

Projet Machine Learning Specialist

Analyse, Classification et Prédiction de la performance des étudiants d'école secondaire

Keven BELLEC

(keven.bellec@supinfo.com)

Celso SANCHEZVIERA










(celso.sanchezviera@supinfo.com)

Campus Rennes

Le travail ci présent expose des études réalisées à propos de deux Data Sets, l'un concernant les caractéristiques et résultats des étudiants pour la matière mathématique, et l'autre pour la matière de la langue Portugaise.

Il était nécessaire de mener aussi les études sur trois quantités de caractéristiques différents pour chaque Data Set.

Dans l'intérêt d'une bonne compréhension, le projet a été structuré de la façon suivante :

 M-30Var	Data Set : Mathématiques – 30 variables
 M-31Var	Data Set : Mathématiques – 31 variables
 M-32Var	Data Set : Mathématiques – 32 variables
 P-30Var	Data Set : Portugais – 30 variables
 P-31Var	Data Set : Portugais – 31 variables
 P-32Var	Data Set : Portugais – 32 variables
 DataAnalysis.ipynb	Analyse des Données communs
 student-mat.csv	
 student-por.csv	

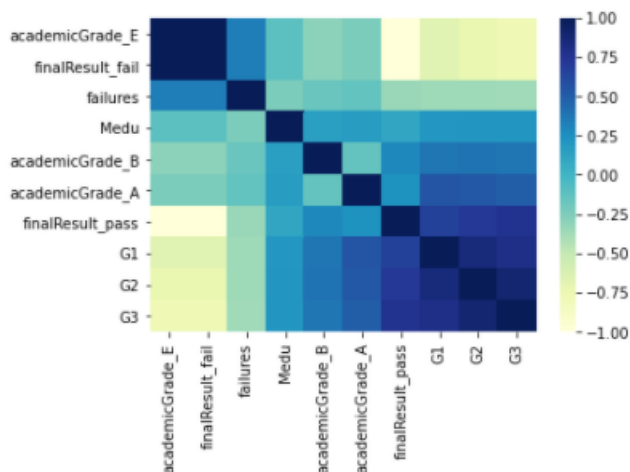
Contents

Classification binaire sur la variable "finalResult"	4
Classification binaire sur la variable "finalResult"	5
Classification binaire sur la variable "finalResult"	6
Multivariate classification on the variable "academicGrade" – 30	7
Multivariate classification on the variable "academicGrade" - 31	8
Multivariate classification on the variable "academicGrade"-32	9
Régression sur la variable "G3" - 31	10
Régression sur la variable "G3" - 32	10
Régression sur la variable "G3" - 32	11
Conclusion.	11

1. Charger l'ensemble de données.
2. Ajouter une variable catégorielle binaire appelée "finalResult", avec comme niveau "pass" si la variable G3 est supérieure ou égale à 10 est "fail" dans le cas contraire.
3. Ajouter une variable catégorielle appelée "AcademicGrade", avec cinq niveaux "A", "B", "C", "D", "E" selon que la variable G3 est comprise entre 16 et 20, 14 et 15, 12 et 13, 10 et 11, 0 et 9.
4. Effectuer une analyse exploratoire complète.

On a créé une méthode pour calculer et filtrer les features qui présentent une corrélation avec Target(G3) d'une valeur absolue majeure que 0.2 et les trier par leur importance. Ou comme on l'a appelée la corrélation significative.

Ce Heat Map de corrélation en est le résultat. On voit plus clairement les corrélations et on peut conclure que les échecs antérieurs, l'éducation de la mère, G1 et G2 sont les variables avec les corrélations les plus fortes.



Métriques :

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

		ACTUAL	
		Negative	Positive
PREDICTION	Negative	TRUE NEGATIVE	FALSE NEGATIVE
	Positive	FALSE POSITIVE	TRUE POSITIVE

Puis que la métrique F1 est une métrique comportant les résultats du reste.

On va l'utiliser comme mesure générale de qualité du model.

Classification binaire sur la variable "finalResult"

30 variables Maths

KNeighbors Score	0.6456
KNeighbors Precision Score	0.5527
KNeighbors Recall Score	0.5771
KNeighbors F1 Score	0.5460
Support Vector Machine Score	0.6835
Support Vector Machine Precision Score	0.5726
Support Vector Machine Recall Score	0.6478
Support Vector Machine F1 Score	0.5589
Linear Discriminant Analysis Score	0.5823
Linear Discriminant Analysis Precision Score	0.5046
Linear Discriminant Analysis Recall Score	0.5055
Linear Discriminant Analysis F1 Score	0.5003
Logistic Regression Score	0.5823
Logistic Regression Precision Score	0.5135
Logistic Regression Recall Score	0.5152
Logistic Regression F1 Score	0.5119
Decision Tree Classifier Score	0.5823
Decision Tree Classifier Precision Score	0.5224
Decision Tree Classifier Recall Score	0.5239
Decision Tree Classifier F1 Score	0.5223
Voting Score	0.5949
Voting Precision Score	0.5053
Voting Recall Score	0.5071
Voting F1 Score	0.4960
Random Forest Classifier Score	0.5823
Random Forest Classifier Precision Score	0.5135
Random Forest Classifier Recall Score	0.5152
Random Forest Classifier F1 Score	0.5119
Gradient Boosting Classifier Score	0.6076
Gradient Boosting Classifier Precision Score	0.5506
Gradient Boosting Classifier Recall Score	0.5538
Gradient Boosting Classifier F1 Score	0.5512
Ada Boost Classifier Score	0.6076
Ada Boost Classifier Precision Score	0.5417
Ada Boost Classifier Recall Score	0.5467
Ada Boost Classifier F1 Score	0.5415

30 variables Portugais

KNeighbors Score	0.8000
KNeighbors Precision Score	0.5030
KNeighbors Recall Score	0.5120
KNeighbors F1 Score	0.4797
Support Vector Machine Score	0.8077
Support Vector Machine Precision Score	0.5760
Support Vector Machine Recall Score	0.6321
Support Vector Machine F1 Score	0.5873
Linear Discriminant Analysis Score	0.8000
Linear Discriminant Analysis Precision Score	0.5884
Linear Discriminant Analysis Recall Score	0.6261
Linear Discriminant Analysis F1 Score	0.5993
Logistic Regression Score	0.8077
Logistic Regression Precision Score	0.5760
Logistic Regression Recall Score	0.6321
Logistic Regression F1 Score	0.5873
Decision Tree Classifier Score	0.8077
Decision Tree Classifier Precision Score	0.6443
Decision Tree Classifier Recall Score	0.6614
Decision Tree Classifier F1 Score	0.6517
Voting Score	0.8154
Voting Precision Score	0.5977
Voting Recall Score	0.6581
Voting F1 Score	0.6131
Random Forest Classifier Score	0.8077
Random Forest Classifier Precision Score	0.5760
Random Forest Classifier Recall Score	0.6321
Random Forest Classifier F1 Score	0.5873
Gradient Boosting Classifier Score	0.8077
Gradient Boosting Classifier Precision Score	0.5760
Gradient Boosting Classifier Recall Score	0.6321
Gradient Boosting Classifier F1 Score	0.5873
Ada Boost Classifier Score	0.7846
Ada Boost Classifier Precision Score	0.5449
Ada Boost Classifier Recall Score	0.5726
Ada Boost Classifier F1 Score	0.5486

On trouve que la classification Binaire avec 30 variables donne des résultats plutôt médiocres.

La répartition de données de Test et Train a été faite à un taux de 80% Test et 20% Train afin de maximiser la quantité de données utilisées par le modèle.

Meilleures performances pour les notes de Portugais.

Meilleures performances Score F1 :

Maths		Portugais	
Support Vector Machine	0.5589	Decision Tree Classifier	0.6517
Gradient Boost	0.5512	Voting Classifier	0.6131

Classification binaire sur la variable "finalResult"

31 variables Maths

KNeighbors Score	0.6456
KNeighbors Precision Score	0.5578
KNeighbors Recall Score	0.5612
KNeighbors F1 Score	0.5590
Support Vector Machine Score	0.7722
Support Vector Machine Precision Score	0.7240
Support Vector Machine Recall Score	0.7240
Support Vector Machine F1 Score	0.7240
Linear Discriminant Analysis Score	0.7722
Linear Discriminant Analysis Precision Score	0.7368
Linear Discriminant Analysis Recall Score	0.7259
Linear Discriminant Analysis F1 Score	0.7307
Logistic Regression Score	0.7975
Logistic Regression Precision Score	0.7162
Logistic Regression Recall Score	0.7642
Logistic Regression F1 Score	0.7322
Decision Tree Classifier Score	0.7975
Decision Tree Classifier Precision Score	0.7803
Decision Tree Classifier Recall Score	0.7571
Decision Tree Classifier F1 Score	0.7659
Voting Score	0.8228
Voting Precision Score	0.7725
Voting Recall Score	0.7882
Voting F1 Score	0.7795
Random Forest Classifier Score	0.7975
Random Forest Classifier Precision Score	0.7162
Random Forest Classifier Recall Score	0.7642
Random Forest Classifier F1 Score	0.7322
Gradient Boosting Classifier Score	0.8481
Gradient Boosting Classifier Precision Score	0.7776
Gradient Boosting Classifier Recall Score	0.8392
Gradient Boosting Classifier F1 Score	0.7992
Ada Boost Classifier Score	0.8354
Ada Boost Classifier Precision Score	0.7814
Ada Boost Classifier Recall Score	0.8072
Ada Boost Classifier F1 Score	0.7923

31 variables Portugais

KNeighbors Score	0.8615
KNeighbors Precision Score	0.4870
KNeighbors Recall Score	0.4409
KNeighbors F1 Score	0.4628
Support Vector Machine Score	0.8846
Support Vector Machine Precision Score	0.7029
Support Vector Machine Recall Score	0.7155
Support Vector Machine F1 Score	0.7089
Linear Discriminant Analysis Score	0.8923
Linear Discriminant Analysis Precision Score	0.7362
Linear Discriminant Analysis Recall Score	0.7362
Linear Discriminant Analysis F1 Score	0.7362
Logistic Regression Score	0.9000
Logistic Regression Precision Score	0.6536
Logistic Regression Recall Score	0.7715
Logistic Regression F1 Score	0.6900
Decision Tree Classifier Score	0.8538
Decision Tree Classifier Precision Score	0.6855
Decision Tree Classifier Recall Score	0.6587
Decision Tree Classifier F1 Score	0.6703
Voting Score	0.8846
Voting Precision Score	0.6739
Voting Recall Score	0.7119
Voting F1 Score	0.6900
Random Forest Classifier Score	0.9000
Random Forest Classifier Precision Score	0.6536
Random Forest Classifier Recall Score	0.7715
Random Forest Classifier F1 Score	0.6900
Gradient Boosting Classifier Score	0.9000
Gradient Boosting Classifier Precision Score	0.7406
Gradient Boosting Classifier Recall Score	0.7555
Gradient Boosting Classifier F1 Score	0.7477
Ada Boost Classifier Score	0.9000
Ada Boost Classifier Precision Score	0.7986
Ada Boost Classifier Recall Score	0.7555
Ada Boost Classifier F1 Score	0.7744

Meilleurs résultats généraux après l'inclusion de la variable G2. Or, la variable G2 a une valeur significative pour la qualité du modèle c'est-à-dire plus de relation avec la variable cible G3

Meilleures performances pour les notes de Mathématiques.

Meilleures performances Score F1 :

Maths		Portugais	
Gradient Boost	0.7992	Ada Boost	0.7744
Ada Boost	0.7923	Gradient Boost	0.7477

Classification binaire sur la variable "finalResult"

32 variables Maths

KNeighbors Score	0.6962
KNeighbors Precision Score	0.6152
KNeighbors Recall Score	0.6946
KNeighbors F1 Score	0.6108
Support Vector Machine Score	0.9241
Support Vector Machine Precision Score	0.9183
Support Vector Machine Recall Score	0.9183
Support Vector Machine F1 Score	0.9183
Linear Discriminant Analysis Score	0.8734
Linear Discriminant Analysis Precision Score	0.8638
Linear Discriminant Analysis Recall Score	0.8638
Linear Discriminant Analysis F1 Score	0.8638
Logistic Regression Score	0.8734
Logistic Regression Precision Score	0.8566
Logistic Regression Recall Score	0.8682
Logistic Regression F1 Score	0.8617
Decision Tree Classifier Score	0.8861
Decision Tree Classifier Precision Score	0.8810
Decision Tree Classifier Recall Score	0.8759
Decision Tree Classifier F1 Score	0.8783
Voting Score	0.8987
Voting Precision Score	0.8983
Voting Recall Score	0.8881
Voting F1 Score	0.8925
Random Forest Classifier Score	0.8734
Random Forest Classifier Precision Score	0.8566
Random Forest Classifier Recall Score	0.8682
Random Forest Classifier F1 Score	0.8617
Gradient Boosting Classifier Score	0.8861
Gradient Boosting Classifier Precision Score	0.8810
Gradient Boosting Classifier Recall Score	0.8759
Gradient Boosting Classifier F1 Score	0.8783
Ada Boost Classifier Score	0.8608
Ada Boost Classifier Precision Score	0.8538
Ada Boost Classifier Recall Score	0.8490
Ada Boost Classifier F1 Score	0.8512

32 variables Portugais

KNeighbors Score	0.8000
KNeighbors Precision Score	0.5562
KNeighbors Recall Score	0.6354
KNeighbors F1 Score	0.5601
Support Vector Machine Score	0.8846
Support Vector Machine Precision Score	0.7457
Support Vector Machine Recall Score	0.8536
Support Vector Machine F1 Score	0.7828
Linear Discriminant Analysis Score	0.8923
Linear Discriminant Analysis Precision Score	0.7505
Linear Discriminant Analysis Recall Score	0.8812
Linear Discriminant Analysis F1 Score	0.7932
Logistic Regression Score	0.8769
Logistic Regression Precision Score	0.7105
Logistic Regression Recall Score	0.8632
Logistic Regression F1 Score	0.7534
Decision Tree Classifier Score	0.9077
Decision Tree Classifier Precision Score	0.8057
Decision Tree Classifier Recall Score	0.8805
Decision Tree Classifier F1 Score	0.8359
Voting Score	0.8923
Voting Precision Score	0.7657
Voting Recall Score	0.8631
Voting F1 Score	0.8012
Random Forest Classifier Score	0.8769
Random Forest Classifier Precision Score	0.7105
Random Forest Classifier Recall Score	0.8632
Random Forest Classifier F1 Score	0.7534
Gradient Boosting Classifier Score	0.9385
Gradient Boosting Classifier Precision Score	0.8552
Gradient Boosting Classifier Recall Score	0.9422
Gradient Boosting Classifier F1 Score	0.8906
Ada Boost Classifier Score	0.9538
Ada Boost Classifier Precision Score	0.8952
Ada Boost Classifier Recall Score	0.9533
Ada Boost Classifier F1 Score	0.9208

Encore meilleurs résultats les variables G1 et G2 ont un fort rapport avec la variable G3, comme il était visible avec les graphiques de corrélation.

Meilleures performances pour les notes de Mathématiques.

Meilleures performances Score F1 :

Maths		Portugais	
Support Vector Machine	0.9183	Ada Boost	0.9208
Voting Classifier	0.8925	Gradient Boost	0.8906

Multivariate classification on the variable "academicGrade" – 30

30 variables Maths

KNeighbors Score	0.2785
KNeighbors Precision Score	0.1817
KNeighbors Recall Score	0.1751
KNeighbors F1 Score	0.1700
Support Vector Machine Score	0.3291
Support Vector Machine Precision Score	0.2336
Support Vector Machine Recall Score	0.2270
Support Vector Machine F1 Score	0.1520
Linear Discriminant Analysis Score	0.3165
Linear Discriminant Analysis Precision Score	0.2136
Linear Discriminant Analysis Recall Score	0.1938
Linear Discriminant Analysis F1 Score	0.1933
Logistic Regression Score	0.3418
Logistic Regression Precision Score	0.2336
Logistic Regression Recall Score	0.2085
Logistic Regression F1 Score	0.1957
Decision Tree Classifier Score	0.2532
Decision Tree Classifier Precision Score	0.1817
Decision Tree Classifier Recall Score	0.1738
Decision Tree Classifier F1 Score	0.1742
Voting Score	0.3291
Voting Precision Score	0.2083
Voting Recall Score	0.1733
Voting F1 Score	0.1608
Random Forest Classifier Score	0.3418
Random Forest Classifier Precision Score	0.2336
Random Forest Classifier Recall Score	0.2085
Random Forest Classifier F1 Score	0.1957
Gradient Boosting Classifier Score	0.3038
Gradient Boosting Classifier Precision Score	0.2186
Gradient Boosting Classifier Recall Score	0.1898
Gradient Boosting Classifier F1 Score	0.1942
Ada Boost Classifier Score	0.3418
Ada Boost Classifier Precision Score	0.2269
Ada Boost Classifier Recall Score	0.2141
Ada Boost Classifier F1 Score	0.2144

30 variables Portugais

KNeighbors Score	0.2385
KNeighbors Precision Score	0.2274
KNeighbors Recall Score	0.2437
KNeighbors F1 Score	0.2288
Support Vector Machine Score	0.2769
Support Vector Machine Precision Score	0.2096
Support Vector Machine Recall Score	0.1038
Support Vector Machine F1 Score	0.1384
Linear Discriminant Analysis Score	0.3615
Linear Discriminant Analysis Precision Score	0.3351
Linear Discriminant Analysis Recall Score	0.4327
Linear Discriminant Analysis F1 Score	0.3506
Logistic Regression Score	0.3154
Logistic Regression Precision Score	0.2951
Logistic Regression Recall Score	0.3519
Logistic Regression F1 Score	0.3005
Decision Tree Classifier Score	0.3077
Decision Tree Classifier Precision Score	0.2927
Decision Tree Classifier Recall Score	0.3351
Decision Tree Classifier F1 Score	0.3060
Voting Score	0.2615
Voting Precision Score	0.2246
Voting Recall Score	0.3956
Voting F1 Score	0.2105
Random Forest Classifier Score	0.3154
Random Forest Classifier Precision Score	0.2951
Random Forest Classifier Recall Score	0.3519
Random Forest Classifier F1 Score	0.3005
Gradient Boosting Classifier Score	0.3385
Gradient Boosting Classifier Precision Score	0.3085
Gradient Boosting Classifier Recall Score	0.3155
Gradient Boosting Classifier F1 Score	0.2966
Ada Boost Classifier Score	0.3462
Ada Boost Classifier Precision Score	0.3108
Ada Boost Classifier Recall Score	0.3413
Ada Boost Classifier F1 Score	0.2880

Résultats très bas, la classification binaire résulte bien plus effective pour cette problématique.

Meilleures performances pour les notes de Portugais.

Meilleures performances Score F1 :

Maths		Portugais	
Ada Boost	0.2144	Linear Discriminant A.	0.3506
Random Forest	0.1957	Decision Tree Classifier	0.3060

Multivariate classification on the variable "academicGrade" - 31

31 variables Maths

KNeighbors Score	0.3291
KNeighbors Precision Score	0.2483
KNeighbors Recall Score	0.2516
KNeighbors F1 Score	0.2401
Support Vector Machine Score	0.4304
Support Vector Machine Precision Score	0.3269
Support Vector Machine Recall Score	0.3382
Support Vector Machine F1 Score	0.3230
Linear Discriminant Analysis Score	0.4557
Linear Discriminant Analysis Precision Score	0.4243
Linear Discriminant Analysis Recall Score	0.4042
Linear Discriminant Analysis F1 Score	0.3840
Logistic Regression Score	0.4051
Logistic Regression Precision Score	0.3507
Logistic Regression Recall Score	0.3098
Logistic Regression F1 Score	0.2999
Decision Tree Classifier Score	0.5316
Decision Tree Classifier Precision Score	0.4693
Decision Tree Classifier Recall Score	0.4171
Decision Tree Classifier F1 Score	0.4215
Voting Score	0.5316
Voting Precision Score	0.4710
Voting Recall Score	0.4158
Voting F1 Score	0.4211
Random Forest Classifier Score	0.4051
Random Forest Classifier Precision Score	0.3507
Random Forest Classifier Recall Score	0.3098
Random Forest Classifier F1 Score	0.2999
Gradient Boosting Classifier Score	0.4557
Gradient Boosting Classifier Precision Score	0.4224
Gradient Boosting Classifier Recall Score	0.4171
Gradient Boosting Classifier F1 Score	0.3968
Ada Boost Classifier Score	0.4937
Ada Boost Classifier Precision Score	0.4374
Ada Boost Classifier Recall Score	0.4183
Ada Boost Classifier F1 Score	0.4038

31 variables Portugais

KNeighbors Score	0.2846
KNeighbors Precision Score	0.2768
KNeighbors Recall Score	0.3201
KNeighbors F1 Score	0.2861
Support Vector Machine Score	0.4692
Support Vector Machine Precision Score	0.4541
Support Vector Machine Recall Score	0.5049
Support Vector Machine F1 Score	0.4690
Linear Discriminant Analysis Score	0.5538
Linear Discriminant Analysis Precision Score	0.5473
Linear Discriminant Analysis Recall Score	0.5907
Linear Discriminant Analysis F1 Score	0.5604
Logistic Regression Score	0.5000
Logistic Regression Precision Score	0.4726
Logistic Regression Recall Score	0.5514
Logistic Regression F1 Score	0.4842
Decision Tree Classifier Score	0.4769
Decision Tree Classifier Precision Score	0.4705
Decision Tree Classifier Recall Score	0.4903
Decision Tree Classifier F1 Score	0.4753
Voting Score	0.5000
Voting Precision Score	0.4871
Voting Recall Score	0.5375
Voting F1 Score	0.5011
Random Forest Classifier Score	0.5000
Random Forest Classifier Precision Score	0.4726
Random Forest Classifier Recall Score	0.5514
Random Forest Classifier F1 Score	0.4842
Gradient Boosting Classifier Score	0.6077
Gradient Boosting Classifier Precision Score	0.6160
Gradient Boosting Classifier Recall Score	0.6232
Gradient Boosting Classifier F1 Score	0.6148
Ada Boost Classifier Score	0.6154
Ada Boost Classifier Precision Score	0.5918
Ada Boost Classifier Recall Score	0.6208
Ada Boost Classifier F1 Score	0.5929

Meilleurs résultats avec G1. Toujours meilleurs résultats pour la classification binaire.

Meilleures performances pour les notes de Portugais.

Meilleures performances Score F1 :

Maths		Portugais	
Decision Tree Classifier	0.4215	Gradient Boost	0.6148
Voting Classifier	0.4211	Ada Boost	0.5929

Multivariate classification on the variable "academicGrade"-32

32 variables Maths

KNeighbors Score	0.3671
KNeighbors Precision Score	0.2717
KNeighbors Recall Score	0.2641
KNeighbors F1 Score	0.2610
Support Vector Machine Score	0.4937
Support Vector Machine Precision Score	0.3855
Support Vector Machine Recall Score	0.3862
Support Vector Machine F1 Score	0.3749
Linear Discriminant Analysis Score	0.5823
Linear Discriminant Analysis Precision Score	0.5260
Linear Discriminant Analysis Recall Score	0.5248
Linear Discriminant Analysis F1 Score	0.5138
Logistic Regression Score	0.4684
Logistic Regression Precision Score	0.3738
Logistic Regression Recall Score	0.3815
Logistic Regression F1 Score	0.3488
Decision Tree Classifier Score	0.5696
Decision Tree Classifier Precision Score	0.4752
Decision Tree Classifier Recall Score	0.4684
Decision Tree Classifier F1 Score	0.4677
Voting Score	0.5443
Voting Precision Score	0.4102
Voting Recall Score	0.4164
Voting F1 Score	0.4068
Random Forest Classifier Score	0.4684
Random Forest Classifier Precision Score	0.3738
Random Forest Classifier Recall Score	0.3815
Random Forest Classifier F1 Score	0.3488
Gradient Boosting Classifier Score	0.7089
Gradient Boosting Classifier Precision Score	0.6726
Gradient Boosting Classifier Recall Score	0.6659
Gradient Boosting Classifier F1 Score	0.6667
Ada Boost Classifier Score	0.7089
Ada Boost Classifier Precision Score	0.6583
Ada Boost Classifier Recall Score	0.5547
Ada Boost Classifier F1 Score	0.5896

32 variables Portugais

KNeighbors Score	0.3385
KNeighbors Precision Score	0.3315
KNeighbors Recall Score	0.3821
KNeighbors F1 Score	0.3424
Support Vector Machine Score	0.6462
Support Vector Machine Precision Score	0.6209
Support Vector Machine Recall Score	0.6686
Support Vector Machine F1 Score	0.6310
Linear Discriminant Analysis Score	0.7615
Linear Discriminant Analysis Precision Score	0.7496
Linear Discriminant Analysis Recall Score	0.7864
Linear Discriminant Analysis F1 Score	0.7539
Logistic Regression Score	0.5769
Logistic Regression Precision Score	0.5453
Logistic Regression Recall Score	0.6187
Logistic Regression F1 Score	0.5566
Decision Tree Classifier Score	0.7077
Decision Tree Classifier Precision Score	0.7212
Decision Tree Classifier Recall Score	0.7185
Decision Tree Classifier F1 Score	0.7117
Voting Score	0.7000
Voting Precision Score	0.7051
Voting Recall Score	0.7164
Voting F1 Score	0.7026
Random Forest Classifier Score	0.5769
Random Forest Classifier Precision Score	0.5453
Random Forest Classifier Recall Score	0.6187
Random Forest Classifier F1 Score	0.5566
Gradient Boosting Classifier Score	0.7769
Gradient Boosting Classifier Precision Score	0.7580
Gradient Boosting Classifier Recall Score	0.7980
Gradient Boosting Classifier F1 Score	0.7635
Ada Boost Classifier Score	0.7000
Ada Boost Classifier Precision Score	0.6855
Ada Boost Classifier Recall Score	0.7560
Ada Boost Classifier F1 Score	0.6747

Toujours meilleurs résultats, les variables G1 et G2 ont un fort rapport avec la variable G3,
Globalement une plus basse performance pour la classification multiple par rapport à la binaire.

Meilleures performances pour les notes de Portugais.

Meilleures performances Score F1 :

Maths		Portugais	
Gradient Boost	0.6667	Gradient Boost	0.7635
Ada Boost	0.5896	Linear Discriminant A.	0.7539

Régression sur la variable "G3" - 31

30 variables Maths

Linear Regression R2 Score	-0.2850
Linear Regression Mean Squared Error Score	17.5682
KNeighbors Regressor R2 Score	0.1219
KNeighbors Regressor Mean Squared Error Score	12.0052
Decision Tree Regressor R2 Score	-0.3073
Decision Tree Regressor Mean Squared Error Score	17.8734

30 variables Portugais

Linear Regression R2 Score	0.2257
Linear Regression Mean Squared Error Score	10.4003
KNeighbors Regressor R2 Score	0.1965
KNeighbors Regressor Mean Squared Error Score	10.7919
Decision Tree Regressor R2 Score	-0.2176
Decision Tree Regressor Mean Squared Error Score	16.3538

Des mauvais résultats, les modèles ne trouvent pas un rapport important parmi les variables analysées et G3.
Les variables analysées semblent être plus en rapport avec les résultats en portugais qu'en mathématiques.

Meilleures performances :

Maths		Portugais	
KNeighbors R2	0.1219	Linear Regression R2	0.2257
KNeighbors squared e.	12.0052	Linear R. squared error.	10.4003

Régression sur la variable "G3" - 32

31 variables Maths

Linear Regression R2 Score	0.5917
Linear Regression Mean Squared Error Score	5.5824
KNeighbors Regressor R2 Score	0.2564
KNeighbors Regressor Mean Squared Error Score	10.1670
Decision Tree Regressor R2 Score	0.2936
Decision Tree Regressor Mean Squared Error Score	9.6582

31 variables Portugais

Linear Regression R2 Score	0.6370
Linear Regression Mean Squared Error Score	4.8750
KNeighbors Regressor R2 Score	0.2819
KNeighbors Regressor Mean Squared Error Score	9.6447
Decision Tree Regressor R2 Score	0.5138
Decision Tree Regressor Mean Squared Error Score	6.5308

Meilleurs résultats, les variables G1,

Globalement une plus basse performance pour la classification multiple par rapport au binaire.

Meilleures performances pour les notes de Portugais.

Meilleures performances Score F1 :

Maths		Portugais	
Linear Regression R2	0.5917	Linear Regression R2	0.6370
Linear R squared e.	5.5824	Linear R squared e.	4.8750

Régression sur la variable "G3" - 32

32 variables Maths

Linear Regression R2 Score	0.7391
Linear Regression Mean Squared Error Score	3.5676
KNeighbors Regressor R2 Score	0.3461
KNeighbors Regressor Mean Squared Error Score	8.9394
Decision Tree Regressor R2 Score	0.6926
Decision Tree Regressor Mean Squared Error Score	4.2025

32 variables Portugais

Linear Regression R2 Score	0.8571
Linear Regression Mean Squared Error Score	1.9189
KNeighbors Regressor R2 Score	0.3342
KNeighbors Regressor Mean Squared Error Score	8.9420
Decision Tree Regressor R2 Score	0.7245
Decision Tree Regressor Mean Squared Error Score	3.7000

Enorme amélioration avec les variables G1 et G2 fait que réaffirme encore une fois l'importance du rapport entre ces deux variables et G3,

Globalement une plus basse performance pour la classification multiple par rapport à la binaire.

Meilleures performances pour les notes de Portugais.

Meilleures performances Score F1 :

Maths		Portugais	
Linear Regression R2	0.7391	Linear Regression R2	0.8571
Linear R squared e.	3.5676	Linear R squared e.	1.9189

Conclusion.

Après les études menés on peut en tirer quelques conclusions :

1. Même si les premiers 30 variables peuvent avoir de la valeur, à cause de la malédiction de la dimensionnalité, elles auront tendance à empêcher la bonne performance du modèle.
2. Les résultats scolaires dans le passé (G1, G2) ont une grande corrélation avec les résultats actuelles ou futures (G3).
3. Les variables analysées ont une plus grande influence pour les résultats de la matière Langue portugaise, et par conséquent les résultats de classification et régression, pour le Data Set concernant ces résultats, sont de manière générale plus précis.
4. Pour les Data Sets étudiées les algorithmes les plus performantes et alors ceux qui modèlent mieux les données ont été les suivants : Linear Regression pour la régression et Gradient Boost et Ada Boost pour la classification