Cas Kaggle:
Student Grade Prediction

Raúl Villar – NIU:1596830

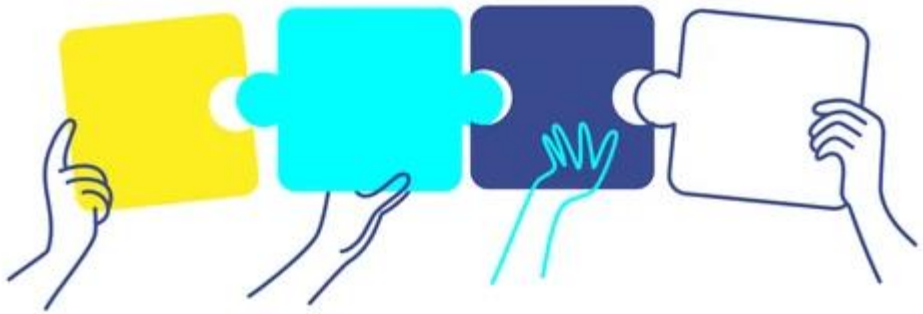
Github: https://github.com/NIU1596830/Cas-Kaggle

OBJECTIUS

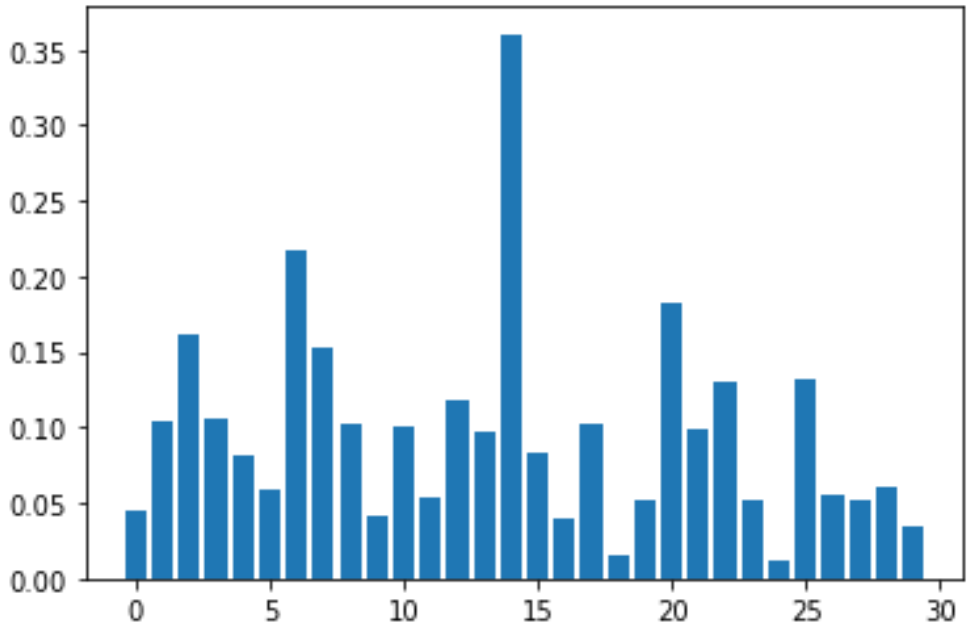
- Analitzar una base de dades sobre l'información dels estudiants i la seva nota.
- Crear un model que pugui predir la nota.

Introducció al dataset

- 29 atributs categòrics i 4 atributs numerics
- El target és l'atribut 'G3'
- Total de 395 mostres

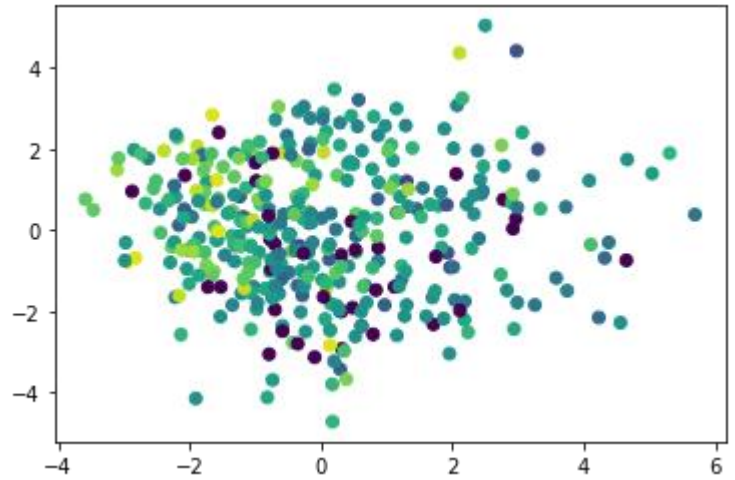


PREPROCESSING (NORMALITZACIÓ, OUTLIER REMOVAL, FEATURE SELECTION)



Taula de la correlació dels atributs amb G3

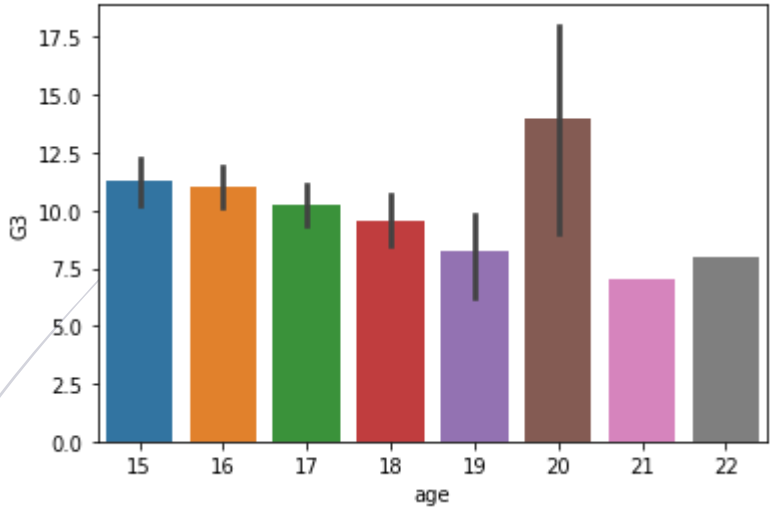
- Normalitzem ‘absences’
- No hi ha dades sense informació
- Eliminació de ‘G1’ i ‘G2’

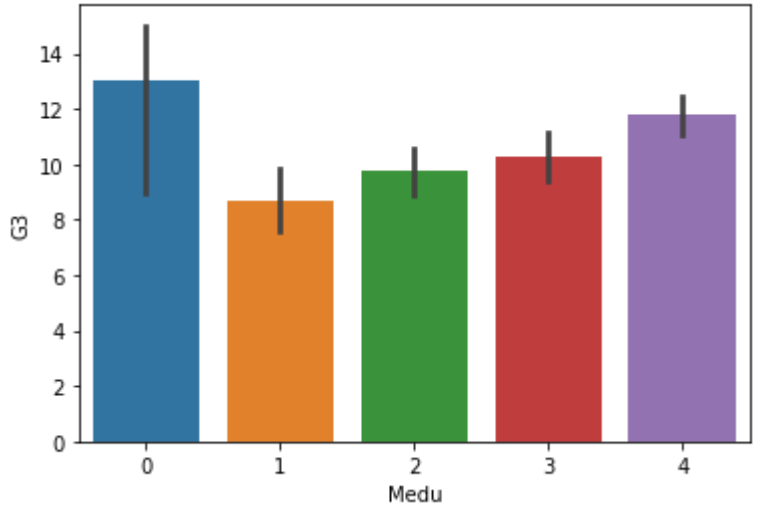


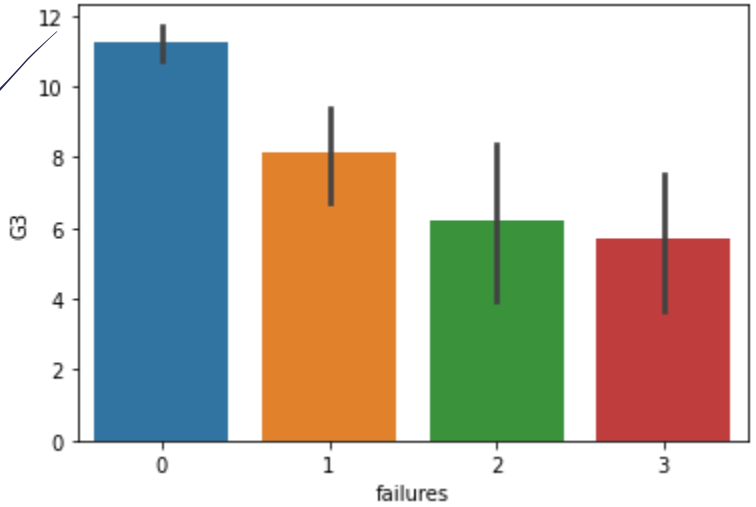
PCA de 2 components

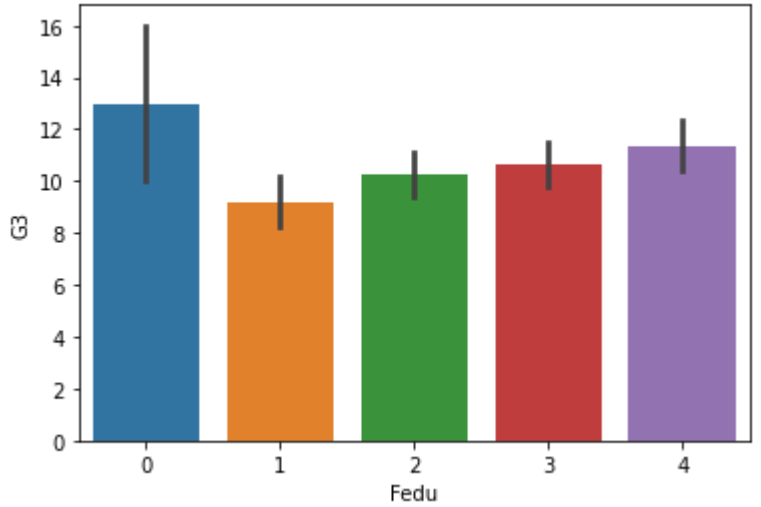
# age	# Medu	# Fedu	# failures
student's age (numeric: from 15 to 22)	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary	number of past class failures (numeric: n if 1<=n<3, else 4)

Atributs amb més correlació



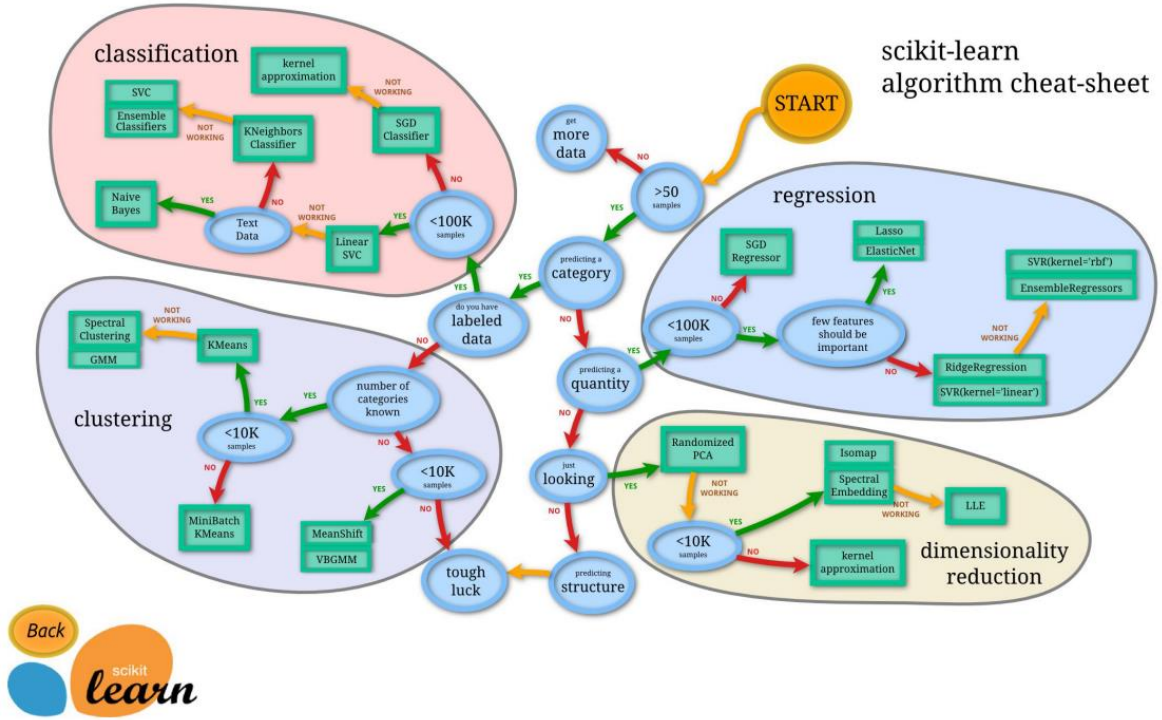






Tenim: 3 mostres amb Medu=0 de 395 mostres
Tenim: 2 mostres amb Fedu=0 de 395 mostres
Tenim: 3 mostres amb age=20 de 395 mostres

MODEL SELECTION



MODELS A CONSIDERAR

- SGDRegressor
- Regressió lineal simple
- Random forest
- Kneighbors

SGDRegressor

```
R-squared: -0.5898457278834739
CV mean score:  -0.42480406604700854
MSE:  28.714277962381093
RMSE:  5.358570514827727
```

Dataset alta correlació

```
R-squared: -0.12958784100796694
CV mean score:  -0.21441775074861136
MSE:  23.317715657346785
RMSE:  4.8288420617521535
```

Dataset alta i mitja correlació

```
R-squared: -0.5950925624331187
CV mean score:  -0.190549715330487
MSE:  36.171199640629624
RMSE:  6.014249715519769
```

Dataset amb tots els atributs

```
R-squared: 0.06753545451472598
CV mean score:  0.03653920750935029
MSE:  15.995397360633802
RMSE:  3.9994246286977084
```

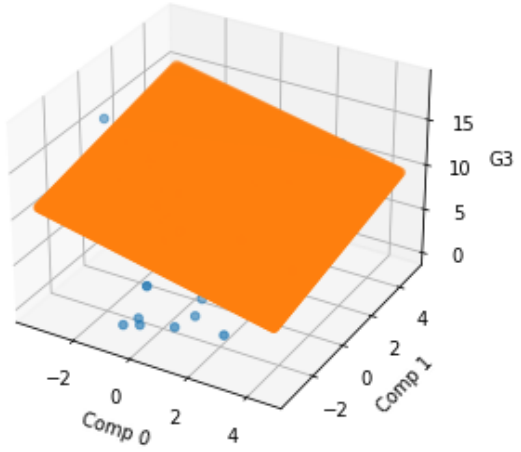
Dataset PCA

Regressió lineal simple

```
        age :
Mean squeared error:  21.67211703017247
R2 score:  -0.018396961634633335
        Medu :
Mean squeared error:  18.375831765579477
R2 score:  0.08251754245711496
        Fedu :
Mean squeared error:  24.116308533937907
R2 score:  0.019761161776679637
        failures :
Mean squeared error:  19.049028224953805
R2 score:  0.1152220066171733
```

```
        PC0 :
Mean squeared error:  21.734970967929293
R2 score:  -0.055223586897303
        PC1 :
Mean squeared error:  19.73795716334648
R2 score:  0.00745193510176001
```

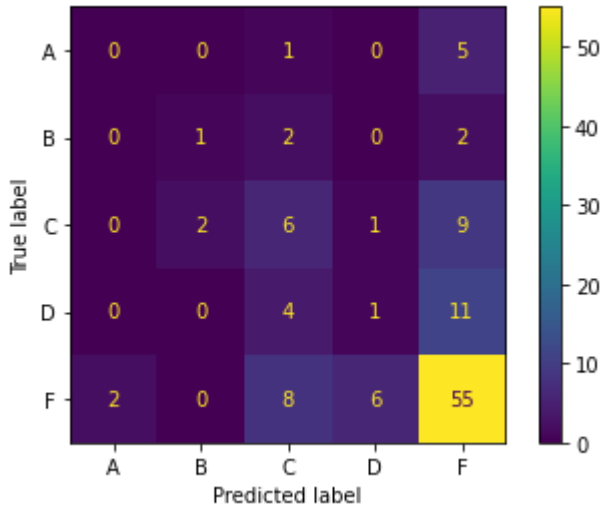
```
Mean squeared error:  15.10704094166262
R2 score:  0.11013682844067518
```



ENSEMBLE RANDOM FOREST

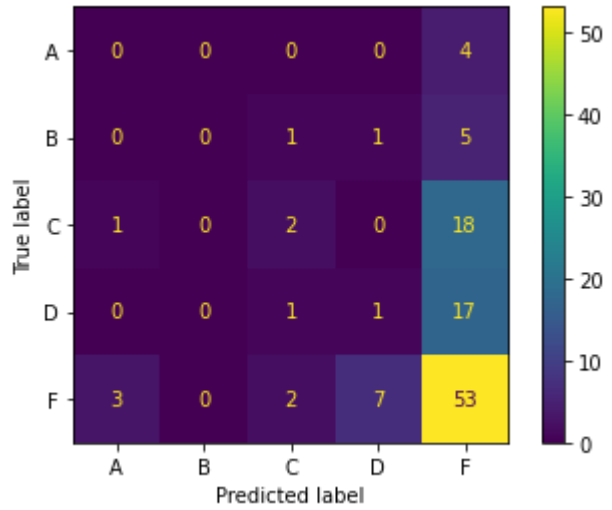
Letter Grade	Percentage	GPA
A	90–100%	4.0
B	80–89%	3.0
C	70–79%	2.0
D	60–69%	1.0
F	0–59%	0.0

Atributs alta correlació



Accuracy test: 54.31%

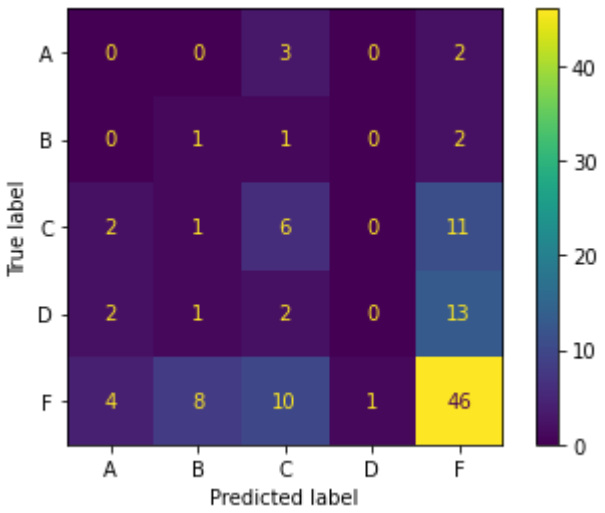
PCA



Accuracy test: 48.27%

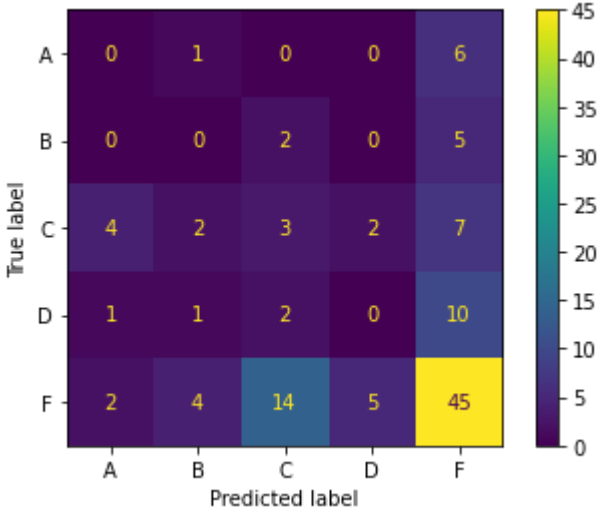
KNEIGHBORS

Atributs alta correlació



Accuracy test: 45.68%

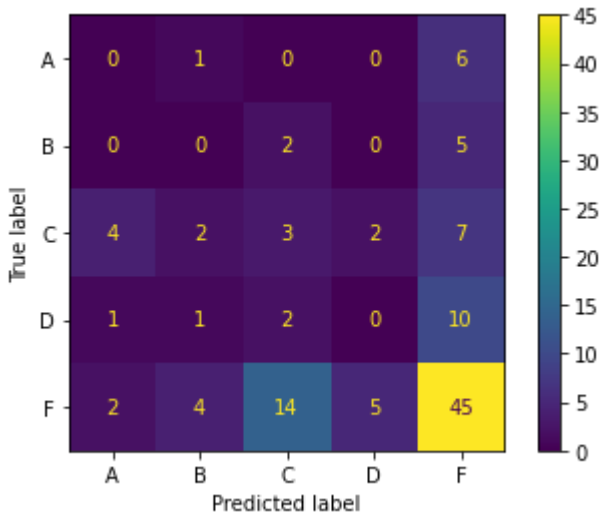
PCA



Accuracy test: 41.37%

Regressió lineal a categoric

Atribut ‘failures’



Accuracy test: 66%

CONCLUSIONS