APC - CAS KAGGLE
Brain Tumor Data Set

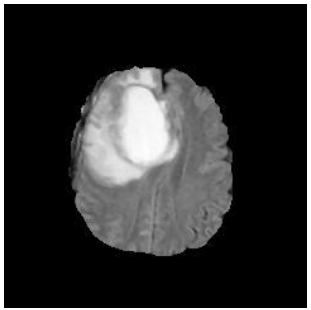
https://www.kaggle.com/jakeshbohaju/brain-tumor
https://github.com/marcgarrofe/ApC-Kagle

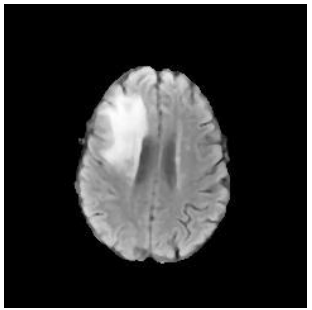
Marc Garrofé, 1565644

INTRODUCCIÓ

En el nostre problema ens presenten un dataset amb informació referent a les característiques d'imatges cerebrals.

La nostra fita és generar un model que classifiqui i identifiqui aquelles imatges que tenen tumor.



DADES

El nostre dataset està format per 15 columnes. Les dues primeres fan referència a la ID de la imatge i la segona a la classificació final (estat Objectiu). Les 13 restants són propietats de la imatge.

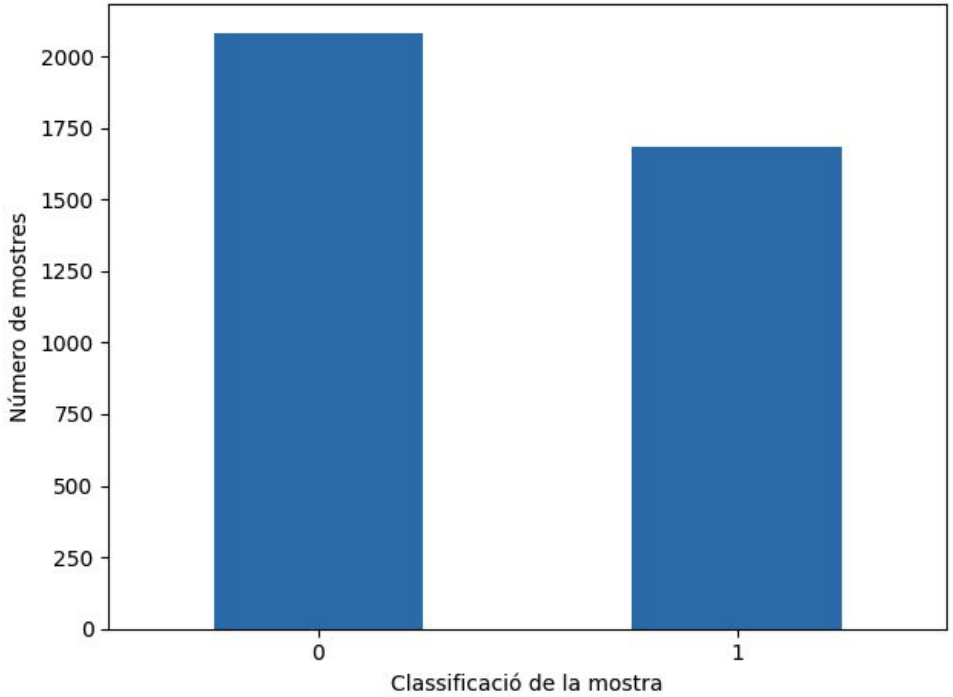
```
dataset.head()
```

	Image	Class	Mean	Variance	Standard Deviation	Entropy	Skewness	Kurtosis	Contrast	Energy	ASM	Homogeneity	Dissimilarity	Correlation	Coarseness
0	Image1	0	6.535	619.588	24.892	0.109	4.276	18.901	98.614	0.293	0.086	0.531	4.473	0.982	0.000
1	Image2	0	8.750	805.958	28.389	0.267	3.718	14.465	63.859	0.475	0.226	0.651	3.220	0.989	0.000
2	Image3	1	7.341	1143.808	33.820	0.001	5.062	26.480	81.867	0.032	0.001	0.268	5.982	0.978	0.000
3	Image4	1	5.958	959.712	30.979	0.001	5.678	33.429	151.230	0.032	0.001	0.244	7.701	0.964	0.000
4	Image5	0	7.315	729.541	27.010	0.147	4.283	19.079	174.989	0.344	0.118	0.501	6.835	0.973	0.000

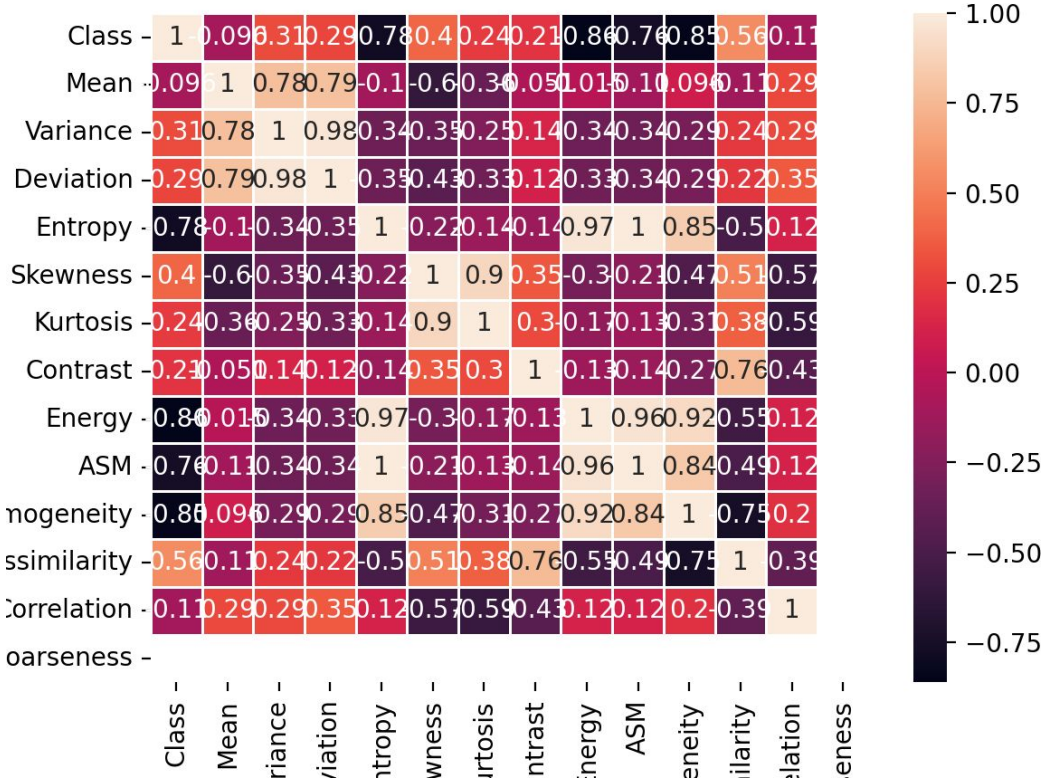
PREPROCESSING

Observem que en el dataset no hi han dades faltants i està balancejat.

A nivell de distribució de les dades és molt divers, algunes si estan estandaritzades pero la majoria d'elles no.



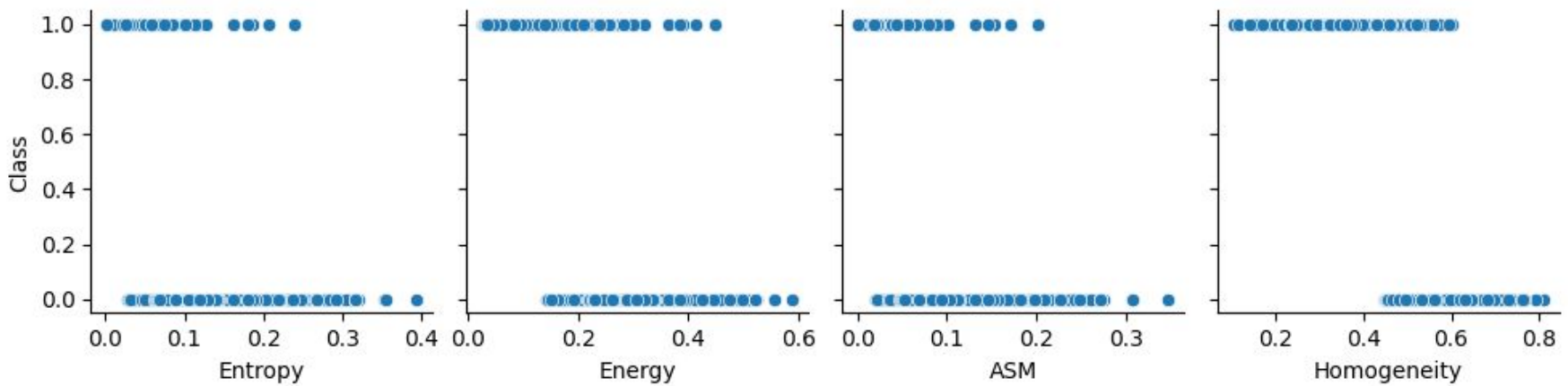
CORRELACIÓ



Atributs principals:

- Entropy (78%)
- Energy (80%)
- ASM (76%)
- Homogenity (85%)

Distribució atributs principals segons la classe



GRID SEARCH

Per trobar la millor combinació de paràmetres i configuracions per definir un model eficaç i òptim, definim una funció per provar amb tots els paràmetres.

- Estandarització
- Feature Selection
- Grid Search (paràmetres model)

MÈTODES UTILITZATS

Regresor Logístic

Gradient Descent

Suported Vector Classfication

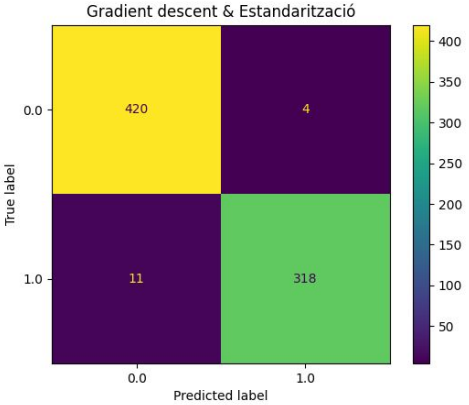
Random Forest

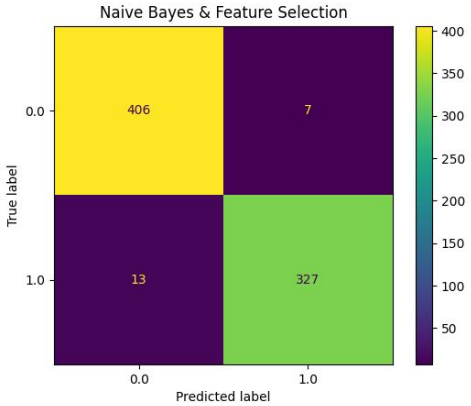
Naive Bayes

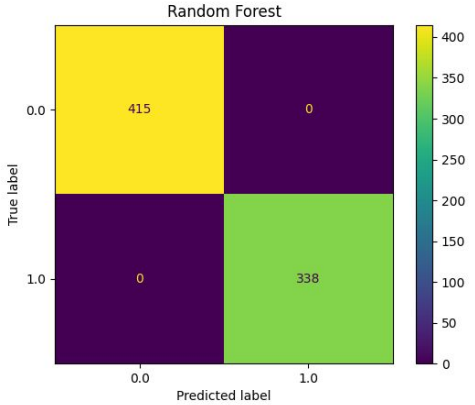
RESULTATS - CrossValidation

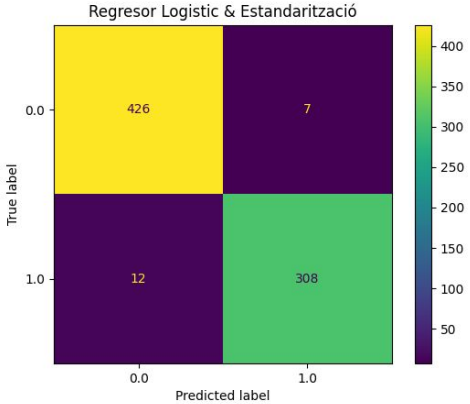
Model	Preprocessing	Hiperparametres	Mètrica	Temps
Regresor Logístic	Estandarització	penalty: l2, solver: newton-cg, warm_start: True	98.30 %	1.87 s
Gradient Descent	Estandarització	'alpha': 0.001, 'class_weight': 'balanced', 'fit_intercept': True, 'learning_rate': 'optimal', 'loss': 'log', 'max_iter': 5000, 'penalty': 'l2', 'shuffle': True, 'warm_start': 'True'	98.24 %	4.78 s
SVC	Estandarització	'class_weight': 'balanced', 'gamma': 'scale', 'kernel': 'linear'	98.30 %	1.37 s
Random Forest	default	'balanced_subsample', 'criterion': 'gini', 'max_features': 'auto', 'n_estimators': 75, 'warm_start': True	98.91 %	101.31 s
Naive Bayes	Feature Selection	default	97.25 %	0.01 s

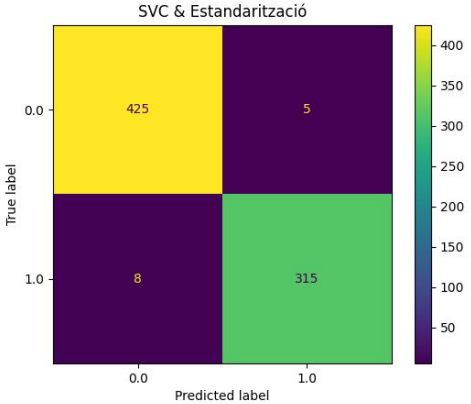
Confussion Matrix - Best Models



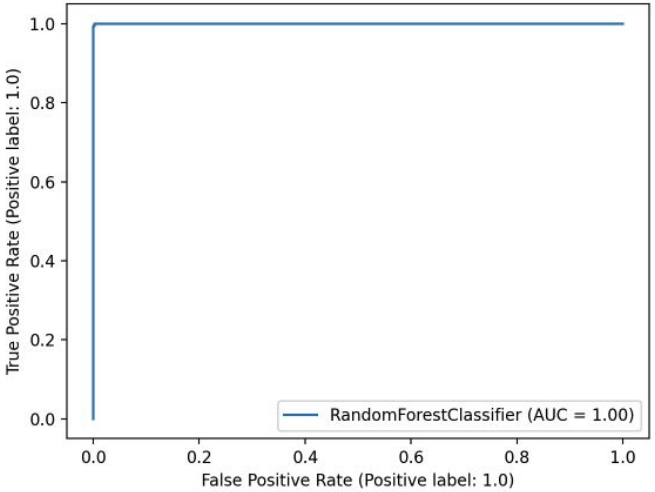


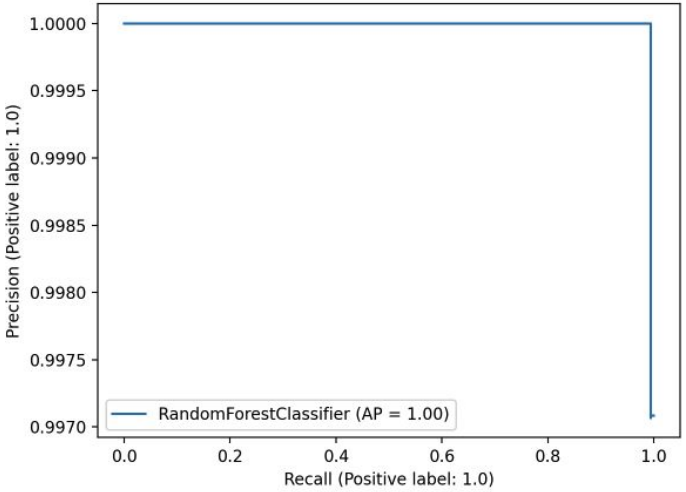






RANDOM FOREST





ROC & Recall

CONCLUSIONS

Concluïm que Random Forest és el model que millor classifica les imatges cerebrals.

Tot i que Random Forest és el més lent, la diferència respecte els altres models és imperceptible i en l'àmbit que s'aplica no repercutiria doncs es busca la màxima precisió.

Per un treball a futur, es poden definir diferents models que analitzessin tot tipus de proves mèdiques com ara analítiques, biopsies, ecografías, etc. per tal de tractar o detectar amb antel.lació malalties.