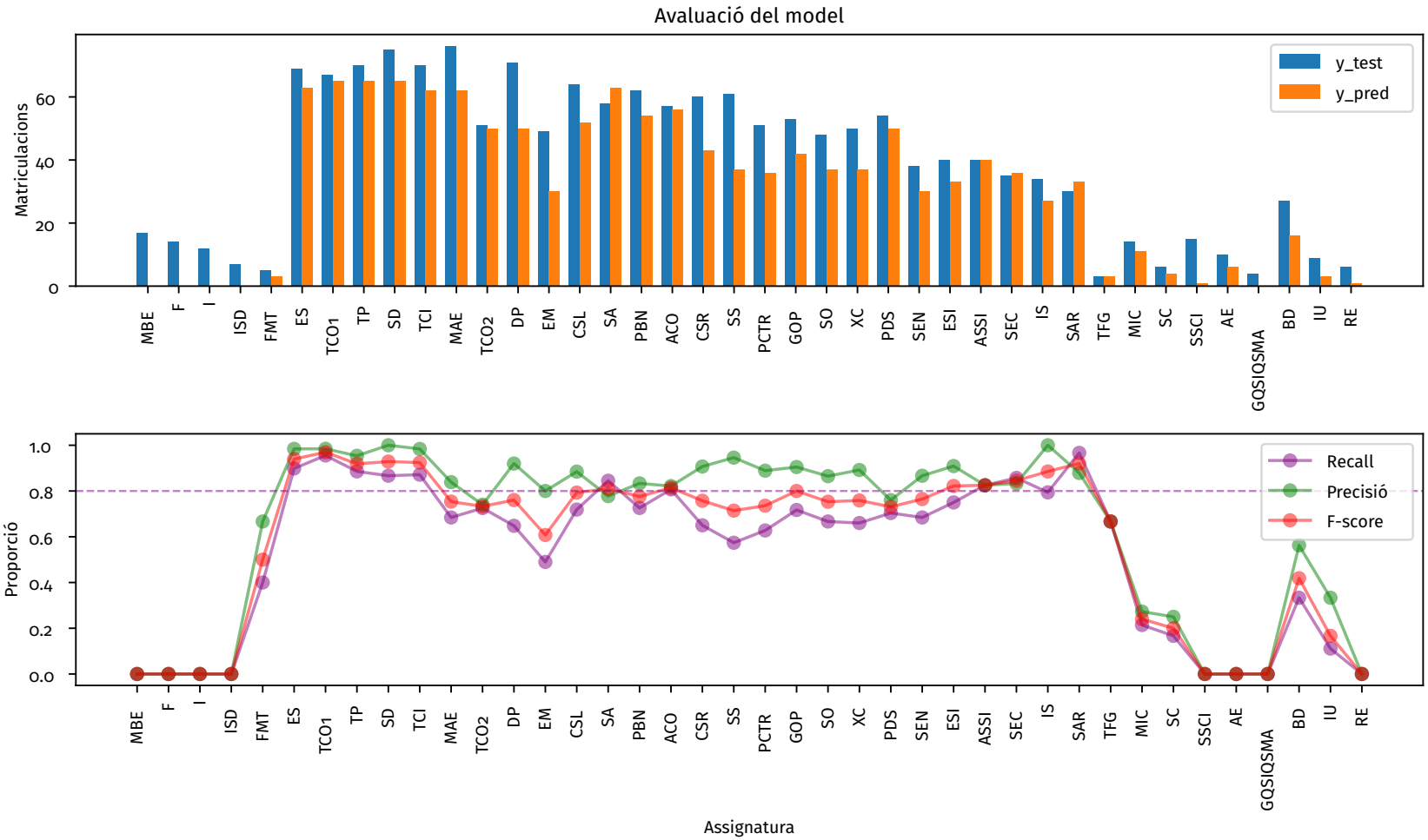
Arquitectura software





Models de predicció de matrícula: Arbres de decisió

Identificador (Id)	*Model*	*Profunditat*	*Transformació*	*Motxilla*
DT3T1	ARBRE	3	1	No
DT4T1	ARBRE	4	1	No
DT5T1	ARBRE	5	1	No



Arbres de decisió

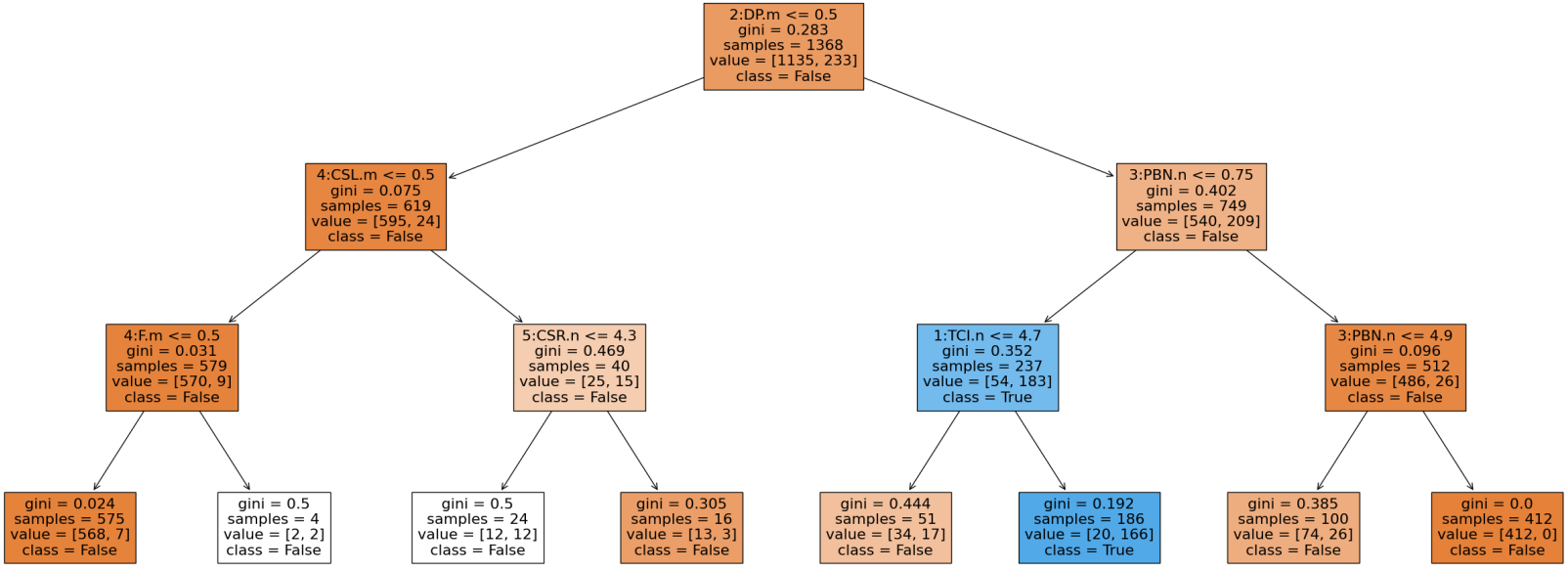


Figura: Arbre de PBN (ID: DT3T1)

- Es matricula si l'estudiant va matricular Dispositius Programables (DP).
- No la matricula si ja té aprobada PBN.

Arbres de decisió

Per matricular-se a Circuits i Sistemes Lineals (CSL), hi ha una gran dependència de l'assignatura de Teoria de Circuits (TCI).

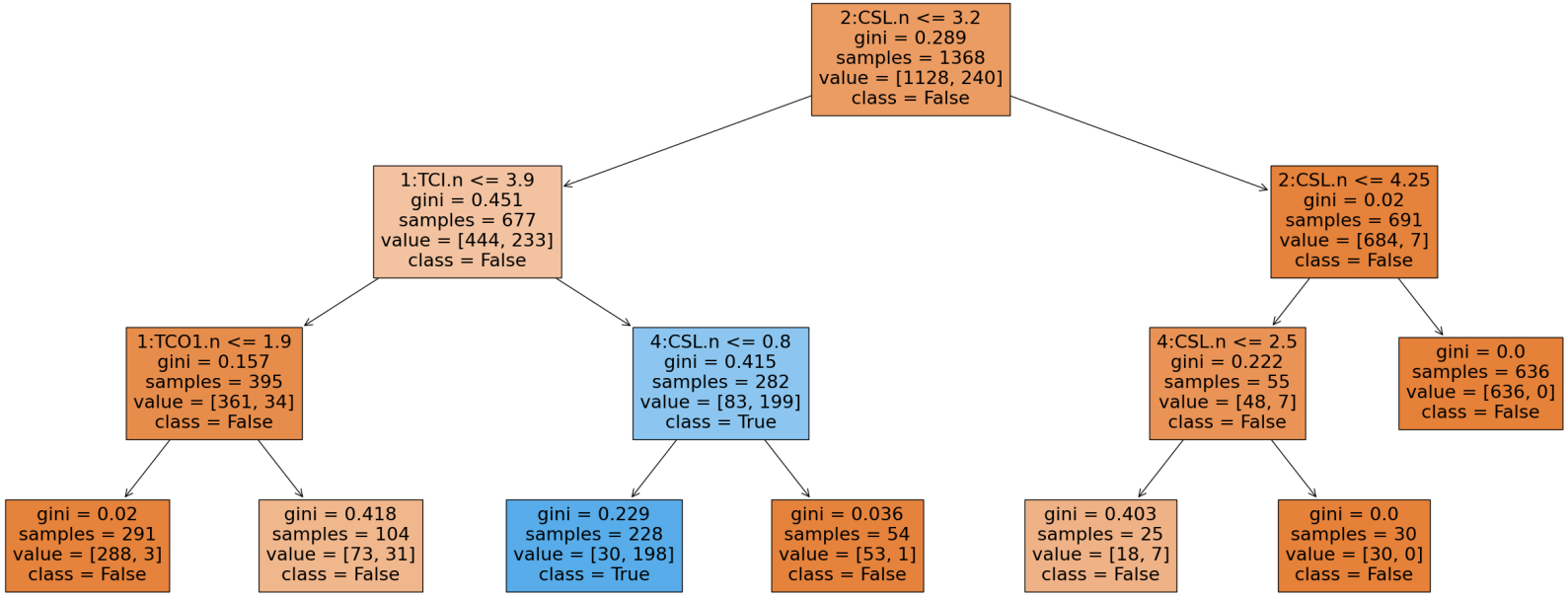
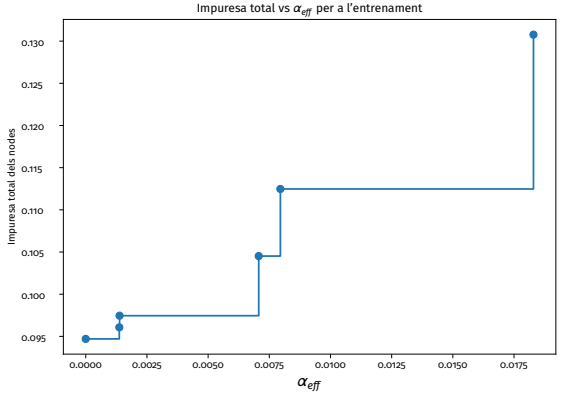
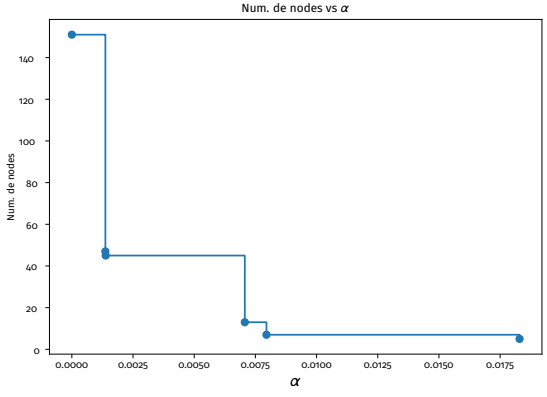
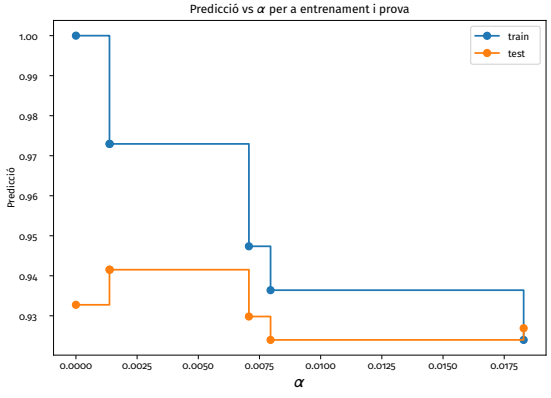


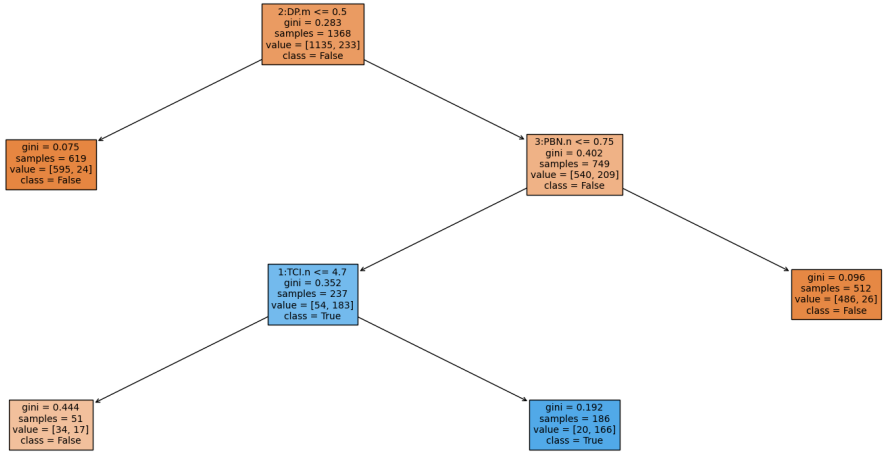
Figura: Arbre de CSL (ID: DT3T1)

Arbres de decisió: Esporgat dels arbres

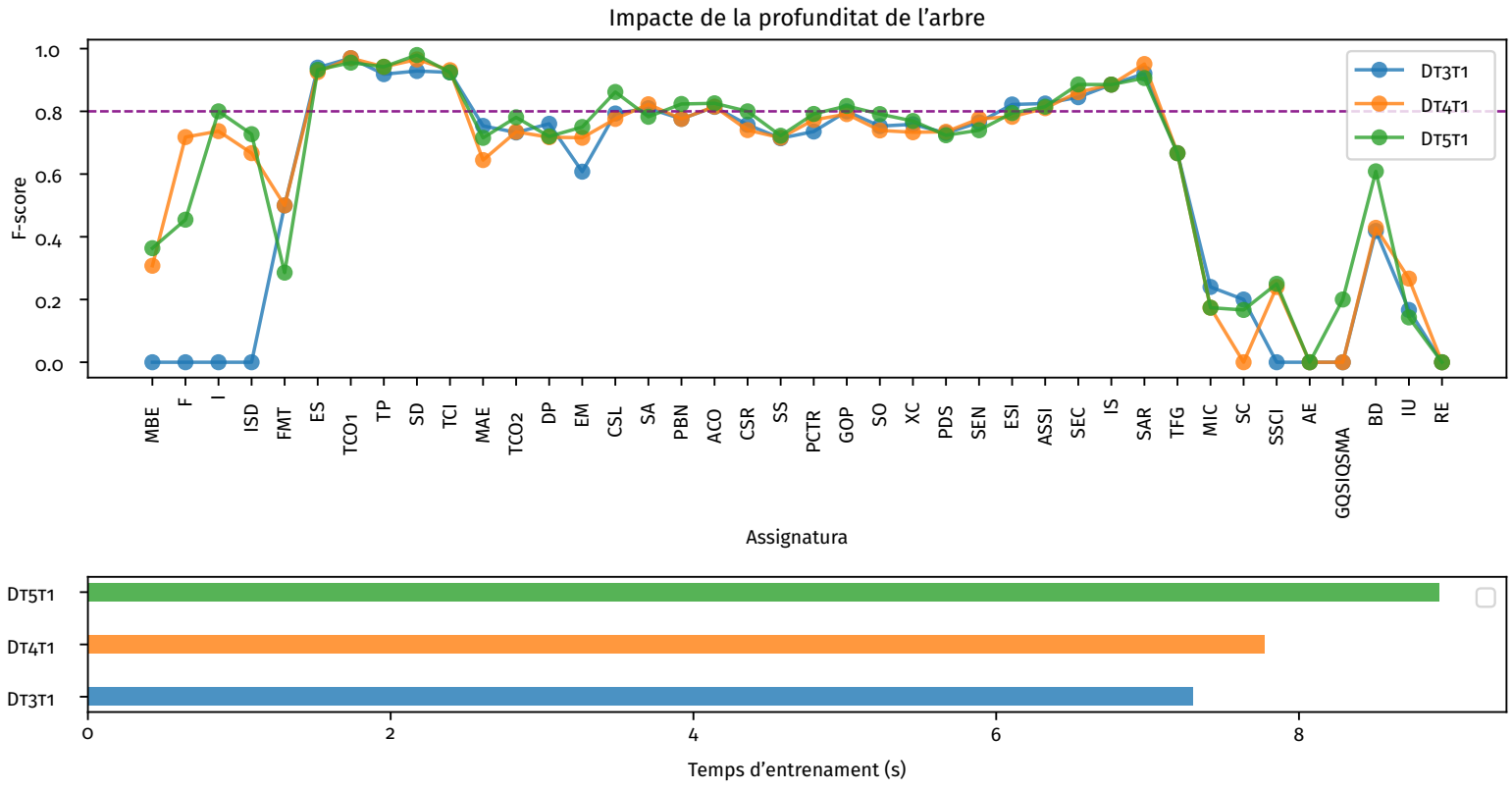








Arbres de decisió: Profunditat de l'arbre



Arbres de decisió: Profunditat de l'arbre

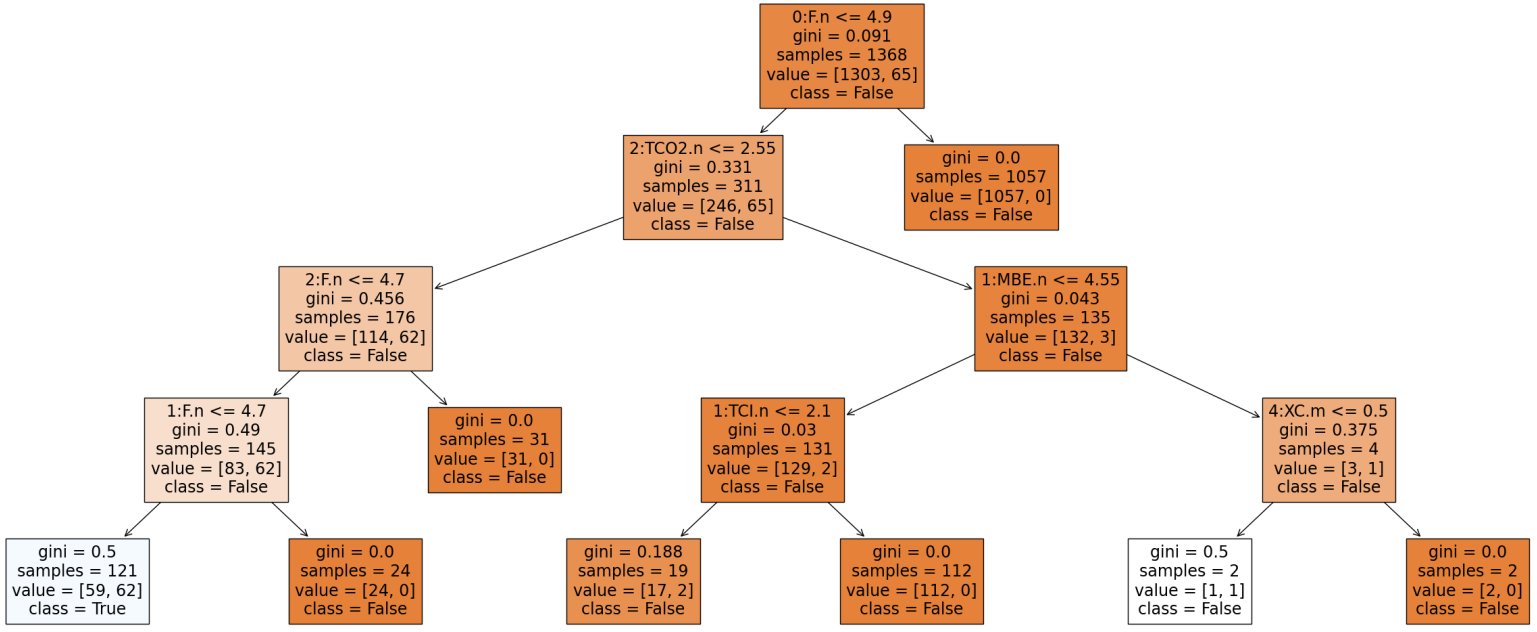
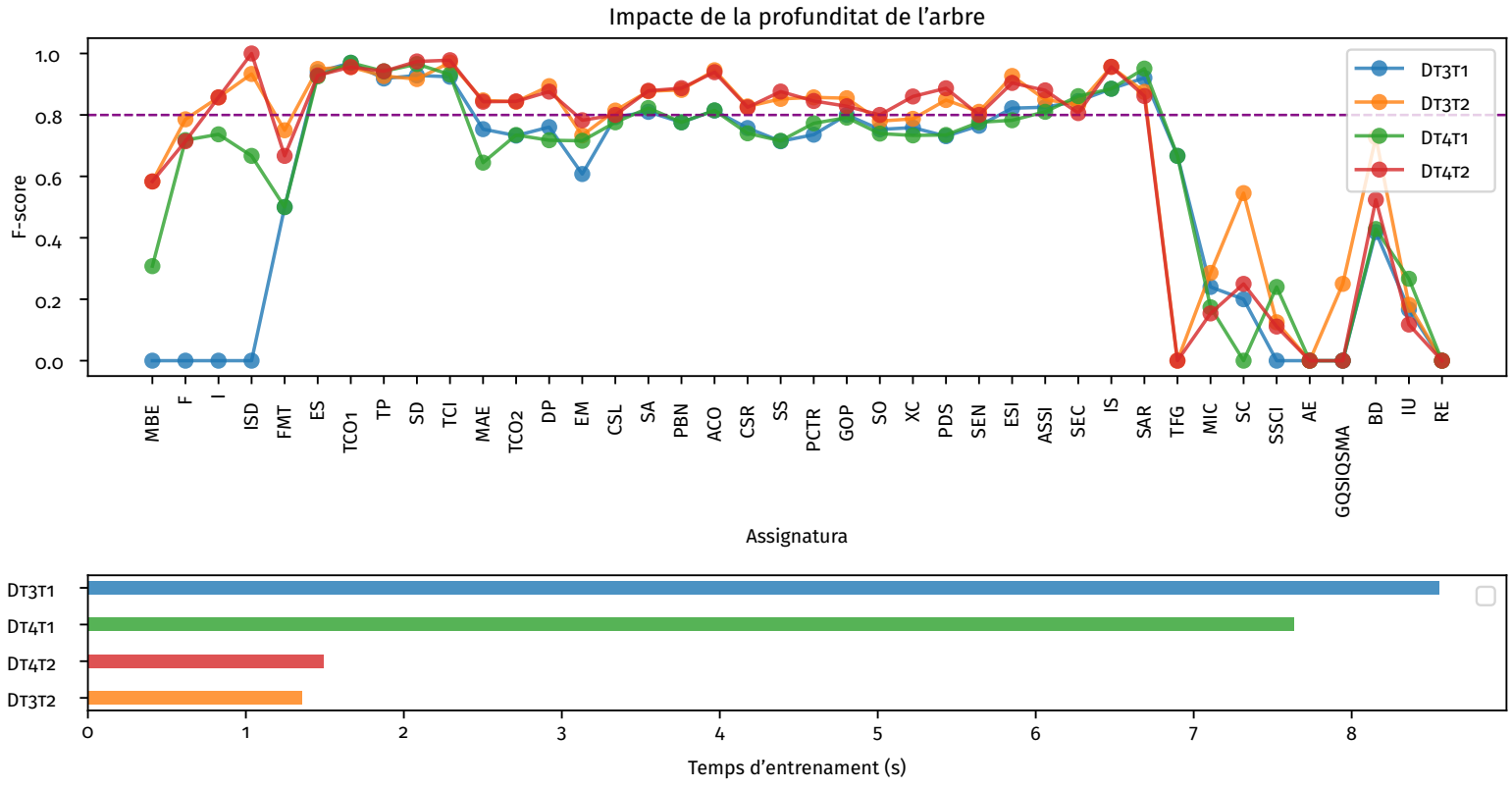


Figura: Arbre de F (ID: DT4T1)

Arbres de decisió: Transformació de les dades

Identificador (Id)	*Model*	*Profunditat*	*Transformació*	*Motxilla*
DT3T1	ARBRE	3	1	No
DT3T2	ARBRE	3	2	No
DT4T1	ARBRE	4	1	No
DT4T2	ARBRE	4	2	No



Arbres de decisió: Transformació de les dades

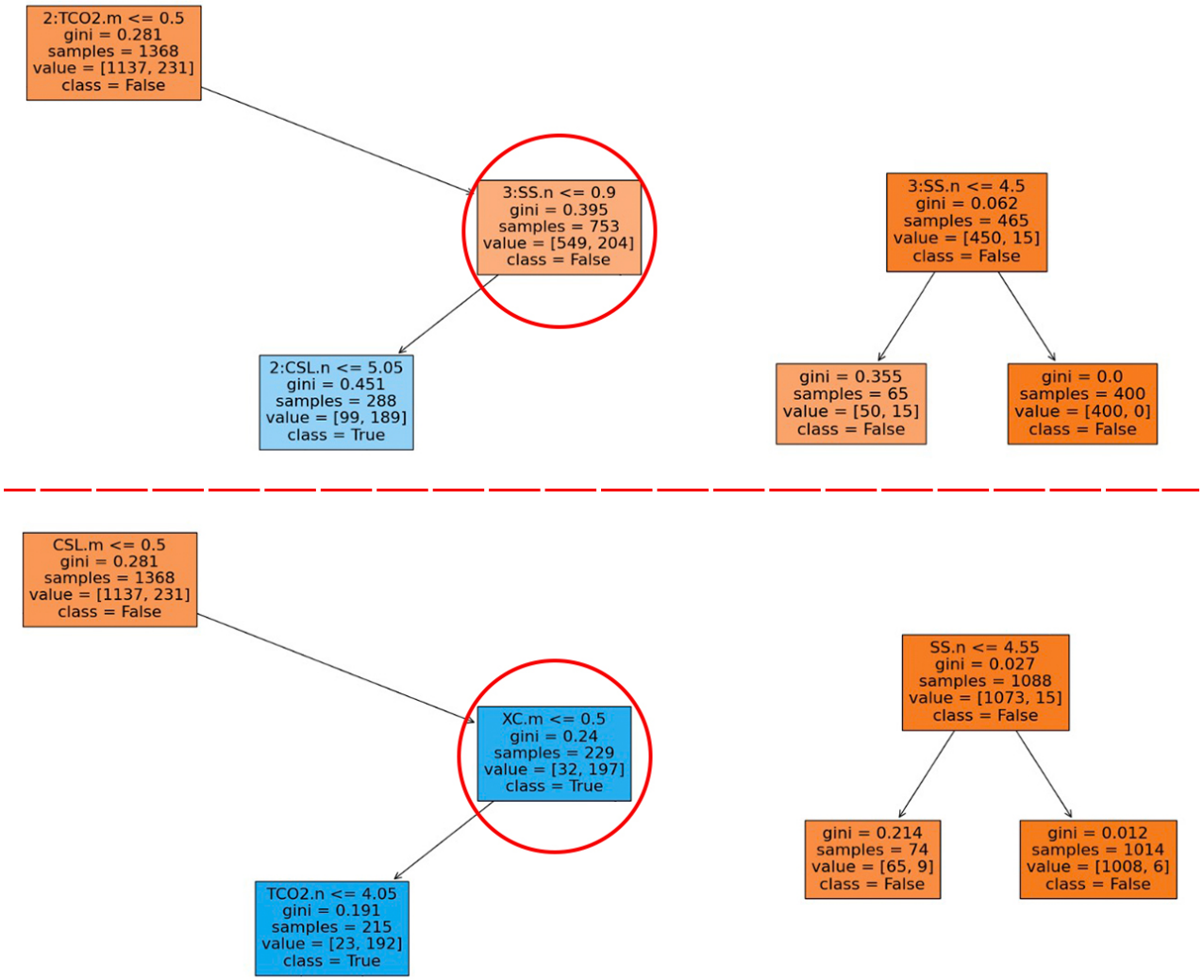
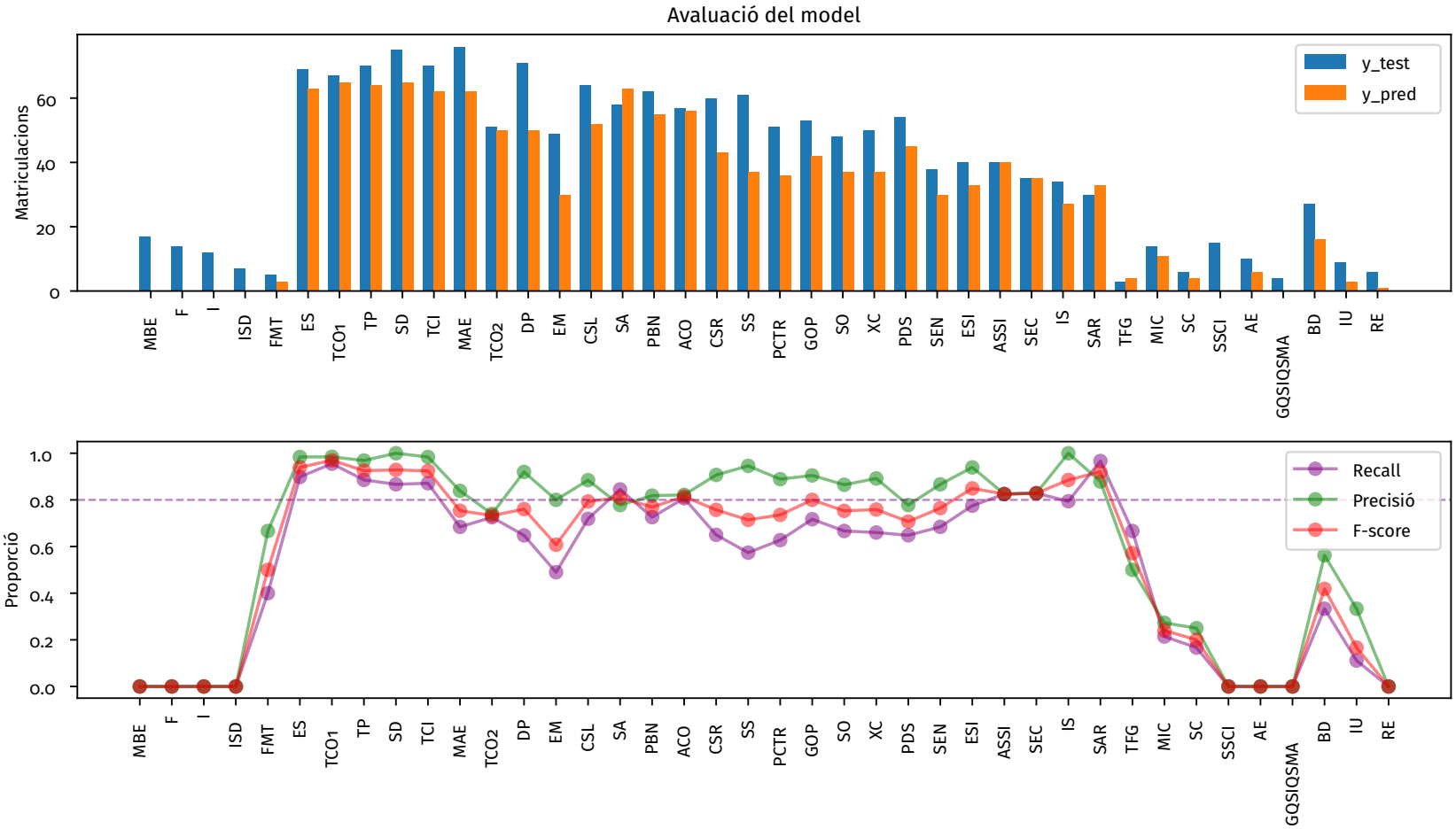


Figura: Diferències entre DT3T1 i DT3T2 de l'assignatura de SS[2]

[2] La part superior representa el model DT3T1 i la inferior el model DT3T2

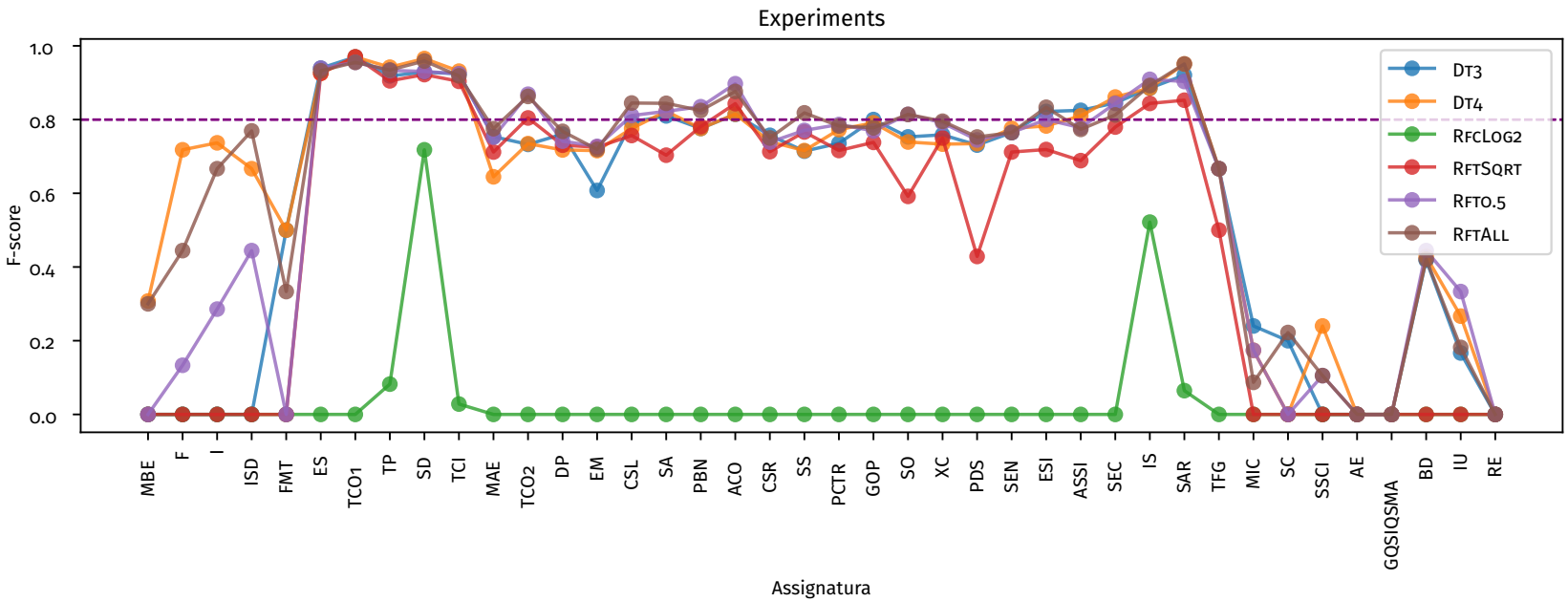
Arbres de decisió: Motxilla de l'estudiant

Identificador (Id)	*Model*	*Profunditat*	*Transformació*	*Motxilla*
DT3T1	ARBRE	4	1	No
DT3T2	ARBRE	4	2	No
DT3T1M	ARBRE	4	1	Sí
DT3T2M	ARBRE	4	2	Sí



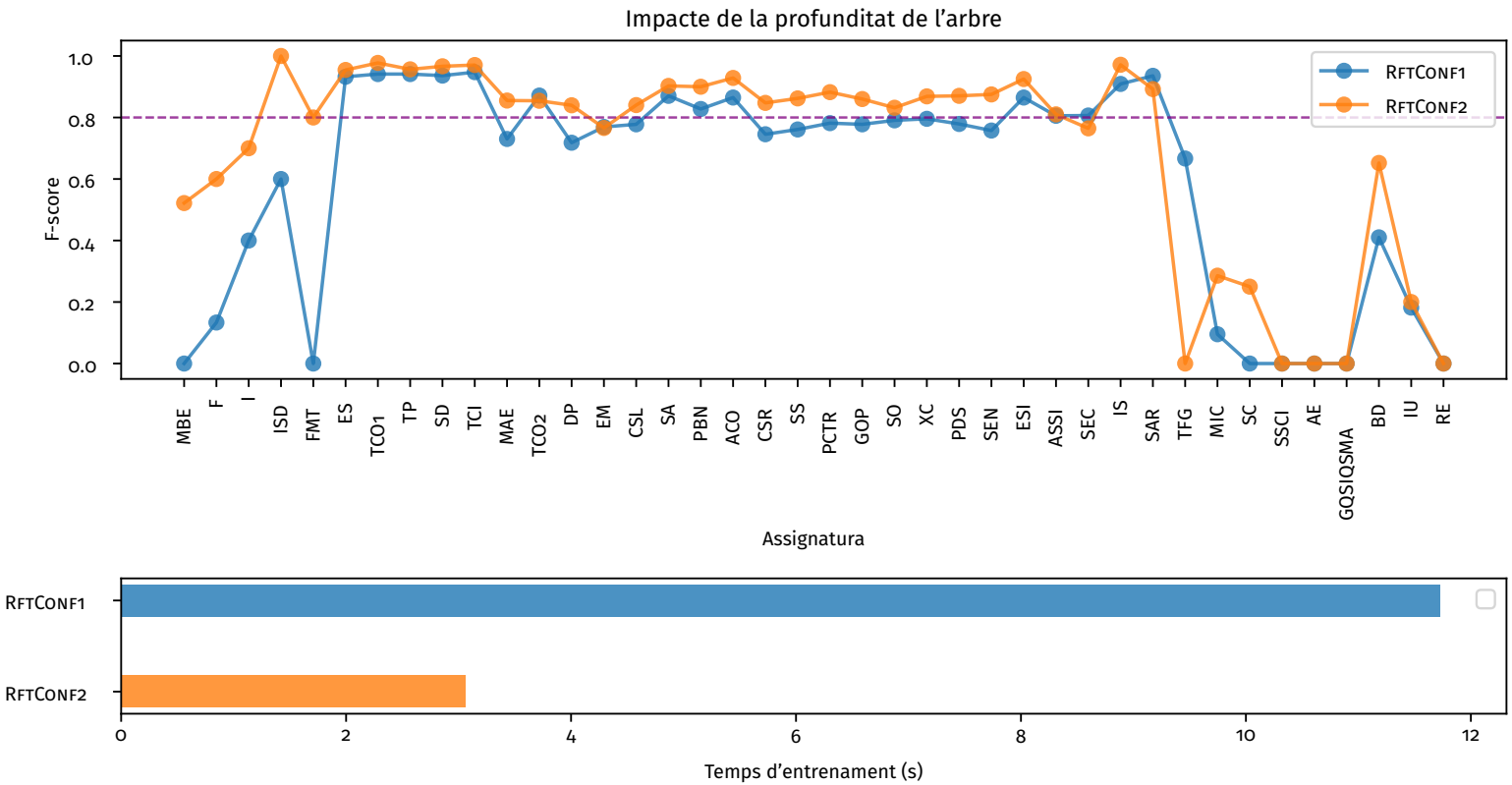
Boscos d'arbres aleatoris

Identificador	*Model*	*Profunditat*	*Transformació*	*Estimadors*	*Features*	*Mostres*
DT3	ARBRE	3	1	1	n_features	n_samples
DT4	ARBRE	4	1	1	n_features	n_samples
RFTLOG2	BOSC	4	1	20	log_2	n_samples
RFTSQRT	BOSC	4	1	20	sqrt	n_samples
RFTO.5	BOSC	4	1	20	0.5	n_samples
RFTALL	BOSC	4	1	20	n_features	n_samples



Boscos d'arbres aleatoris

Identificador	*Model*	*Profunditat*	*Transformació*	*Estimadors*	*Features*	*Mostres*
RFTCONF1	BOSC	4	1	15	0.5	0.8
RFTCONF2	BOSC	4	2	15	0.5	0.8



Boscos d'arbres aleatoris

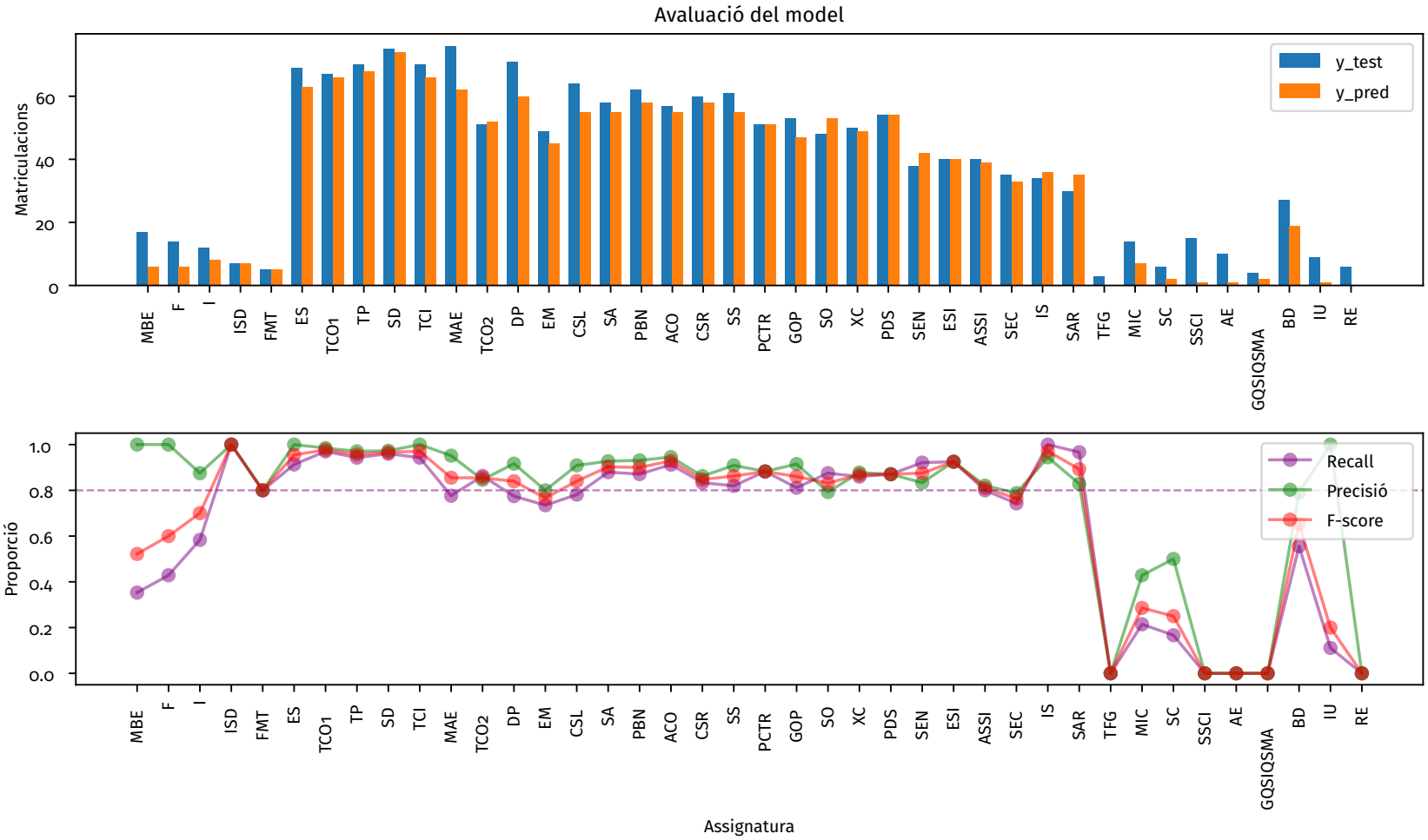


Figura: Avaluació dels models de les assignatures (ID: RFTCONF2)

Obtenció del millor bosc

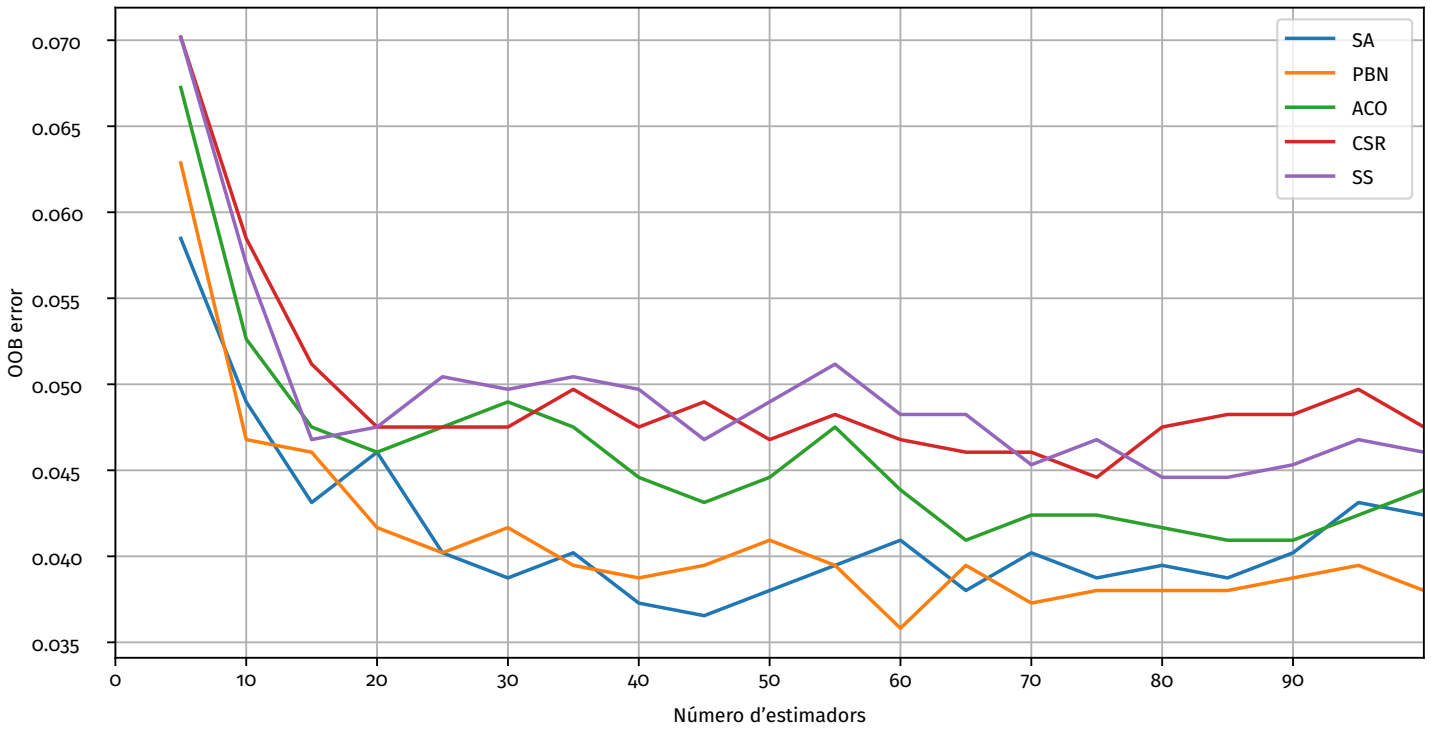


Figura: Obtenció del millor bosc d'arbres de les assignatures de Q4 (ID: RFTCONF2)

Obtenció del millor bosc

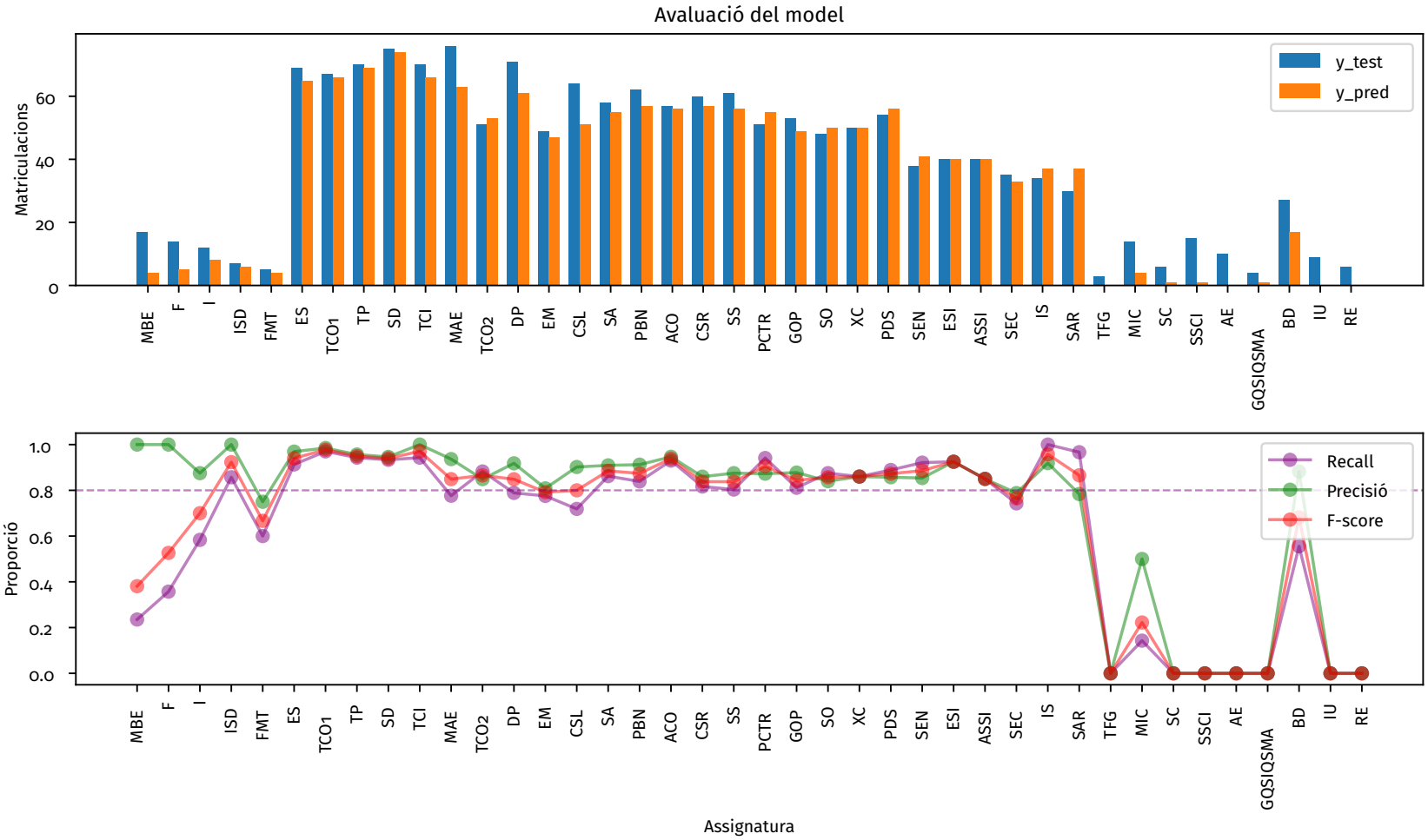
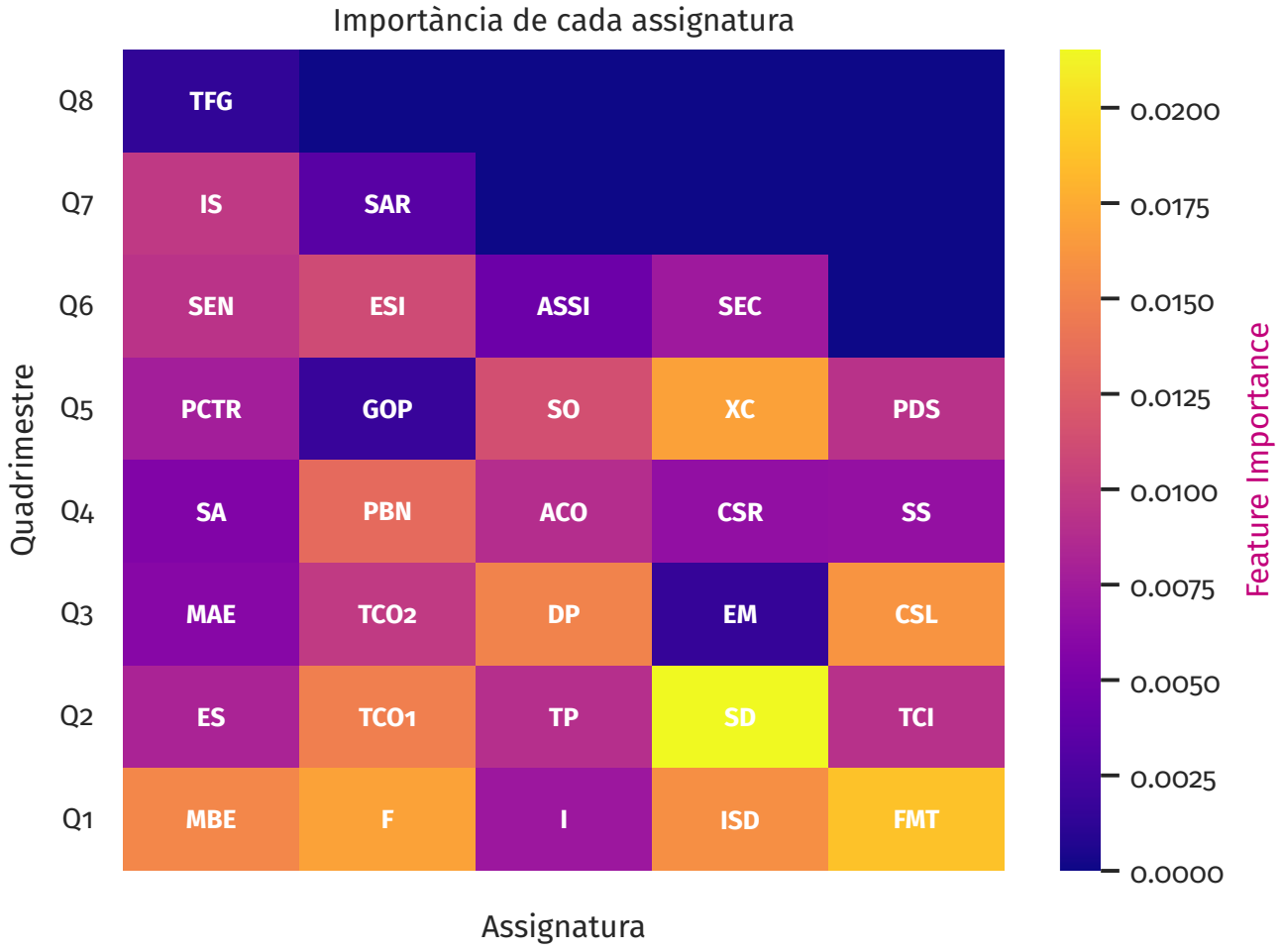


Figura: Avaluació del millor bosc d'arbres de cada assignatura (ID: RFTCONF2)

Importància de les assignatures



Conclusions i treball futur

- El model DT3T1, amb una profunditat de 3 nodes, proporciona prediccions precises en la majoria dels casos.
- El model DT4T1, amb una profunditat de 4 nodes, millora especialment en les assignatures del primer quadrimestre, obtenint regles implícites com la regla dels repetidors.
- S'ha observat que utilitzar arbres de decisió amb una profunditat de 5 nodes, com el model DT5T1, pot conduir a problemes de overfitting.
- També, s'ha vist que la incorporació de dades addicionals de l'estudiant, com ara la nota d'accés, la via d'accés, l'assignació de beca, entre altres, no altera el patró de matriculació.

Conclusions i treball futur

- S'ha observat que amb els boscos d'arbres es poden realitzar tots els experiments possibles, a més, s'obtenen millors resultats. Ofereixen avantatges com ara la capacitat de calcular la importància de les característiques de les dades.
- Els boscos d'arbres suggereixen que les assignatures dels primers quadrimestres tenen una importància significativa per a la continuïtat dels estudiants en la carrera. A més, s'han identificat diverses assignatures que destaquen per sobre de les altres.
- Cal tenir en compte que el comportament dels estudiants pot canviar amb el temps i seguir tendències que no es poden preveure amb anticipació.
- Seria interessant provar altres possibles transformacions i diferents algorismes d'aprenentatge automàtic.
- Finalment, seria interessant desenvolupar un sistema complet que integri la «solució» definitiva.