Année universitaire 2019-2020

Rapport CheckYourSmile

Tutorial Project

LE Viet Minh Thong, PHAN Dinh Triem

Mai 2020

Institut National des Sciences Appliquées de Toulouse
Département Génie Mathématique et Modélisation

Contents

1 Introduction

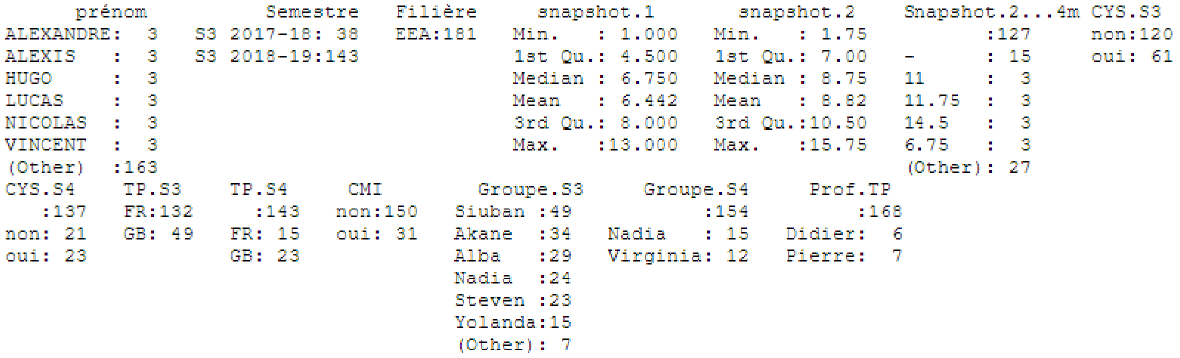
CheckYourSmile (CYS) is a web platform project for learning speciality vocabulary in foreign languages (eg IT English: networks / databases), led by Dr Nadia Yassine-Diab, where users can learn through a set of ”serious” games. The objective is to provide a complement to face-to-face language courses in higher education courses, where few hours can be devoted to teaching speciality vocabulary, despite its importance for the professional integration of students. Note that one of the innovations of CYS is to offer a collaborative system to propose and validate lexical entries: thus, everyone participates in the construction of knowledge (cf. crowdsourcing).

The first CYS prototype was released in 2014 (currently online at www.checkyoursmile.fr). IDEX (Initiative of Excellence) funding from the University of Toulouse in 2016 made it possible to hire several developers, trainees and post-docs to develop the site; a new version was released in January 2017, including new games and new features. The platform is free and licensed under the Creative Commons license. The previous prototypes have already served us to demonstrate the concept and to propose a stable and functional version of the site which now includes 6 games that work on the 4 skills of learning a language (French as a foreign language and English for French).

Our objective is to obtain indicators on the plus-value of Check Your Smile in a university context and on the combinations of variables which make it possible to obtain the best results in order to improve the effects of the determined tool.

2 Materials

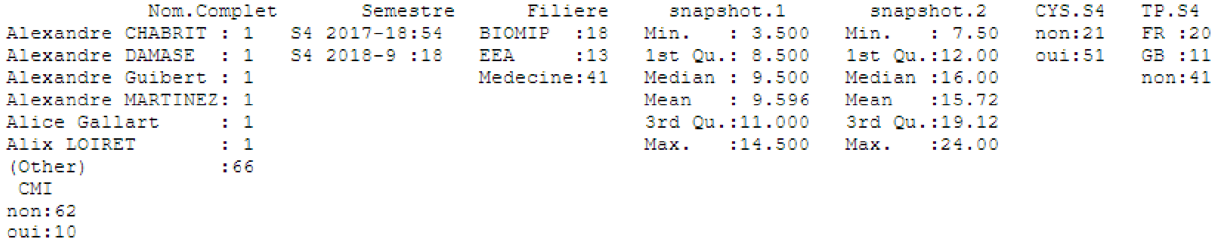
The subject aims to study a database acquired during 3 academic years (2016-7, 2017-8 and 2018-9). It contains the evaluation results of students of different UPS courses as well as details on their courses and the particularities of the received language teaching (English or French TP, CMI engineering courses, use from CYS or not ...)



```
      prénom          Semestre    Filière     snapshot.1       snapshot.2    Snapshot.2...4m CYS.S3
ALEXANDRE:  3   S3 2017-18: 38    EEA:181   Min.   : 1.000   Min.   : 1.75             :127     non:120
ALEXIS   :  3   S3 2018-19:143              1st Qu.: 4.500   1st Qu.: 7.00   -      : 15     oui: 61
HUGO     :  3                               Median : 6.750   Median : 8.75   11     :  3
LUCAS    :  3                               Mean   : 6.442   Mean   : 8.82   11.75  :  3
NICOLAS  :  3                               3rd Qu.: 8.000   3rd Qu.:10.50   14.5   :  3
VINCENT  :  3                               Max.   :13.000   Max.   :15.75   6.75   :  3
(Other)  :163                                                                (Other): 27
CYS.S4    TP.S3    TP.S4     CMI         Groupe.S3     Groupe.S4     Prof.TP
   :137   FR:132     :143   non:150   Siuban :49           :154         :168
non: 21   GB: 49   FR: 15   oui: 31   Akane  :34   Nadia   : 15   Didier:  6
oui: 23            GB: 23             Alba   :29   Virginia: 12   Pierre:  7
                                       Nadia  :24
                                       Steven :23
                                       Yolanda:15
                                       (Other): 7
```

Figure 1: Summary of data semester 3

In fact, in semester 3, we carried out an assessment of 181 students including:
- 38 in 2017-2018 and 143 in 2018-2019
- All 181 in the sector EEA
- 120 used the CYS's tool while 41 did not use it
- 132 had practical works in French while 49 used English
- 160 were in CMI while 31 were not



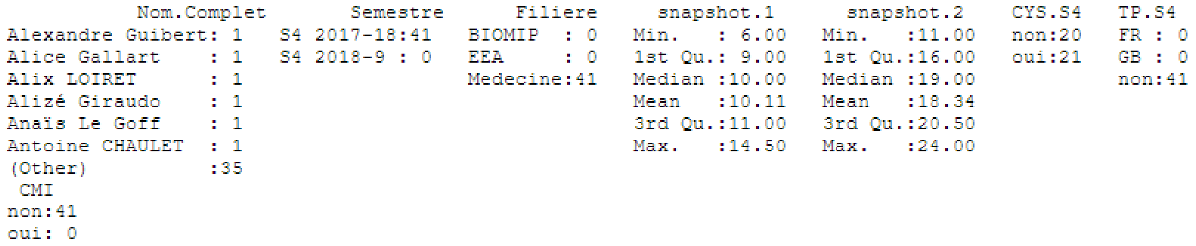
```
            Nom.Complet       Semestre      Filiere     snapshot.1       snapshot.2    CYS.S4   TP.S4
Alexandre CHABRIT : 1   S4 2017-18:54   BIOMIP  :18   Min.   : 3.500   Min.   : 7.50   non:21   FR :20
Alexandre DAMASE  : 1   S4 2018-9 :18   EEA     :13   1st Qu.: 8.500   1st Qu.:12.00   oui:51   GB :11
Alexandre Guibert : 1                   Medecine:41   Median : 9.500   Median :16.00            non:41
Alexandre MARTINEZ: 1                                 Mean   : 9.596   Mean   :15.72
Alice Gallart     : 1                                 3rd Qu.:11.000   3rd Qu.:19.12
Alix LOIRET       : 1                                 Max.   :14.500   Max.   :24.00
(Other)           :66
 CMI
non:62
oui:10
```

Figure 2: Summary of data semester 4

In fact, medecins do not practise TP and they are not in engineering class(CMI). Thus, we separated data into 2 groups: medecin and non-medecin.

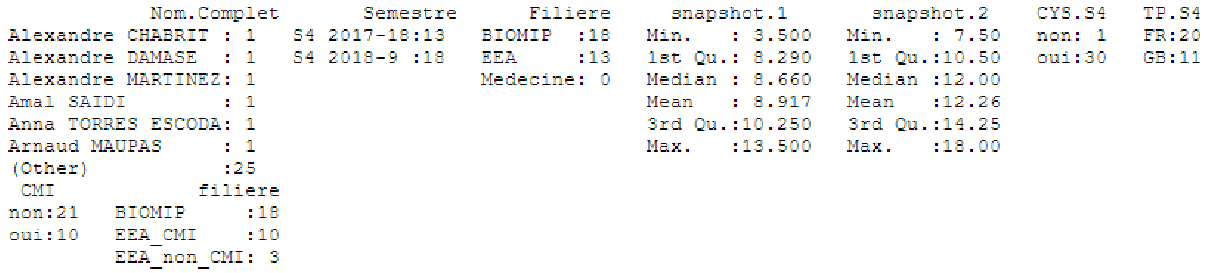
In semester 4, we carried out an assessment of 41 medecin students including:
- All 41 in 2017-2018
- 21 used the CYS's tool while 20 did not use it
- None practised TPs
- None was in CMI

In semester 4, we carried out an assessment of 31 students including:
- 13 in 2017-2018 and 18 in 2018-2019
- 18 in BIOMIP and 13 in EEA
- 30 used the CYS's tool while 1 did not use it
- 20 had practical works in French while 11 used English
- 10 were in CMI (all in EEA) while 21 were not



```
           Nom.Complet        Semestre      Filiere      snapshot.1      snapshot.2    CYS.S4   TP.S4
Alexandre Guibert: 1   S4 2017-18:41   BIOMIP  : 0   Min.   : 6.00   Min.   :11.00   non:20   FR : 0
Alice Gallart    : 1   S4 2018-9 : 0   EEA     : 0   1st Qu.: 9.00   1st Qu.:16.00   oui:21   GB : 0
Alix LOIRET      : 1                   Medecine:41   Median :10.00   Median :19.00            non:41
Alizé Giraudo    : 1                                 Mean   :10.11   Mean   :18.34
Anaïs Le Goff    : 1                                 3rd Qu.:11.00   3rd Qu.:20.50
Antoine CHAULET  : 1                                 Max.   :14.50   Max.   :24.00
(Other)          :35
  CMI
non:41
oui: 0
```

Figure 3: Summary of data semester 4 for medecins



```
            Nom.Complet        Semestre       Filiere      snapshot.1       snapshot.2     CYS.S4   TP.S4
Alexandre CHABRIT : 1   S4 2017-18:13   BIOMIP  :18   Min.   : 3.500   Min.   : 7.50   non: 1   FR:20
Alexandre DAMASE  : 1   S4 2018-9 :18   EEA     :13   1st Qu.: 8.290   1st Qu.:10.50   oui:30   GB:11
Alexandre MARTINEZ: 1                   Medecine: 0   Median : 8.660   Median :12.00
Amal SAIDI        : 1                                 Mean   : 8.917   Mean   :12.26
Anna TORRES ESCODA: 1                                 3rd Qu.:10.250   3rd Qu.:14.25
Arnaud MAUPAS     : 1                                 Max.   :13.500   Max.   :18.00
(Other)           :25
  CMI           filiere
non:21   BIOMIP     :18
oui:10   EEA_CMI    :10
         EEA_non_CMI: 3
```

Figure 4: Summary of data semester 4 for non-medecins

3 Methods

We applied statistical tests on data of both semesters (3 and 4 respectively)

Firstly, descriptive statistics were carried out to determine the most influential factors among considered variables. "Summary" in R provided a range of descriptive statistics at once. Moreover, charts like "boxplot" illustrated which variables should be more important than others. Furthermore, "interaction.plot" showed how variables interacted mutually.

Secondly, linear regression led to a linear formula to study how multiple variables affected the progressions of students simultaneously including their mutual interactions. We have used AIC/BIC as a criterion to choose the best fitting model to the data. Model ANOVA pointed out the effect of 3 qualitative variables on the progression of students as a term of difference between 2 Snapshots and as a term of ratio. On another hand, model ANCOVA pointed out the effect of 3 qualitative variables and Snapshot1 on Snapshot2. By comparing the R^2 values, we maintained the model whose R^2 value is higher.

Thirdly, non-linear regression (decision tree) with the tree graph demonstrated how multiple variables affected the progressions of students.

According to cross-validated predictions, we kept the model which possessed the minimal complexity parameter. As the same method we applied on linear regression, we studied 2 cases:
* The effect of 3 qualitative variables on the progression of students as a term of difference between 2 Snapshots.
* The effect of 3 qualitative variables and Snapshot1 on Snapshot2
We calculated the cross-validation errors of those models to reach the proper models.

Developments were carried out in R.

4 Results

4.1 Descriptive statistics

4.1.1 Semester 3

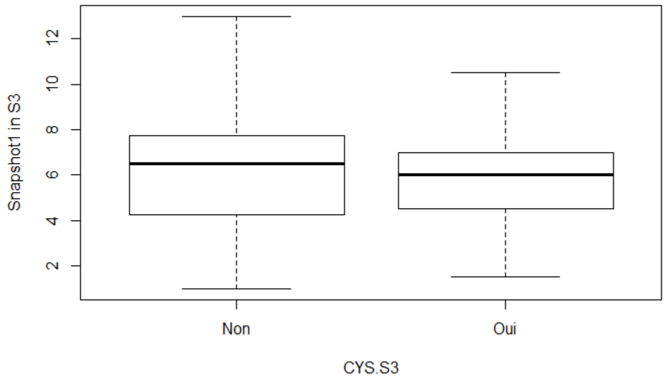


Figure 5: Snapshot1 of non-CMIs

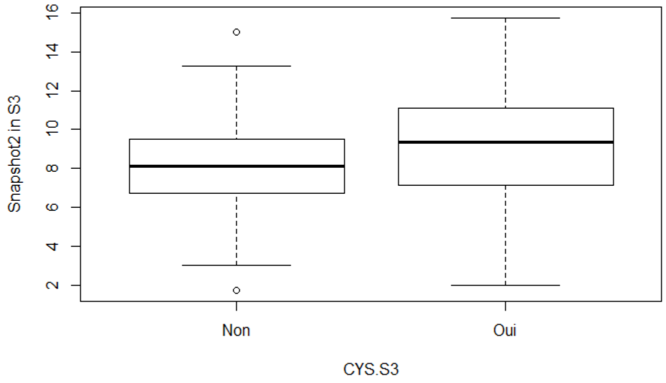


Figure 6: Snapshot2 of non-CMIs

According to the boxplots (Figure 2,3), we found the fact that:
- Generally, a non-CMI student had a higher Snapshot 1 score when he/she used the CYS tool.
- However, among non-CMI students, those who used CYS had higher Snapshot 2 scores than those who did not use it.

We have observed a positive effect of the CYS tool applied to non-CMI students

According to the boxplots(Figure 4,5), we found the fact that:
- Generally, a CMI student had a lower Snapshot 1 score when he/she did not use the CYS tool.
- Among non-CMI students, those who used CYS had lower Snapshot 2 scores than those who did not use it.

We will study for this case the result evolution

Thus, CMI students progressed better when they did not use CYS.

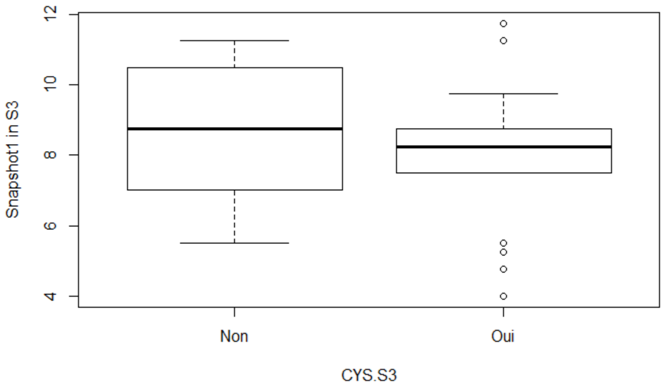


Figure 7: Snapshot1 of CMIs

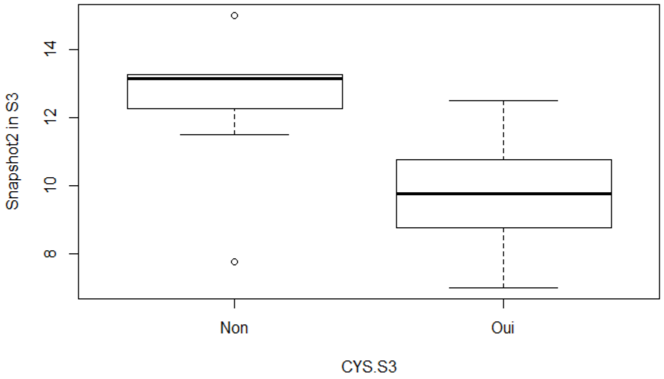


Figure 8: Snapshot2 of CMIs

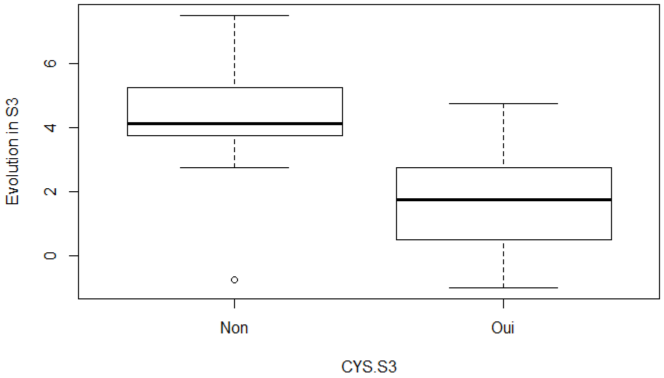


Figure 9: Evolution of CMIs

We found that:

- The interaction between CYS and CMI was important: A student using CYS progressed less if he was in CMI. On the other hand, a student not using CYS progressed more if he was in CMI.

- The interaction between TP and CMI did not exist because no CMI student was using French for TPs.

- The interaction between CYS and CMI existed: A student using CYS progressed a little more if he practised TPs in English. However, a student not using CYS progressed much more if he practised TPs in English.

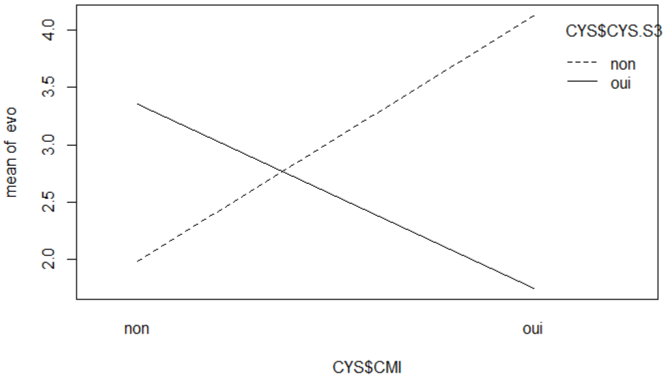


Figure 10: Interaction between CMI and CYS

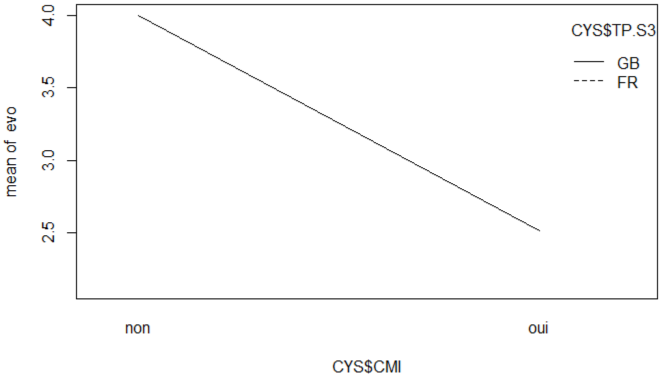


Figure 11: Interaction between CMI and TP

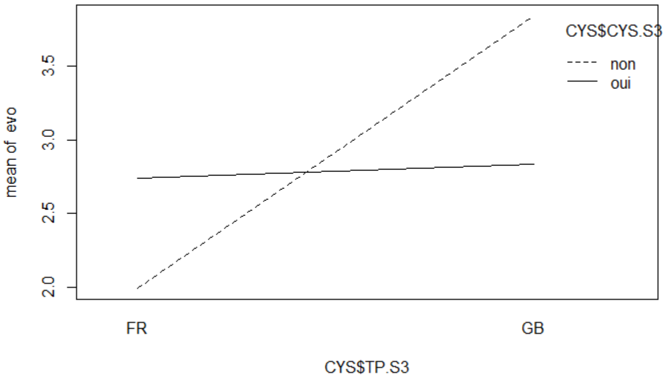


Figure 12: Interaction between CYS and TP

4.1.2 Semester 4

4.1.2.1 Medecin

According to the boxplots (Figure 2,3), we found the fact that:

- Generally, a medecin student had no remarkable difference on Snapshot 1 score whether he/she used the CYS tool.

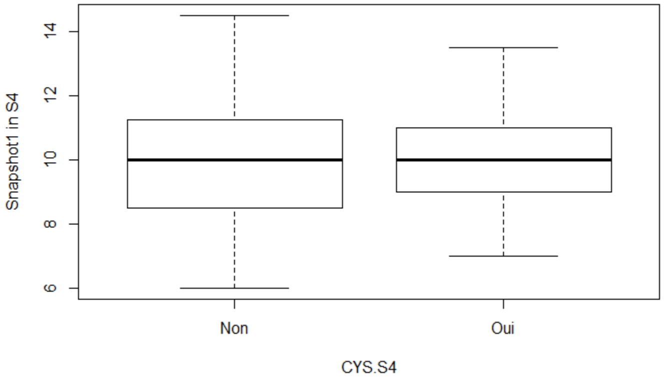


Figure 13: Snapshot1 of medecins

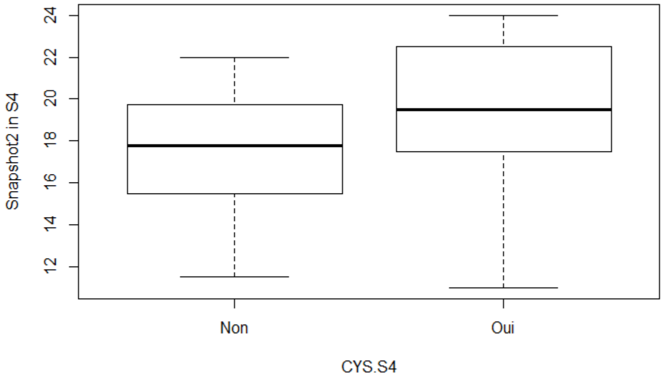


Figure 14: Snapshot2 of medecins

- However, among medecin students, those who used CYS had slightly higher Snapshot 2 scores than those who did not use it.

We have observed a positive effect of the CYS tool applied to non-medecin students

We can visualise the data of medecin student based on the use of CYS:

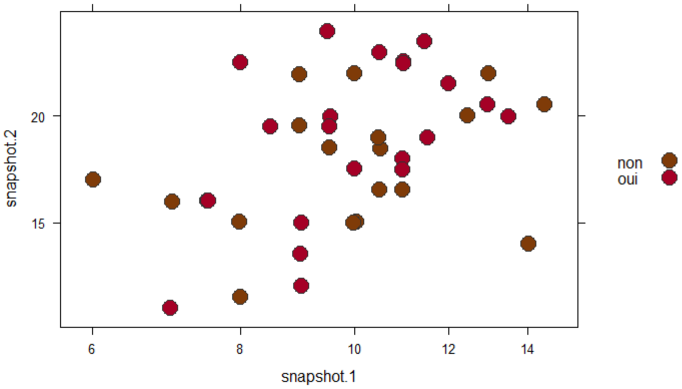


Figure 15: Cloud of points according to the use of CYS

4.1.2.2 Non-medecin

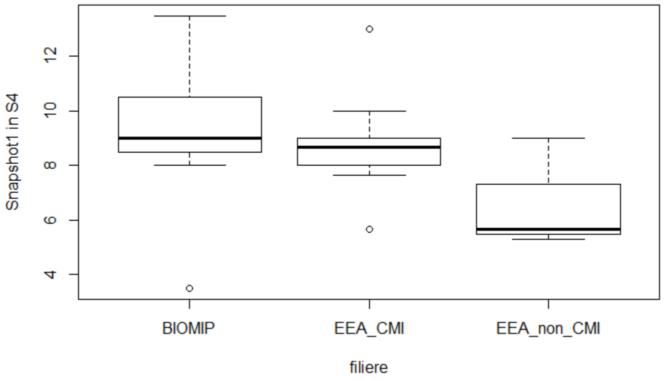


Figure 16: Snapshot1 of non-medecins

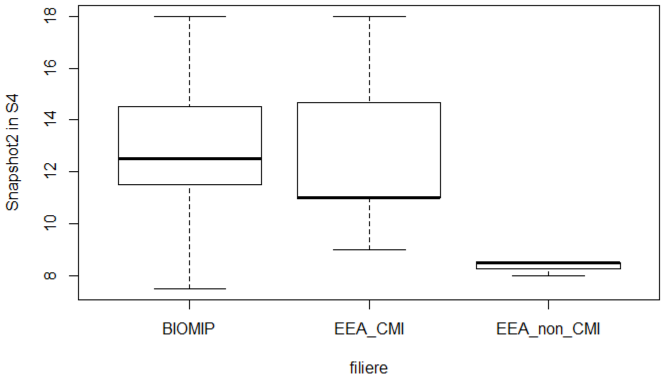


Figure 17: Snapshot2 of non-medecins

According to the boxplots(Figure 4,5), we found the fact that:

- Generally, a non-CMI student in EEA had the lowest Snapshot 1 score while BIOMIP students and CMI students in EEA had mostly same Snapshot 1 scores.

- A non-CMI student in EEA had the lowest Snapshot 2 score while a BIOMIP student had the greatest score.

We can visualise the data of non-medecin student based on faculties:

We can also visualise the data of non-medecin student based on language of TPs:

Only 1 among 31 non-medecin students did not use CYS so we could eliminate the variable CYS in the model.

We found that BIOMIP students always practised TPs in french while CMI-students in EEA practised TPs in English. The number of non-CMI students in EEA was just 3.

Thus, there is no interaction between faculties and TPs.

4.2 Linear regression

We started to conduct ANOVA models to study the impact of CMI, CYS and TP language factors on the evolution of results between Snapshot 1 and Snapshot 2 in semester 3

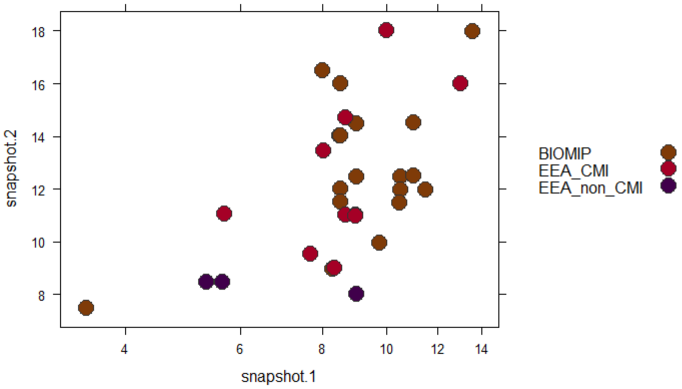


Figure 18: Cloud of points for non-medecin based on faculties

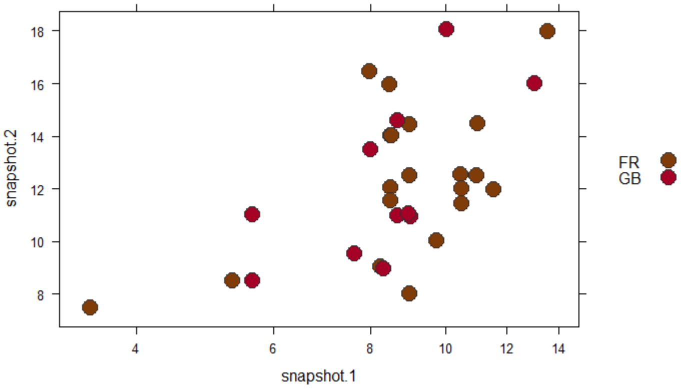
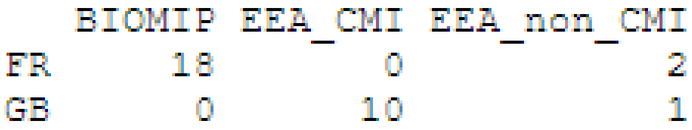


Figure 19: Cloud of points for non-medecin based on language of TPs



```
   BIOMIP EEA_CMI EEA_non_CMI
FR     18       0           2
GB      0      10           1
```

Figure 20: Cross table of non-medecin students

4.2.1 ANOVA

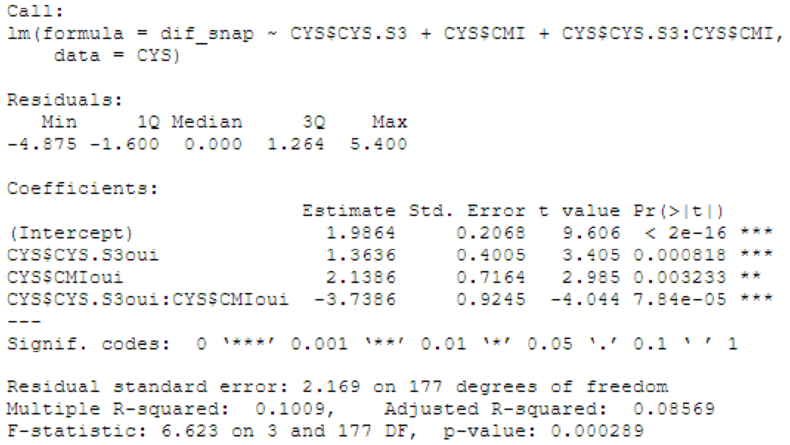
We applied ANOVA just for data in semester 3 due to the fact that the R^2 -value in semester 4 was far lower than ANCOVA.

By applying the BIC criterion on the complete model, we found the ”modBIC2” model, where we did not find the impact of the TP language:

- When we considered the difference between 2 snapshots we observed the modBIC1 model where this difference depended on the use of CYS, the fact of being in CMI and an interaction term between these two.

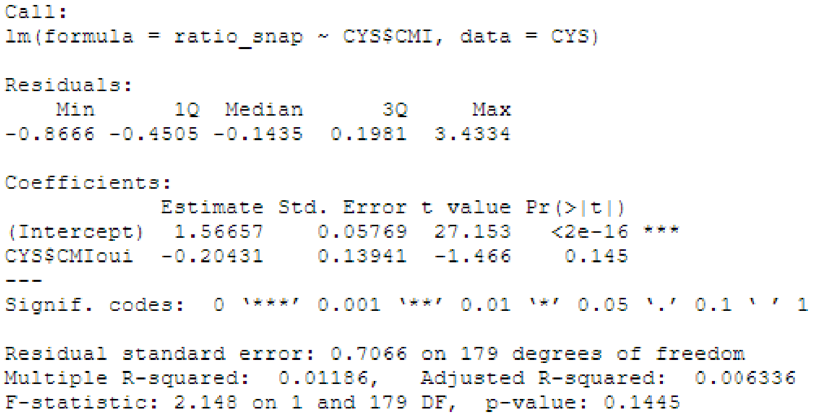
- When we considered the ratio between 2 snapshots we observed the modBIC2 model where this ratio depended only on being in CMI.

However, we had the R-adjusted values too small (less than 0.1) so we decided to go further towards the ANCOVA model



```
Call:
lm(formula = dif_snap ~ CYS$CYS.S3 + CYS$CMI + CYS$CYS.S3:CYS$CMI,
    data = CYS)

Residuals:
   Min     1Q Median     3Q    Max
-4.875 -1.600  0.000  1.264  5.400

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.9864     0.2068   9.606  < 2e-16 ***
CYS$CYS.S3oui               1.3636     0.4005   3.405 0.000818 ***
CYS$CMIoui                  2.1386     0.7164   2.985 0.003233 **
CYS$CYS.S3oui:CYS$CMIoui  -3.7386     0.9245  -4.044 7.84e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.169 on 177 degrees of freedom
Multiple R-squared:  0.1009,    Adjusted R-squared:  0.08569
F-statistic: 6.623 on 3 and 177 DF,  p-value: 0.000289
```

Figure 21: ANOVA model in case difference



```
Call:
lm(formula = ratio_snap ~ CYS$CMI, data = CYS)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8666 -0.4505 -0.1435  0.1981  3.4334

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.56657    0.05769  27.153   <2e-16 ***
CYS$CMIoui  -0.20431    0.13941  -1.466    0.145
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7066 on 179 degrees of freedom
Multiple R-squared:  0.01186,   Adjusted R-squared:  0.006336
F-statistic: 2.148 on 1 and 179 DF,  p-value: 0.1445
```

Figure 22: ANOVA model in case ratio

4.2.2 ANCOVA

4.2.2.1 Semester 3

According to the BIC test, we found the best modBIC2 model:

$$snapshot.2 \sim snapshot.1 + TP.S3 + CYS.S3 + CMI + CYS.S3 : CMI$$

In fact, we estimated the result of Snapshot2 by the model:

$$(modbest) : Snapshot2_{ijkl} = \mu + \alpha Snapshot1_{ijkl} + \beta_i + \gamma_j + \theta_k + \delta_{jk} + \varepsilon_{ijkl}, \forall i = 1,\, 2,\, \forall j = 1, 2, \forall k = 1, 2$$

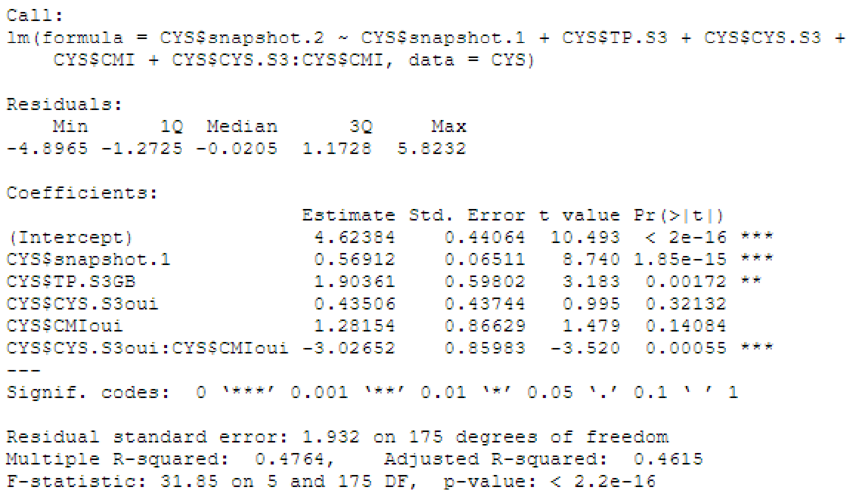
where:

i, j, k are modality indices of the qualitative variables TP.S3, CYS.S3 and CMI, respectively (1 for the answer *No* and 2 for the answer *Yes*, in the case of TP 1 for *FR* and 2 for *GB*)

The index *ijkl* is to indicate the l-th individual having modalities *i, j, k* for TP.S3, CYS.S3 and CMI, respectively. $varepsilon_{ijkl}$ is errors in the estimation of the individual with the index *ijkl*.

From where:

$$\mu = 4.62384$$
$$\alpha = 0.56912$$
$$\beta_1 = \gamma_1 = \theta_1 = \delta_{11} = \delta_{12} = \delta_{21} = 0$$
$$\beta_2 = 1.90361$$
$$\gamma_2 = 0.43506$$
$$\delta_{22} = -3.02652$$



```
Call:
lm(formula = CYS$snapshot.2 ~ CYS$snapshot.1 + CYS$TP.S3 + CYS$CYS.S3 +
    CYS$CMI + CYS$CYS.S3:CYS$CMI, data = CYS)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8965 -1.2725 -0.0205  1.1728  5.8232

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               4.62384    0.44064  10.493  < 2e-16 ***
CYS$snapshot.1            0.56912    0.06511   8.740 1.85e-15 ***
CYS$TP.S3GB               1.90361    0.59802   3.183  0.00172 **
CYS$CYS.S3oui             0.43506    0.43744   0.995  0.32132
CYS$CMIoui                1.28154    0.86629   1.479  0.14084
CYS$CYS.S3oui:CYS$CMIoui -3.02652    0.85983  -3.520  0.00055 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.932 on 175 degrees of freedom
Multiple R-squared:  0.4764,    Adjusted R-squared:  0.4615
F-statistic: 31.85 on 5 and 175 DF,  p-value: < 2.2e-16
```

Figure 23: ANCOVA model

We decided to model the score of Snapshot2 according to Snapshot1, the use of the CheckYourSmile tool, the language of TP and the fact that the student is in CMI or not.

So, under the ANCOVA model, we found that the three qualitative factors had an impact on the Snapshot2 result. However, Snaphot1 had a big effect on the result of Snapshot2. There was also an interaction term between the variable CYS.S3 and the variable CMI and the latter had a significant negative effect on Snapshot2.

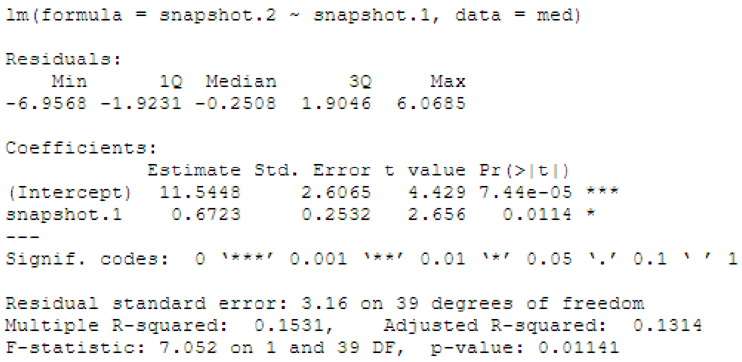
ie, a CMI student using the Check Your Smile tool tended to downgrade by about 2.6 points (-3.02652 + 0.43506) and a non-CMI student using the Check Your Smile tool tended to progress around 0, 4 points (0.43506)

We saw that the ANCOVA model gave us a much better R-fit than the ANOVA model. (0.4615 >> 0.006 and 0.4615 >> 0.09)

To obtain a higher value of R, we will pass under nonlinear models.

4.2.2.2 Semester 4

Medecin

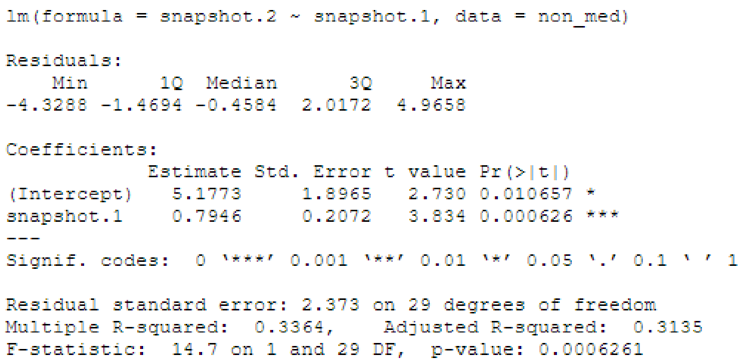


```
lm(formula = snapshot.2 ~ snapshot.1, data = med)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9568 -1.9231 -0.2508  1.9046  6.0685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.5448     2.6065   4.429 7.44e-05 ***
snapshot.1    0.6723     0.2532   2.656   0.0114 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.16 on 39 degrees of freedom
Multiple R-squared:  0.1531,    Adjusted R-squared:  0.1314
F-statistic: 7.052 on 1 and 39 DF,  p-value: 0.01141
```

Figure 24: ANCOVA model

We found that snapshot2 only depends on snapshot1 for medecin students. We did not see the impact of CYS on progression of medecin students. However the R^2-value is such small (0.1531) that we must look for another model.

Non-medecin



```
lm(formula = snapshot.2 ~ snapshot.1, data = non_med)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3288 -1.4694 -0.4584  2.0172  4.9658 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.1773     1.8965   2.730 0.010657 *  
snapshot.1    0.7946     0.2072   3.834 0.000626 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.373 on 29 degrees of freedom
Multiple R-squared:  0.3364,	Adjusted R-squared:  0.3135 
F-statistic:  14.7 on 1 and 29 DF,  p-value: 0.0006261
```

Figure 25: ANCOVA model

We also found that snapshot2 only depends on snapshot1 for non-medecin students. We did not see the impact of CYS on progression of non-medecin students. However the R^2-value is such small (0.3364) that we must look for another model.

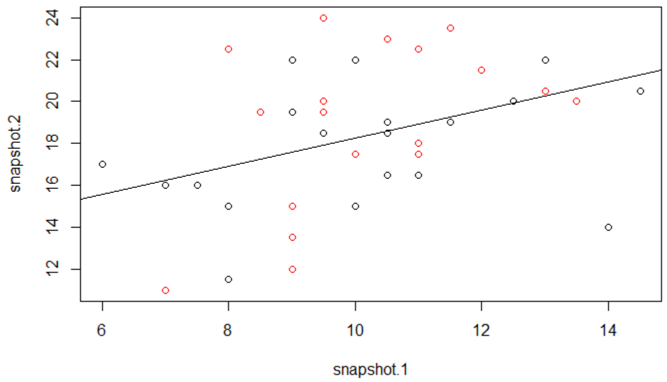


Figure 26: Cloud of points of medecin students and regression line

So we saw no impact of CYS on the progression of students.

4.3 Decision tree

4.3.1 Semester 3

We tried several times to study the behaviour of cross-validation errors of these two types of the tree and we did not see the impact of CYS in the first case.

When we took the second:

- Among students practising TPs in English, a student using the CYS tool progressed less than another not using CYS. (see sheets 5,6,7). Besides, among students practising TPs in English and using the CYS tool, a CMI student progressed less than another non-CMI.

- The binary regression tree allowed us to conclude that the effect of the CYS tool in the progression of students was not remarkable.

- The cross-validation error of the ANCOVA model was smaller than that of the decision tree. However, we found the same phenomenon for CMIs on the effect of CYS on student progression.

In term of cross-validation error:

* The first tree gave us 122.9607

* The second tree gave us 123.3507

By the way, ANCOVA gave us 100.7896 as cross-validation error.

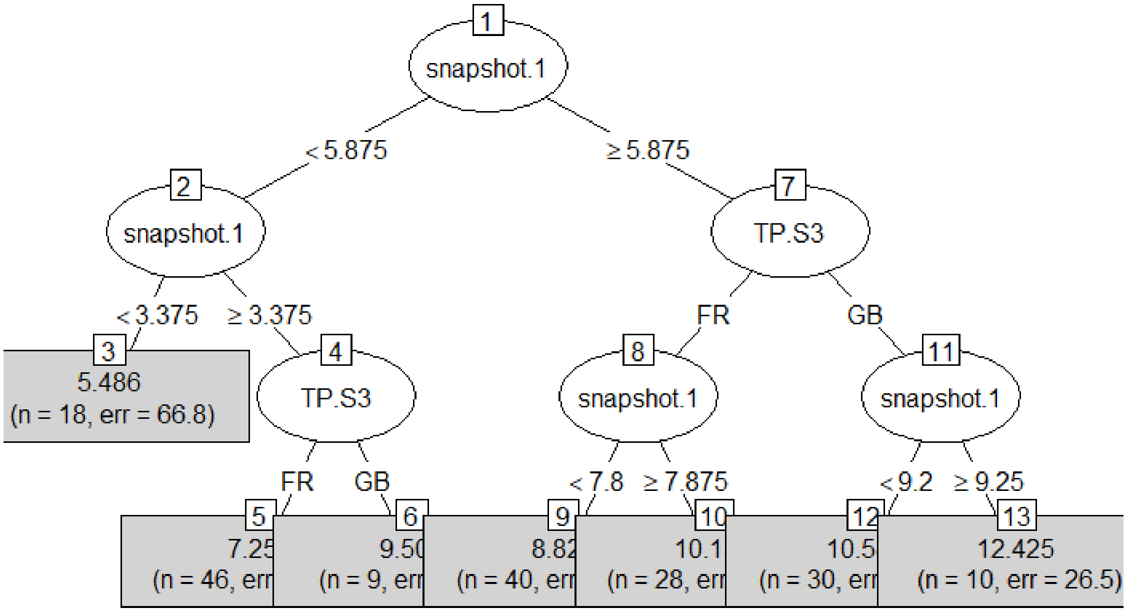


Figure 27: Regression tree

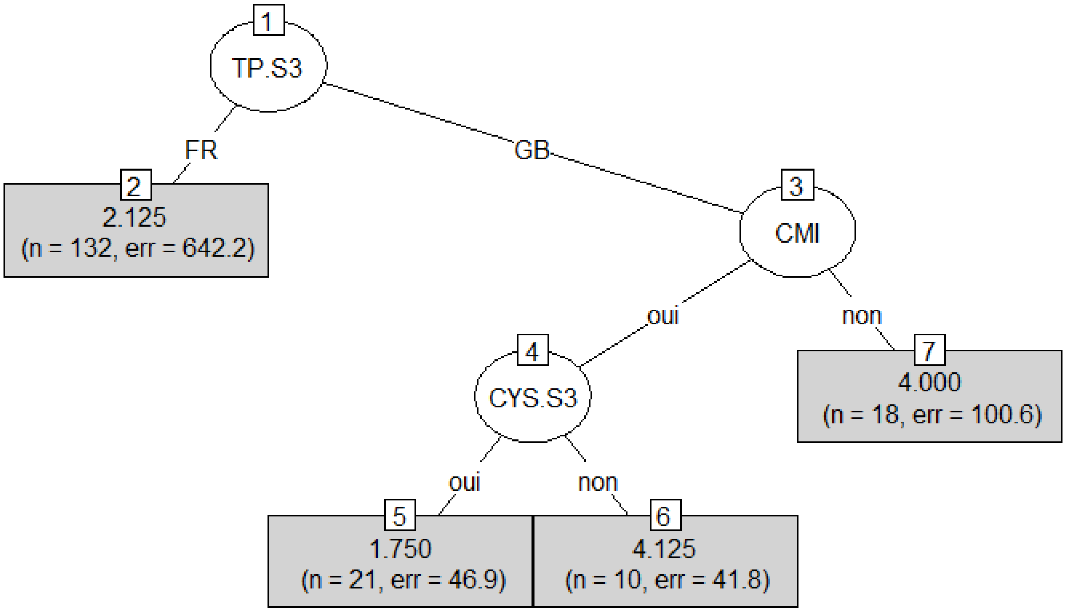


Figure 28: Regression tree of difference

4.3.2 Semester 4

4.3.2.1 Medecin

We saw a positive impact of CYS on medecin students. On average, a medecin student using CYS progressed more than 2 points compared to a medecin student not using CYS.

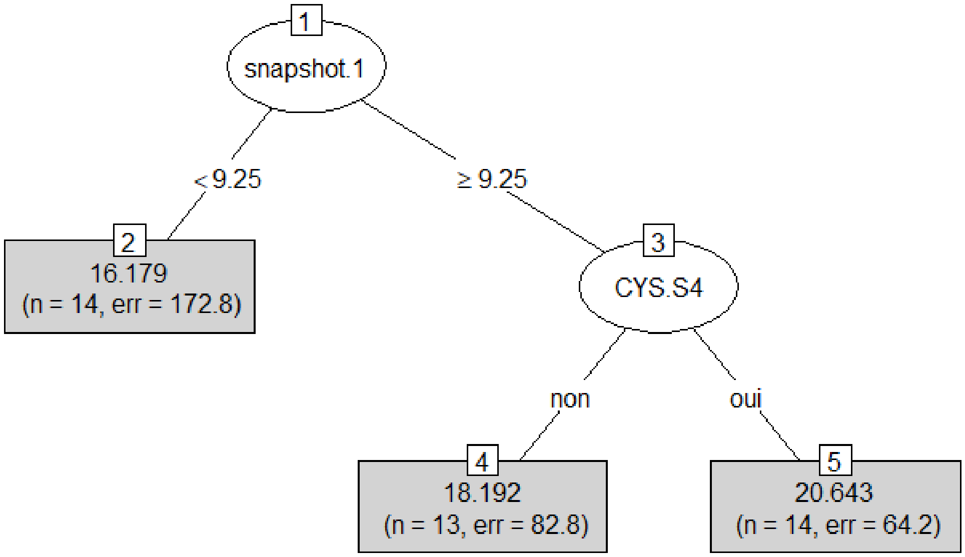


Figure 29: Regression tree of medecin students

4.3.2.2 Non-medecin

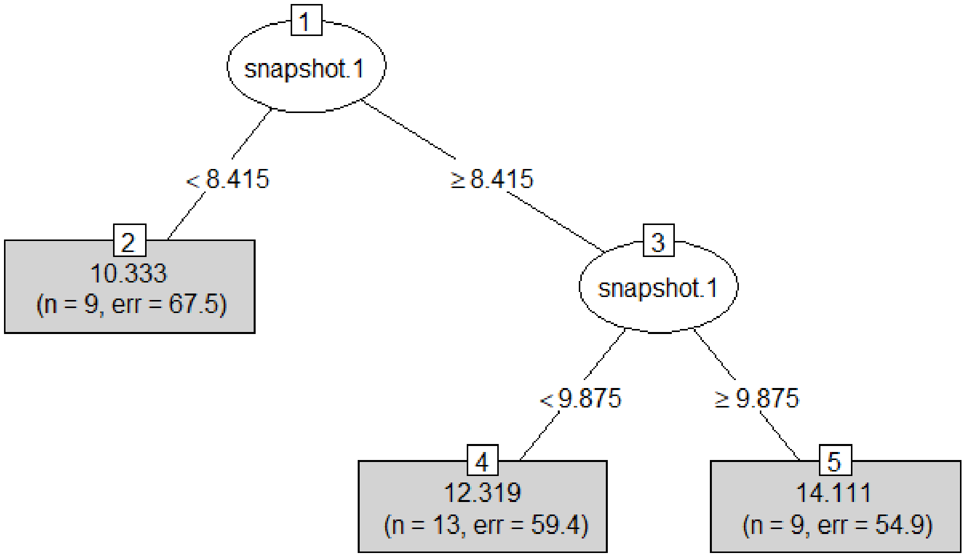


Figure 30: Regression tree of non-medecin students

Only 1 non medecin student did not use CYS so we could not conclude the impact of CYS on progression of students.

In term of cross-validation error:

* The decision tree of medecins gave us 159.01

By the way, ANCOVA gave us 161.46 as cross-validation error.

5 Discussion

In semester 3:

Both linear(ANCOVA) and non-linear(decision tree) models led us to the fact that CYS only had a positive impact on those who were not in CMI while it had a negative impact on those who were in CMI. This result matches up with those found in descriptive statistic

However, the non-linear model showed us how the variable TP impacted on the result of students using CYS while linear model could not.

In semester 4:

The non-linear(decision tree) model led us to the fact that CYS had a positive impact on medecin students. This result matches up with those found in descriptive statistic .

In the case of non-medecin, only 1 among them did not use CYS. Thus, we could not find its impact on the progression of students.

In a word, modeling the score of snapshot2 from snapshot1 and other variables(TP,CYS,CMI) in semester 3 are not good enough to claim the effectiveness of CYS. We prefer to add more variables, which resulted in such small R^2-values. In semester 4, modeling the score of snapshot2 from snapshot1 and CYS do not ensure a good model.

References

LEWIS, R. J. (2000). An introduction to classification and regression tree (cart) analysis. *In Annual meeting of the society for academic emergency medicine in San Francisco, California*, volume 14.

MONTGOMERY, D. C., PECK, E. A. et VINING, G. G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

Montgomery *et al.* (2012) Lewis (2000)