Apprentissage Statistique - TP3

Performance scolaire : cas du Portugal

Jacques Aguilera
Youssouf Camara
Zakarya Ali

29 Janvier 2018

Table des matières

Introduction

L'article 26 de la Déclaration universelle des droits de l'homme dit que « toute personne a droit à l'éducation. L'éducation doit être gratuite, au moins en ce qui concerne l'enseignement élémentaire et fondamental. Elle doit viser au plein épanouissement de la personnalité humaine». Cependant, au-delà de cette gratuité et ce caractère obligatoire de l'éducation, un problème plus profond se cache. En effet, même si l'éducation est un facteur clé du progrès économique, il n'en demeure pas moins que les premiers concernés n'arrivent pas à valider certains acquis fondamentaux pour la poursuite des études. Parmi ces acquis se trouvent des matières telles que les mathématiques et la langue nationale (Portugais dans notre cas). Pour comprendre la disparité entre les élèves qui réussissent et ceux qui échouent, nous nous inspirerons de l'article de Paulo Cortez et Alice Silva (2008) et nous utiliserons leur base de données pour prédire les performances d'élèves portugais.

"Using Data Mining To Predict Secondary School Student Performance", publié en 2008, décrit comment l'application de méthodes d'apprentissage statistique peut permettre d'évaluer la performance scolaire. Grâce à des données récoltées sur des élèves du second degré au Portugal, l'article compare la précision prédictive de différents algorithmes (Arbre de décision (DT), forêt aléatoire (RF), réseau de neurones, et support vector machine (SVM)) face à des prédicteurs dits naïfs, en terme de classification et de régression. Dans notre compte-rendu, nous allons essayer de reproduire ces résultats, avec les méthodes décision tree, random forest et SVM.

En utilisant les techniques précédemment citées, notre objectif est donc de prédire la réussite des élèves et, si possible, d'identifier les variables clés qui affectent le succès ou l'échec scolaire. Puisque le manque de succès dans les classes de base de mathématiques et la langue portugaise est extrêmement grave, les deux principaux cours (Mathématiques et Portugais) seront modélisés sous trois objectifs i) classification binaire (réussite (pass) / échec (fail) ; ii) classification avec cinq niveaux (de I très bon ou excellent à V - insuffisant) ; et iii) la régression, avec une sortie numérique comprise entre zéro et vingt de moyenne (G3). Pour chacune de ces approches, trois configurations seront étudiées. La configuration A) prise en compte de toutes les variables à notre disposition (y compris les résultats de la première et seconde période scolaire (G1 et G2 respectivement). Configuration B) la prise en compte de toutes les variables excepté le résultat de la seconde période d'étude. Enfin, la configuration C) où l'on ne prendra pas en compte les grades G1 et G2.

Afin de mieux comprendre notre démarche, nous allons dans un premier temps décrire les données et présenter les différentes techniques de modélisation. Ensuite, nous allons présenter les résultats avec les trois techniques et configurations. Enfin, nous allons discuter de la pertinence de ces résultats et faire des propositions d'amélioration de ces derniers.

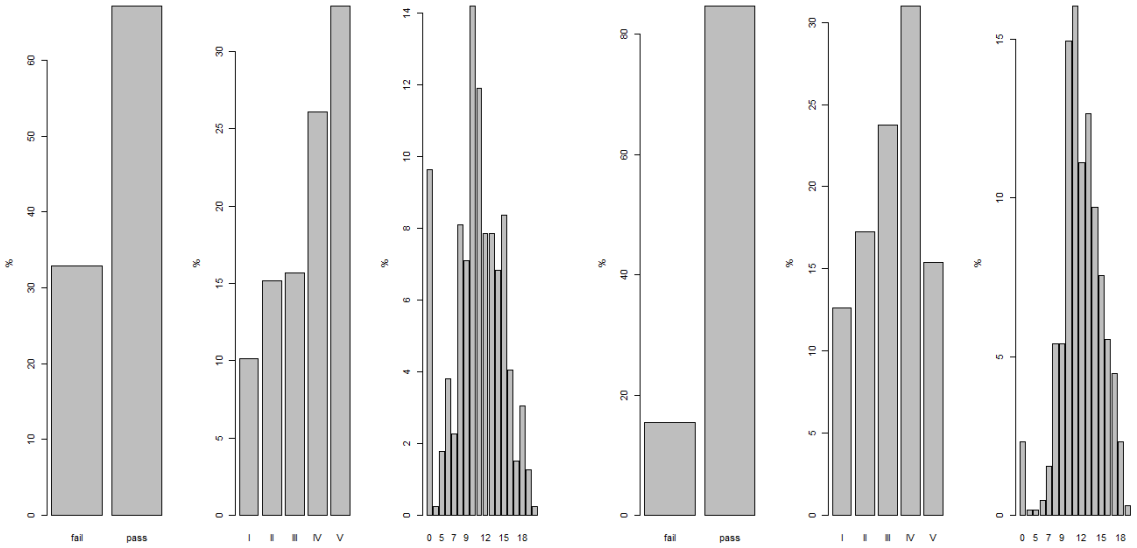
1 Données et Méthodologie

1.1 Description des données

Les données sont disponibles ici :
`https://archive.ics.uci.edu/ml/datasets/Student+Performance`

On a 2 jeux de données : un qui représente les élèves et leurs notes en Mathématiques et l'autre en Portugais. Ces deux datasets ont la même structure. Chaque ligne représente un élève et ses caractéristiques (voir table 4 la liste des caractéristiques). On va évaluer la colonne G3 qui est le résultat de l'individu lors du 3e examen dans la matière considérée. Pour chaque matière, on a quelques centaines d'observations.

Le code qui nous a permis d'obtenir les résultats de la partie 2 est disponible ici : `https://github.com/zakaryaxali/ml_student_performance`



(a) Mathématiques (b) Portugais

FIGURE 1: Répartition des élèves en fonction de leur score

1.2 Techniques d'apprentissage retenues

1.2.1 Decision Tree

Appartenant à la famille des méthodes dites de partitionnement récursif ou de segmentation, les méthodes de construction d'arbres binaires de décision (classification tree) ou de régression (regression tree) sont basés sur une séquence récursive de règles de division, coupes ou splits. Elles sont donc basées sur une hiérarchisation logique en fonction des variables explicatives les plus pertinentes. Cette sélection des variables se fait de manière automatique. Ainsi, seules les variables pertinentes apparaissent lors de la construction de l'arbre. En clair, l'ensemble des observations est regroupé à la racine de l'arbre puis chaque division ou coupe sépare chaque nœud en deux nœuds fils plus homogènes que le nœud père au sens d'un critère à préciser et dépendant du type de la variable Y, quantitative ou qualitative et de l'élagage pour l'obtention d'un modèle parcimonieux.

Toutefois, il faut noter que cet algorithme suit une stratégie pas à pas hiérarchisée. Il peut, comme dans le cas du choix de modèle pas à pas en régression, passer à coté d'un optimum

global ; il se montre par ailleurs très instable et donc sensible à des fluctuations d'échantillon. Cette instabilité ou variance de l'arbre est une conséquence de la structure hiérarchique : une erreur de division en début d'arbre est propagée tout au long de la construction. Elles sont généralement utilisées pour des raisons d'interprétation et de présentation.

Tous nos arbres de décisions ont été construit selon le principe de partition parcimonieuse et avec une particularité mise sur le choix optimale de notre de feuille à l'aide la technique de validation croisée.

1.2.2 Random Forest

L'algorithme du random forest est un algorithme qui vise à être robuste au bruit, dans le sens où de faibles changements dans les données d'entraînement auront peu ou pas d'impact dans les résultats de la modélisation. Les random forests sont performants pour les tâches de classification où une classe est très sous-représentée. Le principe de cet algorithme est d'agréger les résultats obtenus sur un ensemble d'arbres de décision générés indépendamment à partir d'un échantillon aléatoire tiré dans les données d'entrainement. Chaque arbre est divisé jusqu'à sa profondeur maximale et sans qu'aucune opération d'élagage ne soit réalisée. L'aléa est présent dans la sélection des observations et des variables, ce qui le rend robuste au bruit, aux outliers et au surapprentissage comparé à un arbre simple. A chaque noeud, seule une petite partie des variables sont disponibles. Les random forests sont aussi performants lorsqu'il y a un grand nombre de variables et relativement peu d'observations. Il y a en principe peu de travail de préparation des données, puisqu'il n'est pas nécessaire de normaliser les données, de sélectionner des variables et l'algorithme est robuste aux outliers.

Dans le cas de la classification, l'agrégation des résultats est obtenue par l'association de règles majoritaires et de scores pondérés par la qualité, i.e l'accuracy, des arbres individuels. Dans le cas de la régression, on prend la moyenne sur l'ensemble des arbres.

1.2.3 Support Vector Machine

On appelle Support Vector Machine, un modèle capable de :

- Rechercher l'hyperplan optimal capable de séparer 2 classes en maximisant la distance entre les points les plus proches de ces classes (voir figure 2).Les points se trouvant sur la frontière ainsi obtenue se nomment les support vectors.
- Dans le cas où des points seraient du mauvais coté de la frontière, ajuster leurs poids pour réduire leur influence.
- Lorsqu'on ne trouve pas de séparateur linéaire, projetter les points dans un espace de plus grande dimension où ceux-ci deviennent alors linéairement séparables (kernel trick).

Des trois algorithmes avec lesquels nous effectuons l'évaluation, le SVM est le moins interprétable dans la mesure où on ne peut pas expliquer quels paramètres ont le plus d'importance pour créer la frontière.

Cas de la classification en 5 parties

Le SVM ne peut pas séparer plus de 2 catégories en même temps. Lorsque le nombre de

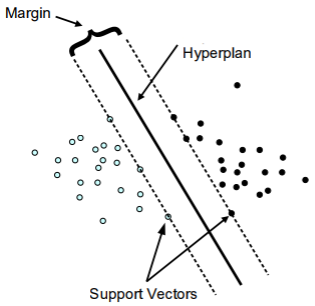


FIGURE 2: Principe du SVM

catégories à classer est supérieur à 2, on applique alors la méthode dite du "One-versus-All". Soit $k > 2$, le nombre de clusters à segmenter. On applique k SVM qui vont séparer à chaque fois une classe et toutes les autres données (voir figure 3).

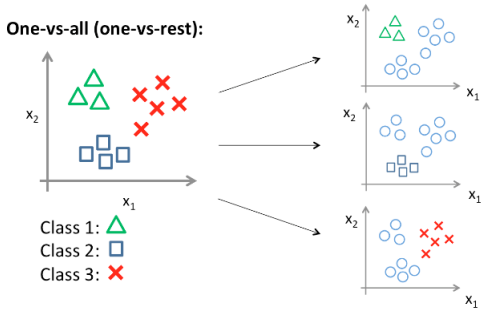


FIGURE 3: SVM - One-Versus-All

Pour la régression, on étend ce principe aux 21 notes possibles.

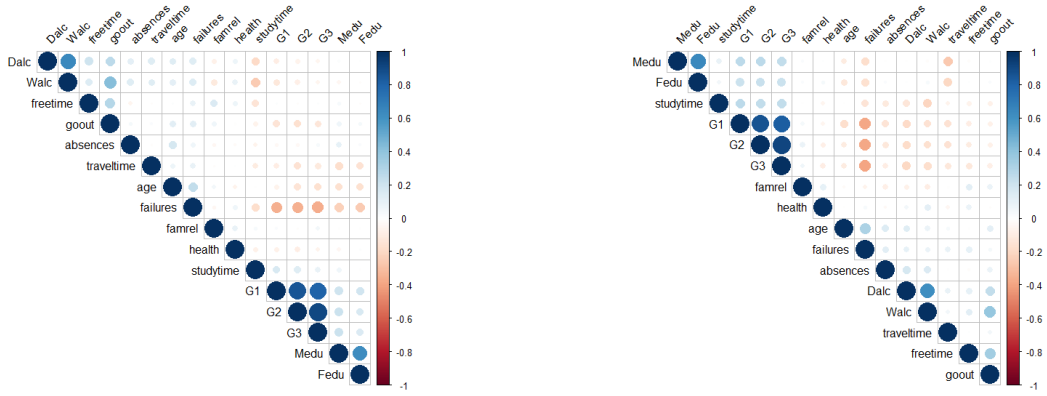
2 Résultats

2.1 Statistiques descriptives

2.1.1 Correlation

Les graphiques (figure 4) de corrélation entre les variables nous apportent des informations très intéressantes et viennent ainsi réconforter notre stratégie de modélisation. En effet, ils nous renseignent que notre variable d'intérêt et les caractéristiques inhérentes aux élèves sont faiblement corrélées. Les corrélations les plus élevées sont celles entre la variable d'intérêt G3 et G2 et G1. Cela indique clairement que seule ces deux variables sont pertinentes pour expliquer la variable à prédire. Cela ne nous surprend point puisque réussir l'année précédente montre la capacité de l'élève à avoir de bonnes moyennes l'année suivante, ce d'autant qu'il existe une corrélation très élevée entre G1 et G2. Toutefois, il faut noter qu'il existe une corrélation négative et significative entre l'échec (failure) et G3, ce qui ne nous étonne guère pour les mêmes raisons précédemment évoquées. Ainsi, en regardant les arbres

de décision, force est de constater que les variables G2 et G1 sont les plus pertinentes pour expliquer la performance des élèves.



(a) Mathématiques (b) Portugais

FIGURE 4: Corrélation des données

2.1.2 Importance des variables

Nous présentons dans la figure 5 la fonctionnalité Importance fourni par la méthode random forest. Nous voyons dans les graphiques, les variables classées par ordre décroissant selon l'indice de Gini.

2.1.3 Naive Predictors

Pour évaluer la qualité de nos prédicteurs, un outil intéressant est de les comparer à des prédicteurs dits "naïfs" dans le sens où ils nécessitent l'application d'éléments rudimentaires pour être utilisés. Ils sont différents selon la méthode et le type de prédiction :

- Communs à la prédiction et la classification :
  - Méthode A : on utilise G2 (exemple du cas binaire : si l'élève a eu plus de 10 au 2e examen, alors on prédit qu'il aura plus de 10 à l'examen 3)
  - Méthode B : on utilise G1
- Pour les deux types de classifications
  - Méthode C : on utilise la catégorie la plus importante (si la majorité des élèves ont plus de 10 à l'examen, on choisit comme prédicteur tous les élèves ont plus de 10 à l'examen)
- Pour la régression :
  - Méthode C : on utilise la moyenne des résultats des élèves

On ajoute les résultats de ces prédicteurs à nos table de score. On les note **NV**.

2.2 Prédiction

Pour les différents algorithmes, on applique le même protocole :

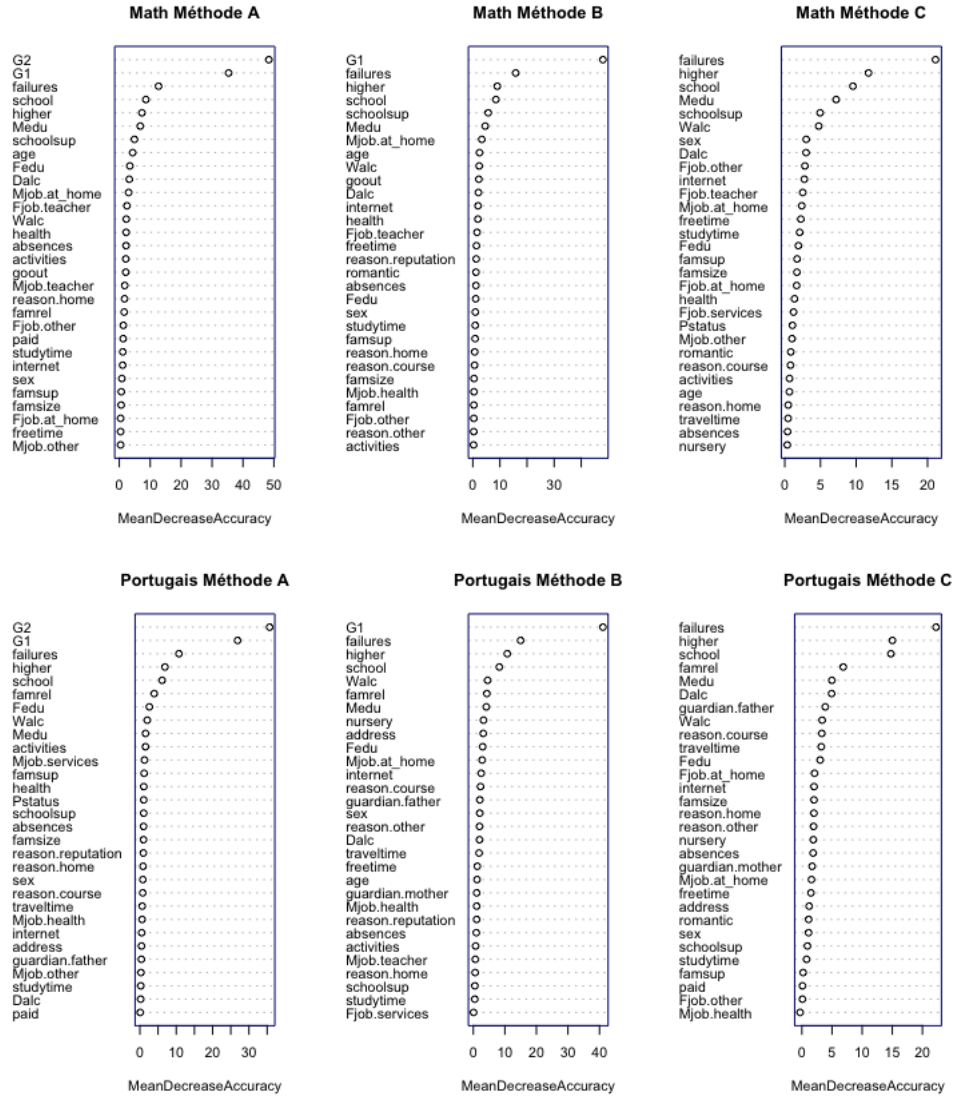


FIGURE 5: Importance des variables dans l'évaluation de la performance

— On entraine sur 70% du jeu de données récupéré.
— On effectue un 10-fold cross validation.
— On obtient nos résultats en appliquant nos modèles aux 30% de données restantes (données de tests).

Pour les algorithmes on utilise des librairies différentes :

— SVM : **caret**
— RF : **randomForest**
— DT : **rpart**

Voir code en annexe.

2.2.1 Classification

Voici les scores obtenus pour les classifications. On marque en gras le meilleur prédicteur dans chaque cas.

	Mathématiques				Portugais			
Méthode	NV	DT	RF	SVM	NV	DT	RF	SVM
A	91.9	**93.26**	90.68	89.83	89.7	**93.61**	91.75	88.14
B	**83.8**	81.90	79.66	77.12	87.5	**90.76**	89.69	86.08
C	67.1	65.52	**71.19**	66.10	84.6	**87.29**	86.08	83.51

TABLE 1: Précision des algorithmes pour la classification binaire

Le premier constat avec le tableau 1 qui présente les scores de la classification binaire, c'est que le SVM est pour chaque cas le moins performant.

Après avoir effectué différentes modélisations, les arbres de décisions semblent fournir des résultats sensiblement supérieurs. Comme nous avons pu le constater à l'aide de la matrice de corrélation, les graphiques fournis par les arbres de décisions confirment notre intuition. Enfin, lorsqu'on choisit la configuration A), force est de constater que seule la variable G2 détermine le mode de classification de la variable cible. Il détermine ainsi en fonction de certaines valeurs (supérieurs ou inférieurs à 9.5) si l'élève passe ou s'il échoue en mathématique ou en portugais. En son absence, la variable G1 prend le relai et ce dernier est chevauché par la variable failure qui détermine de le mode de segmentions. Les arbres permettent ainsi de présenter de manière itérative et intéressante le classement des élèves en fonction de leur résultat précédent.

2.2.2 Régression

Pour les score de prédiction dans le cas de la régression, on choisit d'utiliser l'erreur quadratiques. nos résultats sont présentés dans la table 3.

Les résultats en mathématiques étaient plus compliqués à prédire qu'en portugais. Dans le cas des Maths, le RF s'en sort le mieux dans tous les cas. Il est également les meilleur prédicteur dans les cas A et B pour le portugais. On note également que le SVM est, pour les 2 matières, un bon prédicteur car il fait toujours pratiquement aussi bien, voir mieux que les prédicteurs naïfs. Ca peut paraitre étonnant puisqu'il utilise la méthode du one-versus-all

	Mathématiques				Portugais			
Méthode	NV	DT	RF	SVM	NV	DT	RF	SVM
A	**77.59**	76.17	70.94	65.81	72.9	**74.73**	74.09	63.21
B	**60.5**	55.17	51.28	46.15	58.7	**60.71**	53.88	50.78
C	32.9	31.03	**36.75**	25.64	31.0	35.21	**35.23**	33.16

TABLE 2: Précision des algorithmes pour la classification à 5 classes

Pour la classification à 5 classes (tableau 2) on fait également le constat de la faiblesse du SVM. pour les méthodes A et B, les prédicteurs naïfs sont toujours les meilleurs ou alors meilleurs du meilleur prédicteur. En revanche pour la méthode C, seulement retourner la classe la plus représenté est un peu léger, et nos meilleurs prédicteurs sont sensiblement plus performant.

	Mathématiques				Portugais			
Méthode	NV	DT	RF	SVM	NV	DT	RF	SVM
A	2.01	2.08	**1.8**	1.96	**1.32**	2.55	**1.32**	1.34
B	2.80	2.63	**2.45**	2.66	1.89	2.81	**1.8**	1.87
C	4.59	4.35	**3.78**	4.25	3.23	3.08	3.22	**2.81**

TABLE 3: RMSE des algorithmes pour la régression

et donc crée 21 frontières pour segmenter les notes. Pour la méthode C, on observe un écart important entre le meilleur prédicteur et la méthode naïve, on peut donc justifier le choix de se porter sur un algorithme d'apprentissage dans ce cas. Le DF s'en sort moins bien (voir figure 7.

Conclusion

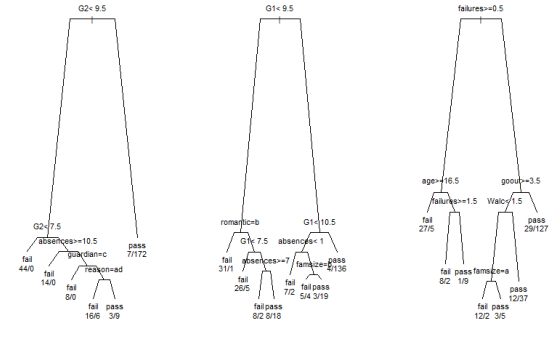
Nos prédicteurs se montrent plus pertinents que les prédicteurs naïfs dans une grande partie des cas. Malgré tout, les écarts avec ces derniers ne sont pas grands et on peut discuter alors de la nécessité de méthodes d'apprentissage statistiques pour rendre compe de la prédiction des performances scolaires. A cela, on peut répondre que les jeux de données n'étaient peut-être pas suffisament grand (centaines d'observations pour chaque matière), ou que les données n'étaient pas assez variées (on récupère les données d'enfants de seulement 2 écoles). Des points techniques peuvent également être soulevés. L'article ne précise par pour le DT et le RF quels hyperparamètres ont été utilisés. Un réglage de ces derniers à l'aide de la méthode du "grid search" pourraient nous permettre d'améliorer la prédiction de ces méthodes.

Un point reste en suspens, comment se servir de ces prédictions pour aider les élèves ? Détecter en amont les élèves suceptibles d'être en difficultés ? Proposer un parcours plus riche aux meilleurs élèves ? ....

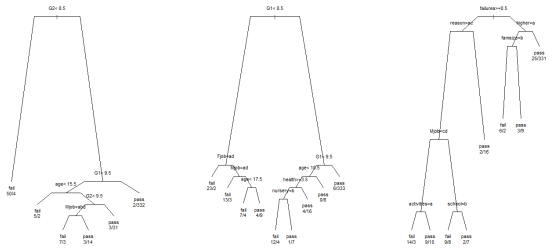
L'interprétation de ces données met en lumière des schémas influençant la réussite scolaire. En effet grâce aux arbres de décision créés (figure 6), on peut, avec précaution, détermi-

ner des motifs qui amènent ou non un élève à réussir en mathématiques et/ou en portugais.

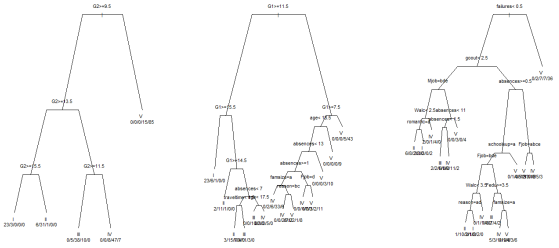
3 Annexe



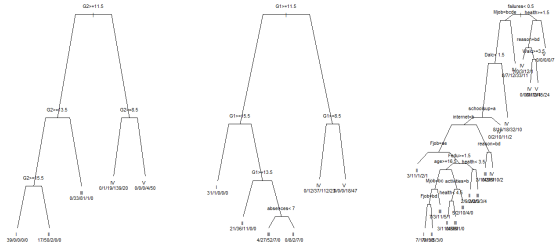
(a) binary/math



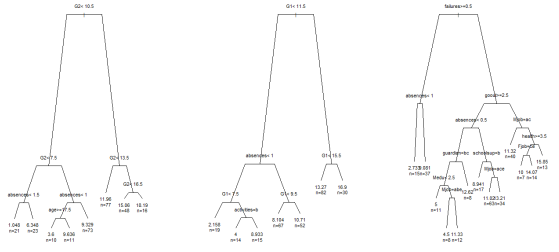
(b) binary/portugais



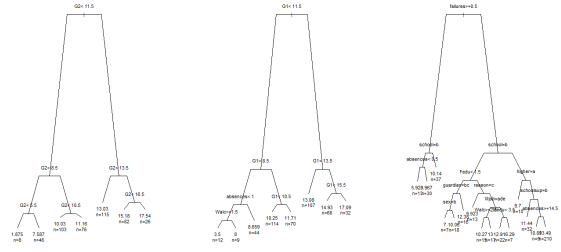
(c) 5-levels/math



(d) 5-levels/portugais



(e) Regressions G3/math



(f) Regressions G3/portugais

FIGURE 6: Arbres de décision pour les méthodes A, B et C

Attribut	Description
school	student's school (binary : 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	student's sex (binary : 'F' - female or 'M' - male)
age	student's age (numeric : from 15 to 22)
address	student's home address type (binary : 'U' - urban or 'R' - rural)
famsize	family size (binary : 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	parent's cohabitation status (binary : 'T' - living together or 'A' - apart)
Medu	mother's education (numeric : 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	father's education (numeric : 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Mjob	mother's job (nominal : 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
Fjob	father's job (nominal : 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at home' or 'other')
reason	reason to choose this school (nominal : close to 'home', school 'reputation', 'course' preference or 'other')
guardian	student's guardian (nominal : 'mother', 'father' or 'other')
traveltime	home to school travel time (numeric : 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric : 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric : n if 1<=n<3, else 4)
schoolsup	extra educational support (binary : yes or no)
famsup	family educational support (binary : yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary : yes or no)
activities	extra-curricular activities (binary : yes or no)
nursery	attended nursery school (binary : yes or no)
higher	wants to take higher education (binary : yes or no)
internet	Internet access at home (binary : yes or no)
romantic	with a romantic relationship (binary : yes or no)
famrel	quality of family relationships (numeric : from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric : from 1 - very low to 5 - very high)
goout	going out with friends (numeric : from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric : from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric : from 1 - very low to 5 - very high)
health	current health status (numeric : from 1 - very bad to 5 - very good)
absences	number of school absences (numeric : from 0 to 93)
these grades are related with the course subject, Math or Portuguese :	
G1	first period grade (numeric : from 0 to 20)
G2	second period grade (numeric : from 0 to 20)
G3	final grade (numeric : from 0 to 20, output target)

TABLE 4: Liste des paramètres des jeux de données

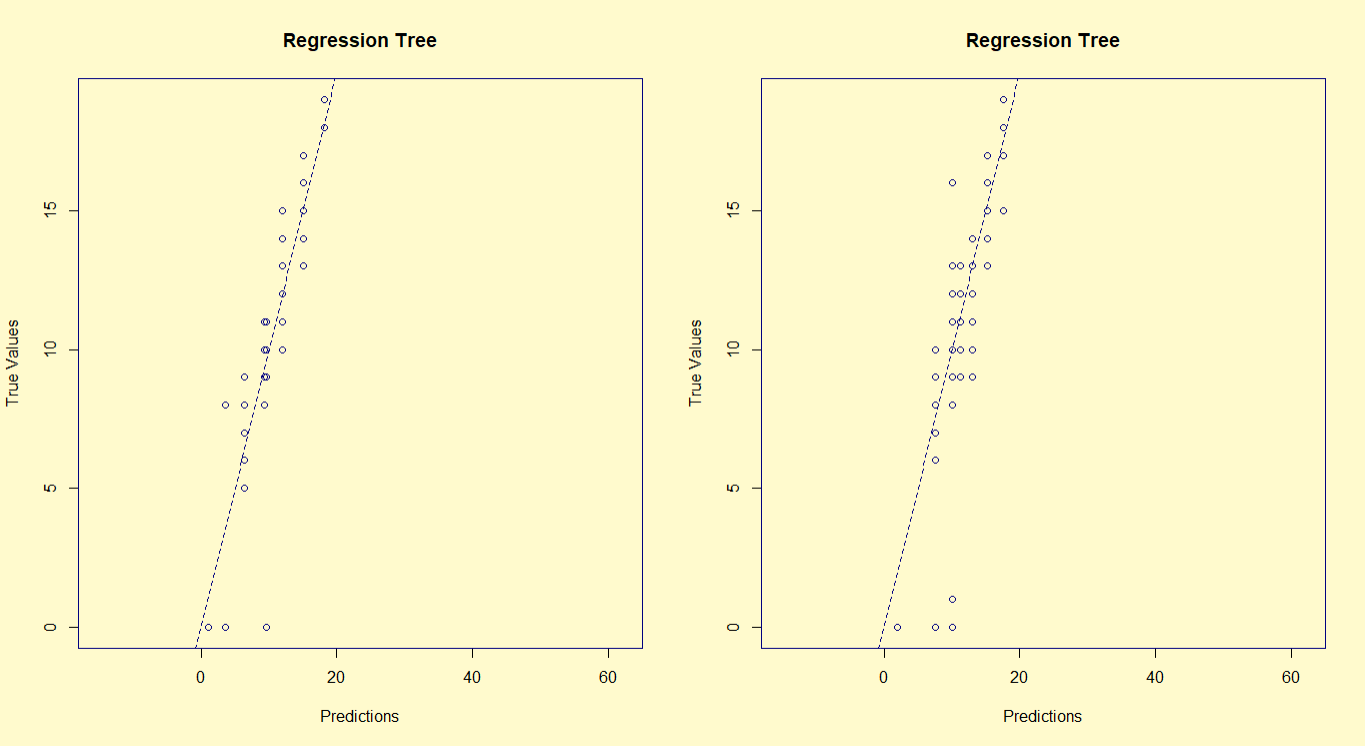


FIGURE 7: Evaluation de la régression pour les arbres de décision - Math (gauche), Portuguais (droite)