Modeling Trajectories to Understand the Delayed Completion of Sequential Curricula Undergraduate Programs

Renato Boegeholz[0000−0002−6826−1210], Julio Guerra[0000−0002−8296−9848], and Eliana Scheihing[0000−0003−1801−9167]

Instituto de Informatica, Universidad Austral de Chile, Chile
renato.boegeholz@uach.cl,{jguerra,escheihi}@inf.uach.cl

Abstract. Taking more time than expected to complete university degree programs is a global and known problem, and in Chile, has relevance because pressure exists to complete degrees on time. In this work, we explore academic delay in higher education programs, in particular an engineering program, and its relation with academic information summarizing the trajectory of students along with the academic program. Academic information is represented by semester-by-semester features that reflect different aspects such as performance, workload, and difficulty. Exploratory analyses of these variables reveal two orthogonal groups: performance and workload; then used to build models predicting the relative delay of a student at her 8th term at the program relative to the expected completion at 8th term. To further explore the trajectory of delay and analyze how the delay relates to other academic aspects such as term by term performance or workload, a sequential model was built. Results show different patterns of behaviors across different levels of delay in the 8th term. The methods and results of this research can be used by educational institutions or the government to support its decisions about the use of resources and attrition rates reduction.

Keywords: academic analytics · curricular analytics · learning analytics, · educational data mining · academic trajectories · time to degree.

1 Introduction

Taking more time than expected to complete university degree programs is a global and known problem. The average time to degree[1] is 1.36 [1] in Latin America, and 1.31 in Chile, a number that has not experienced a relevant variation in the last 10 years. This statistic not only means that obtaining a degree

[1] The ratio between the average time it takes for students to graduate; and the theoretical duration of the study program

takes on average about 31% more time than expected but also reveals that career delays are a serious problem with economic considerations: this situation causes in Chile an additional expense for families and the government estimated at US$ 500 million [2,3]. Delay in completing a degree program may have multiple causes such as the failure and repetition of subjects, temporary suspension of studies, or assuming less course load than expected according to the curricular plan [4]. Special importance is given to academic reasons behind academic delay because pressure exists to complete degrees on time: high education in Chile is financed by the student (or her family) with the help of scholarships or other funding benefits that require good performance and usually do not tolerate delays [5].

In this work, we explore academic delay in higher education programs, in particular an engineering program, and its relation with academic information summarizing the trajectory of students along with the academic program. While similar work focused in performance variables such as grades [6, 17, 19], we include other relevant aspects of the academic information such as the course load taken by the student each term, the difficulty associated with the courses taken, and repetition of courses which academic situation may determine the risk of punitive actions (for example, elimination of the study program by failing a subject more than twice, or failing, in the same semester, more than two subjects), among others. The first research question of this work is:

1) What is the relation between delay in obtaining the degree and academic information along the trajectory of the student?

We build prediction models on academic delay considering different features that can describe academic trajectories. As mentioned before, we seek to represent such trajectories in terms of different academic information spanning performance, course workload, and difficulty. These academic factors are relevant not only because they could predict academic delay, but because they could characterize the trajectories in an *actionable manner*, that is, further analysis of such trajectories could bring insights that could support counseling practices and curricular re-design actions. Thus, we state a second research question:

2) Can the academic trajectories relate to delay be characterized in a manner that provides information about student behavior?

2 Related Work

Researchers have used different approaches to analyze academic information and typically centered around performance measures. Based on grading data and dismissing the background characteristics of the students, [9] identified curriculum subjects that can serve as effective indicators of academic performance. Using X-means [7] to group students yearly, [9] discovered typical progress patterns and evaluated the predictive capacity of the explanatory subjects. Considering the results in an entrance self-assessment test and the academic performance of the first year, [11] used K-means [10] to group students and follow their performance trajectories in the following 2nd and 3rd year, measuring the influence of the first year behavior in the progression of the curriculum [11]. [6] explored multiple personal and social factors that can affect the academic performance of university students and, using the Grade Point Average (GPA) as an explained variable through decision trees [14] proposed a qualitative model to classify and predict it [6].

Dropout has also been investigated. Aiming to relieve dropout, [16] used recommendation systems techniques to predict the grades that students will obtain in future subjects. The predictions were made using personalized multiple linear regressions (PLMR) [15] on student participation data in both traditional classes and Massive Open Online Courses (MOOC) [16]. [13] examined demographic variables as family characteristics; pre-university and university academic performance factors; and the participation or not in recovery courses, to predict the persistence of the students in the study programs. Using as explanatory variables the scores of the ACT standardized test for college admission, the average of grades in high school, the average grades of the first semester of the university; and using analysis of variance (ANOVA), Pearson product-moment correlations, and multiple regression analysis [12] showed that students who were academically prepared to take college-level courses were more likely to persist than students assigned to mandatory recovery courses [13].

Researchers have also focused on performing analyses of students'trajectories to inform the design of curricula. [17] used student performance data from a specific program to perform an analysis of the curriculum design of the program. In particular, they modeled the difficulty of each subject as its contribution (negative or positive) to the students' GPA and then contrasting this measure to a survey of student perception. Using the same performance data they also performed a dropout and enrollment path analysis [17]. Important to consider is that most of the previous work has been carried related to a flexible course–credit systems where the degrees are obtained as the sum of core and optional approved subjects [8, 11, 13, 19]. This context differs from our sequential-non-flexible curricula, where the flow of subjects to take is pre-defined for all the terms of the program.

The representation of the academic trajectories of the students is not trivial, and it is necessary to consider the temporal dependence of the variables under study. Following this idea, [19] used frequent pattern mining [18] to reveal academic trajectories to understand the sequences of subjects to take that could improve student performance. In [20], a sequential data model is proposed that explicitly captures the temporal dependencies of the academic performance characteristics to form fingerprints or *signatures* that are constructed allowing different analytical interpretations and the development of predictive models for the risk of academic delay.

As presented, several of the proposed models -all of the regression type- considered the performance of the students as explanatory variables, without taking into account the dependence that exists in the performance of a student in various subjects within the same semester, as well as in successive semesters. Our work distinguished from previous work in several aspects: in its multivariate nature considers in addition to academic performance aspects such as the academic workload and the difficulty of the subjects in each semester; incorporating the temporal dimension in the modeling of the students' trajectories; and exploring such trajectories in the context of a curriculum with a sequential structure, where about 90% of the courses to be completed are mandatory.

3 Methods

3.1 Academic Features and Data Descriptions

Academic information at the level of the degree program includes data of courses taken, passed, failed (with their grades), and dropped in each term of the academic life of the student. We combine these records with historic data and the *expected* curricular progress [2] to generate a series of academic features in each term of the student academic life. These features represent different dimensions related to academic performance, academic workload (courseload), the relative difficulty of the courses, and the consistency between the courses taken and their theoretical order in the curriculum. Nine features for each term of each student were computed and are defined in Table 1. More details about the definition of the variables can be found in Appendix A.

We understand by academic trajectory to all this information organized in a term by term sequence. To allow comparisons between trajectories, we considered only the activity of the first 8 semesters (8 terms) for each student, counted from their first enrollment. These 8 semesters also represent a milestone of the study program because contain the required subjects to obtain the *Licenciatura* degree[3]. Considering this, the explained variable will be the delay in the 8th term (DELAY8), which measures how far is the student of having completed all courses of the first 8 terms of the plan in her first 8 terms of academic life. A student who passed all planned courses of the first 8 semesters of the program in her first 8 semesters of academic life, has delay zero.

To reduce inconsistencies in the comparisons of the trajectories (because of the dependency between the features and the structure of the program curricula), it was decided to analyze only one study program: Engineering in Computer Science. For the period of available data, this program implements three curriculum versions (2008, 2010, and 2015) each with 11 semesters of duration and an average of 6 subjects per semester. To study the performance of the students during their first 8 semesters, only whose admitted between 2008 and 2015 were

[2] In Chile most of higher education programs have a semi-flexible curricular plan, where the study sequence in pre-defined term by term.

[3] In Chile, "Licenciatura" is similar to a bachelor's degree. To obtain this grade is necessary to complete 8 semesters of subjects that are part of an academic major. This degree allows you to continue an academic career. To be qualified for professional practice there are necessary between 2 and 4 semesters of additional subjects.

Table 1: Description and possible values of the features built for this study. Definition of the variables can be found in Appendix A.

Feature	Description	Possible values
GPA	Not cumulative weighted grade average for the semester.	$[1.0, 7.0]$ *with* 4.0 *the passing grade*
$PASSRATE$	Passing rate of the semester (the ratio of passed subjects to passed plus failed subjects).	$[0, 1]$
$FIRSTIME$	The proportion of subjects enrolled for the very first time in the semester.	$[0, 1]$
$PROGRE$	Contribution of subjects passed in the semester to the total of subjects required to obtain the degree.	$[0, 1]$
$WKLOAD$	Academic workload rate for the semester measured in CST to the average semester CST of the program.	$[0, LEN_{Prog}]$
$DIFFIC\_A$	Difficulty of the semester, as an additive measure ("Alpha difficulty").	$[0, max(HFR_j)SUB_i]$
$DIFFIC\_B$	Difficulty of the semester, as a geometric measure ("Beta difficulty").	$[0, 1]$
$DISPAR$	The disparity of subjects enrolled in the semester: the difference, in semesters, between the highest level subject and the lowest level subject.	$[0, 1]$
$DELAY$	Measurement of the delay between the theoretical and actual (average) semester of the student given their date of admission.	$[0, 1]$

considered. The resulting data set was composed of 14,199 records of academic activity of 365 different students.

3.2 Data Modeling

We are interested in modeling the behavior of the 8th-semester delay (DELAY8) as a function of the other eight features defined for each semester. We limit the scope of the independent variables (the features) to the first 4 semesters because of two reasons. First, the idea of predicting delay (and also predicting dropout) gain relevance if a prediction can be done early, thus it seems reasonable to predict a delay in the 8th semester with information from the four first terms. Second, the four initial semesters correspond to the "Bachillerato" milestone in the engineering programs of the university, where the foundational courses of math and physics are concentrated and which have the higher failure rates, thus are strongly related with academic failure or success. The analyses will be performed in three steps.

The first step is to perform an exploratory data analysis (EDA) for all the variables. We summarize the main characteristics of each variable (mean, median, quantiles, and range) together with their box plots, following with correlation matrix and principal component analysis (PCA) to gain an understanding of the structure of the set of variables and identify the most significant variables which can explain the academic delay.

The second step includes building predictive models using two supervised algorithms, linear regression (LR) and support vector machine (SVM) on the delay at 8th semester with other features of the first 4 semesters, such as difficulty and workload as predictors. These analyses target research question 1.

The third step is to characterize academic trajectories in terms of the relation of the term features and delay. An adaptation of the sequential model proposed by [20] is implemented. In our case, this model is built as follows. Each student trajectory is represented as a sequence of 4 nodes (semesters 1 to 4), where each node is a single value representing a *delay score*. To compute the i semester's delay score, first all students are clustered using k-means on their i semester's features. Then the score of the i semesters is the average delay of all students in the cluster (cluster members).

4 Results and Discussion

4.1 Exploratory Data Analysis

We explored the behavior of all the features in their semester-by-semester progression. Box plots and summary statistics for the variables can be found in Appendix B. Figure 2 presented here as a sample shows box plots of the features GPA (not cumulative weighted grade average) and DIFFIC_A (an additive measure of difficulty) for the first 8 semesters.

From the analysis can be observed that the DELAY variable shows distributions with increasing medians and variability through the terms, as students on average accumulate more delay as they stay more terms. It can be observed that the distributions of the average grade (GPA) have medians that increase slightly but consistently through the terms. PASSRATE shows the lowest median in the 2nd semester with a value of 0.5. In the following semesters, the values of the medians increase progressively meanwhile the median passing rate for the 1st semester is much higher than the rest, with a value of 0.7. Dispersion is similar in all semesters.

In the case of the academic workload (WKLOAD), the median distribution in semesters 2nd to 4th is below the academic workload defined by the curriculum and it approaches that value in semester 5. The dispersion of these distributions increases between the 6th and 8th semesters. The relative difficulty, both DIFFIC_A, and DIFFIC_B, show medians that descend as students

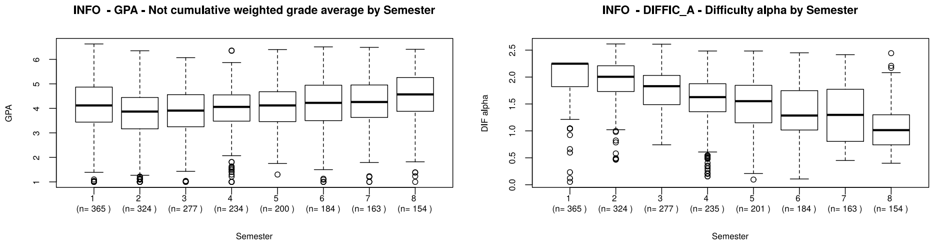

Fig. 2: Variability in the samples across the 8 semesters for (a) The not cumulative weighted grade average (GPA) and (b) The alpha difficulty (DIFFIC_A).

progress in their semesters, with quite similar dispersion. The disparity (DISPAR) shows only two median values: 0.143 for semesters 2 through 5; and 0.286, for semesters 6 through 8. Dispersion appears biased (positively or negatively) for all semesters, except for semester 8 in which symmetry is appreciated.

Correlation matrices and principal component analysis (PCA) were obtained to explore options for reducing the set of variables. The results of the analyzes for semester 3 are described in Figure 4. Biplots and correlograms for all the features can be found in Appendix B. In all semesters it is observed that the delay is negatively correlated with the variables of performance, curricular progress, and academic load, namely: GPA, PASSRATE, PROGRE, AVG, and WKLOAD. On the other hand, it has a very weak positive correlation with the measures of difficulty of the subjects (DIFFIC_A and DIFFIC_B) and weak negative with the measure of disparity (DIS). Two main components of the PCA shows two

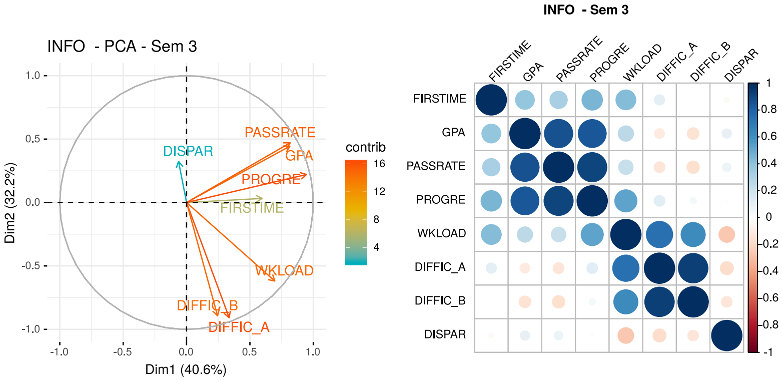

Fig. 4: (a) Biplot and (b) correlogram of the 3rd-semester variables.

groups of variables: i) those that are more related to the individual performance of the students: AVG, PASSRATE, PROGRE and GPA, and ii) those which are related to the characteristics of the study program: WKLOAD, DIFFIC_A, and DIFFIC_B. It is interesting the distinct components represented by performance and workload. To represent each of the groups in the subsequent analyzes of this work, we selected GPA and DIFFIC_A respectively based on a better degree of interpretation than they may have compared to the other features.

Table 2: Summary measures of the predictive capacity of the models. K indicates the number of predictors in the model.

		Regression		SVM		
Model	K	AICc	ΔAICc	RMSE	RSq	MAE
SGD1	2	-123.430	98.913	0.154	0.588	0.127
SGD2	4	-146.596	75.746	0.142	0.645	0.114
SGD3	6	-161.323	61.020	0.139	0.679	0.116
SGD4	8	-177.462	44.881	0.137	0.688	0.115
SA1	8	-118.020	104.322	0.153	0.562	0.127
SA2	16	-153.047	69.296	0.140	0.657	0.117
SA3	24	-180.135	42.208	0.122	0.738	0.102
SA4	32	-222.343	0	0.106	0.805	0.088

4.2 Prediction Models

To analyze the predictive capacity of the explanatory variables, two families of models where built. The SGD_i models use predictors GPA_i and $DIFFIC\_A_i$ adding them incrementally from first to fourth semester. In this way,
$SGD1$ implements: $DELAY_8 \sim GPA_1 + DIFFIC\_A_1$
$SGD2$ implements: $DELAY_8 \sim GPA_1 + DIFFIC\_A_1 + GPA_2 + DIFFIC\_A_2$
and so on.
The SAi models use as predictors all the features in each semester in the same incremental way. Thus,
$SA1$ implements: $DELAY_8 \sim GPA_1 + PASSRATE_1 + WKLOAD_1 + DIFFIC\_A_1 + DIFFIC\_B_1 + DISPAR_1 + FIRSTIME_1 + PROGRE_1$
$SA2$ implements: $DELAY_8 \sim GPA_1 + PASSRATE_1 + WKLOAD_1 + DIFFIC\_A_1 + DIFFIC\_B_1 + DISPAR_1 + FIRSTIME_1 + PROGRE_1 + GPA_2 + PASSRATE_2 + WKLOAD_2 + DIFFIC\_A_2 + DIFFIC\_B_2 + DISPAR_2 + FIRSTIME_2 + PROGRE_2$
and so forth.

Table 2 summarizes the results obtained. Regressions were built using a Generalized Linear Model (GLM) with Gaussian errors and an identity link function. The corrected Akaike Information Criterion (AICc) of every model and the difference with the lower value was computed. We can observe that this indicator declines, as expected, with the inclusion of the variables of the consecutive semesters. An important finding is that the predictive power in the first two semesters is quite similar between the model with two variables and the model with all variables. In contrast, by including semesters 3 and 4, the predictive power of the complete model is much greater. In the case of SVM, the RSME, RSq, and MAE indicators are calculated and ratify the observed with AICc. In particular, with the complete model until the fourth semester, an RSq of 0.805 is obtained which is high.

4.3 Characterization of the Delay Trajectories

Trajectories were built for each student as a sequence of 4 values, one per each of the 4 first terms, each representing the average delay of all students who have similar academic features in a semester. To do this, we performed a clustering at each semester with all its features. Each term, the student falls into one cluster, from which the average delay marks his/her delay trajectory. The number of clusters obtained varied from 3 to 4.

Figure 5a shows the trajectories for all students who completed 8 semesters. To understand the delay behavior along time, the trajectories were presented into 3 groups: those students who reached their 8th semester with 2 semesters of delay or less ($DELAY8 \leq 0.29$), students who reached the 8th semester with a delay between 2 and 4 semesters ($0.29 < DELAY8 \leq 0.57$), and students who have a delay of more than 4 semesters ($DELAY8 > 0.57$).

Figures 5b to 5d shows the trajectories for every one of those groups. It can be observed -as in Figure 5a- that for the first semester, the prediction of delay for all students is 2 or 4 semesters.

Most of the students who complete their 8th semester with a delay lower or equal than 2 semesters (Figure 5b), were projected with 2 semesters or less of delay throughout the entire program.

On the other hand, students who finished the 8th semester with a delay greater than 4 semesters (Figure 5d), maintained similar forecasts during the 4 semesters under study.

It can be seen that in both groups, the less and more delayed, the proportion of “good” and “bad” results for their 1st semester is quite similar, which would make the 1st semester a not reliable indicator of the final delay result.

5 Conclusions

In this research paper, we applied statistical and data mining methods to understand the different behaviors in the progression of students across a sequential curriculum program related to delay in obtaining a degree. First, features that summarize different aspects of the academic records information such as performance, workload, and course difficulty were built for each term a student stays

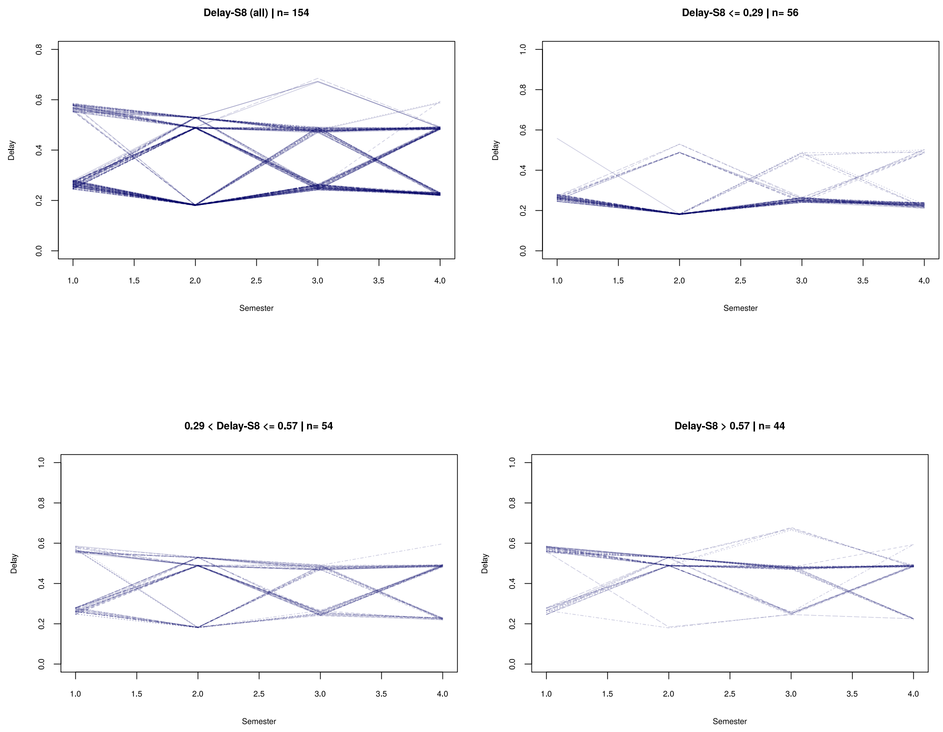

Fig. 6: Delay trajectories for (a) All the students. (b) Students with an 8th-semester delay less or equal than 2 semesters. (c) Students with an 8th-semester delay between 2 and 4 semesters. (d) Students with an 8th-semester delay greater than 4 semesters.

in the academic program. A measure of the delay was custom-made in relation to the expected progress at term 8th. Second, we performed an exploratory data analysis of the features which revealed two different sets of variables that appeared orthogonal within the two principal components of a PCA: a group with all performance variables, and a group with workload and difficulty. The weighted average grade (GPA) and one measure of difficulty (DIFFIC_A) were selected to represent these groups. Third, the predictive capacity of the features was explored, revealing that the selected two variables can predict the 8th-semester delay close enough as a model using all predictors. Fourth, delay sequences were modeled and represented as trajectories showing distinguishable groups of students'behavior. The evidence provided shows that the delay is not strictly determined by the students'performance during their first semester. Low initial performances can follow a path of progressive improvement and reduce their potential delay while ends the program. In conclusion, this work provides valuable insight into a better understanding of the dynamics of the progress in a sequential curricular program, potentially contributing to the decision-making of institutions, directors, and students.

6 Acknowledgments

Work funded by Universidad Austral de Chile and the LALA project (grant no. 586120-EPP-1-2017-1-ES-EPPKA2-CBHE-JP). This project has been funded with support from the European Commission. This publication reflects only the views of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

1. Ferreyra, M., Avitabile C., Botero, J., Paz, F., & Urzúa, S. (2017). At a Crossroads: Higher Education in Latin America and the Caribbean. Directions in Development. Washington, DC: World Bank.
2. SIES Servicio de Información de Educacion Superior del Ministerio de Educación de Chile. (2018). Informe Duracion Real y Sobreduracion de las carreras de Educación Superior (2013-2017) (Spanish).
3. Aequalis Foro de Educación Superior. (2019). Estimación del gasto fiscal y familiar para financiar la sobre-duración de los estudiantes en las carreras: caso chileno (Spanish).
4. Himmel, E. Modelo de análisis de la deserción estudiantil en la educación superior (Spanish). (2002). Facultad de Educación/ Pontificia Universidad Católica de Chile. (2018). Calidad en la Educación, (17), 91-108.
5. Trevino, E., Valdes, H., Castro, M., Costilla, R., Pardo, C., & Donoso Rivas, F. (2010). Factores asociados al logro cognitivo de los estudiantes de America Latina y el Caribe (Spanish). OREALC/UNESCO.
6. Saa, A. A. (2016). Educational data mining and students' performance prediction. International Journal of Advanced Computer Science and Applications, 7(5), 212-220.
7. Pelleg, D., & Moore, A. W. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In Icml (Vol. 1, pp. 727-734).
8. Robinson, R. (2004). Pathways to completion: Patterns of progression through a university degree. Higher Education, 47(1), 1-20. doi:10.1023/B:HIGH.0000009803.70418.9c
9. Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113, 177-194.
10. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).
11. Campagni R., Merlini D., Verri M.C. (2018) The Influence of First Year Behaviour in the Progressions of University Students. Computers Supported Education. CSEDU 2017. Communications in Computer and Information Science, vol 865. Springer, Cham.
12. Maxwell, S. E., and Delaney, H. D. (2004). Designing experiments and analyzing data: a model comparison perspective. Lawrence Erlbaum Associates, Inc.

13. Stewart, S., Lim, D. H., & Kim, J. (2015). Factors influencing college persistence for first-time students. Journal of Developmental Education, 12-20.
14. Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.
15. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models (Vol. 5). Boston: McGraw-Hill Irwin.
16. Elbadrawy, A., Polyzou, A., Ren, Z., Sweeney, M., Karypis, G., & Rangwala, H. (2016). Predicting student performance using personalized analytics. Computer, 49(4), 61-69.
17. Mendez, G., Ochoa, X., Chiluiza, K., & de Wever, B. (2014). Curricular Design Analysis: A Data-Driven Perspective. Journal of Learning Analytics, 1(3), 84-119.
18. Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In Acm sigmod record (Vol. 22, No. 2, pp. 207-216). ACM.
19. Almatrafi, O., Johri, A., Rangwala, H., & Lester, J. (2016), Identifying Course Trajectories of High Achieving Engineering Students through Data Analytics. ASEE Annual Conference and Exposition, New Orleans, Louisiana. doi:10.18260/p.25519
20. Mahzoon, M. J., Maher, M. L., Eltayeby, O., Dou, W., & Grace, K. (2018). A Sequence Data Model for Analyzing Temporal Patterns of Student Data. Journal of Learning Analytics, 5(1), 55-74.
21. van Eck, M.L., Lu, X., Leemans, S.J.J., van der Aalst, W.M.P. (2015). PM2: a Process Mining Project Methodology. CAiSE 2015, LNCS 9097, pp. 297-313.

Appendix A Definition of Explanatory Variables

The following equations describe how the features were built for each semester i of the student's stay and every subject j enrolled in that semester:

$$GPA_i = \frac{\sum_{j=1}^{SUB_i}(GRA_j \cdot CTS_j)}{\sum_{j=1}^{SUB_i} CTS_j} \in [1.0, 7.0] \tag{1}$$

$$PASSRATE_i = \frac{PASS_i}{PASS_i + FAIL_i} \in [0, 1] \tag{2}$$

$$FIRSTIME_i = \frac{SUB1T_i}{SUB_i} \in [0, 1] \tag{3}$$

$$PROGRE_i = \frac{PASS_i}{SUB_{Prog}} \in [0, 1] \tag{4}$$

$$WKLOAD_i = \frac{\sum_{j=1}^{SUB_i} CTS_j}{CTS_{Avg}} \in [0, LEN_{Prog}] \tag{5}$$

$$DIFFIC\_A_i = \sum_{j=1}^{SUB_i} HFR_j \in [0, SUB_i \cdot max(HFR_j)] \tag{6}$$

$$DIFFIC\_B_i = 1 - \prod_{j=1}^{SUB_i} (1 - HFR_j) \in [0, 1] \tag{7}$$

$$DISPAR_i = \frac{max(SEM_j) - min(SEM_j)}{LEN_{Prog} - 1} \in [0, 1] \; where \; j = 1..SUB_i \tag{8}$$

$$DELAY_i = \begin{cases} 0 & i \leq AVG_i \\ \frac{i - AVG_i}{LEN_{Prog} - 1} & i > AVG_i \end{cases} \in [0, 1] \tag{9}$$

Where:

LEN_{Prog}= The total number of semesters of the study program.
SUB_{Prog}= The total number of subjects to be completed in the program to obtain the degree.
CTS_{Prog}= The total number of CTS credits of the study program.
SUB_i = The number of subjects enrolled in semester i.
$SUB1T_i$ = The number of subjects enrolled in the semester i for the very first time.
$PASS_i$ = The number of passed subjects in semester i.
$FAIL_i$ = The number of failed subjects in semester i.
GRA_j = The final grade obtained by the student in the subject j.
SEM_j = Semester in which the subject j is located within the study program.
AVG_i = Average semester in which the student is, with $AVG_i = \frac{\sum_{j=1}^{SUB_i} SEM_j}{SUB_i} \in [1, LEN_{Prog}]$

CTS_j = The number of CTS credits of the subject j.
CTS_{Avg} = The average number of CTS credits per semester of the study program.
HFR_j = The historical failure rate of the subject j.

Appendix B Exploratory Data Visualization

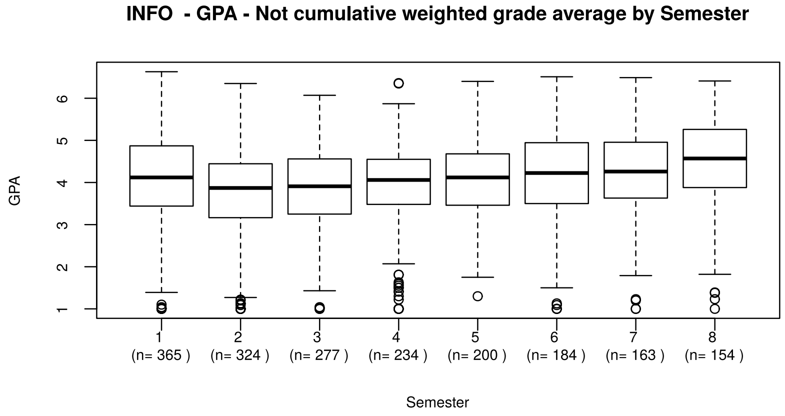

Fig. B.1: Variability in the samples for the weighted semiannual average (GPA), throughout the 8 semesters.

Table of Fig. B.1: Not cumulative weighted grade average (GPA)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	4.087	1.099	1.000	3.440	4.120	4.870	6.630
2	324	3.718	1.106	1.000	3.173	3.870	4.442	6.350
3	277	3.818	0.997	1.000	3.250	3.910	4.560	6.070
4	234	3.896	1.041	1.000	3.482	4.060	4.545	6.360
5	200	4.110	0.981	1.300	3.470	4.120	4.680	6.400
6	184	4.143	1.158	1.000	3.500	4.225	4.942	6.510
7	163	4.305	1.028	1.000	3.630	4.260	4.955	6.490
8	154	4.531	1.030	1.000	3.883	4.570	5.255	6.410

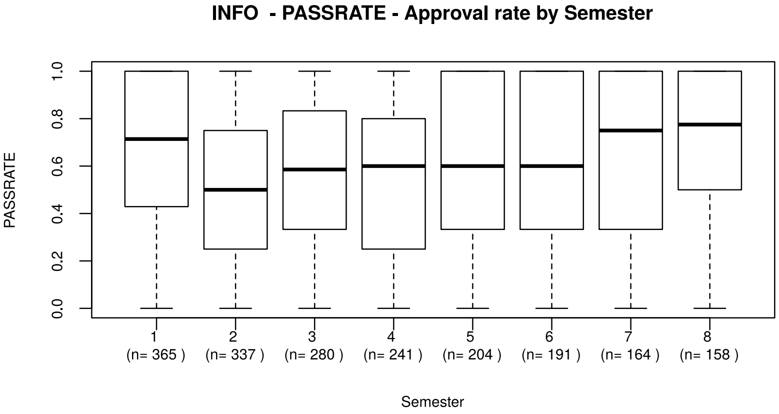

Fig. B.2: Variability in samples for the passing rate (PASSRATE), over the 8 semesters.

Table of Fig. B.2: Passing rate (PASSRATE)

Semester	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	0.672	0.295	0	0.4	0.7	1	1
2	337	0.526	0.321	0.000	0.250	0.500	0.750	1.000
3	280	0.554	0.338	0.000	0.333	0.585	0.833	1.000
4	241	0.558	0.331	0.000	0.250	0.600	0.800	1.000
5	204	0.618	0.336	0.000	0.333	0.600	1.000	1.000
6	191	0.593	0.354	0.000	0.333	0.600	1.000	1.000
7	164	0.661	0.338	0.000	0.333	0.750	1.000	1.000
8	158	0.693	0.314	0.000	0.500	0.775	1.000	1.000

Table of Fig. B.3: First-time-enrolled subjects rate (FIRSTIME)

Semester	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	0.672	0.295	0	0.4	0.7	1	1
2	337	0.526	0.321	0.000	0.250	0.500	0.750	1.000
3	280	0.554	0.338	0.000	0.333	0.585	0.833	1.000
4	241	0.558	0.331	0.000	0.250	0.600	0.800	1.000
5	204	0.618	0.336	0.000	0.333	0.600	1.000	1.000
6	191	0.593	0.354	0.000	0.333	0.600	1.000	1.000
7	164	0.661	0.338	0.000	0.333	0.750	1.000	1.000
8	158	0.693	0.314	0.000	0.500	0.775	1.000	1.000

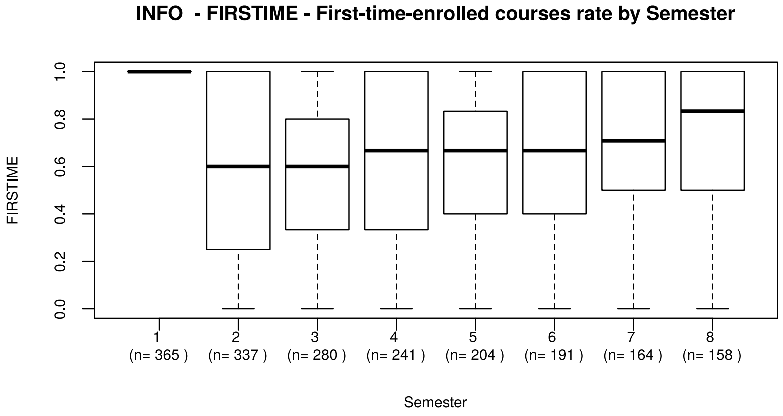

Fig. B.3: Variability in samples for the first-time-enrolled subjects rate (FIRSTIME), over the 8 semesters.

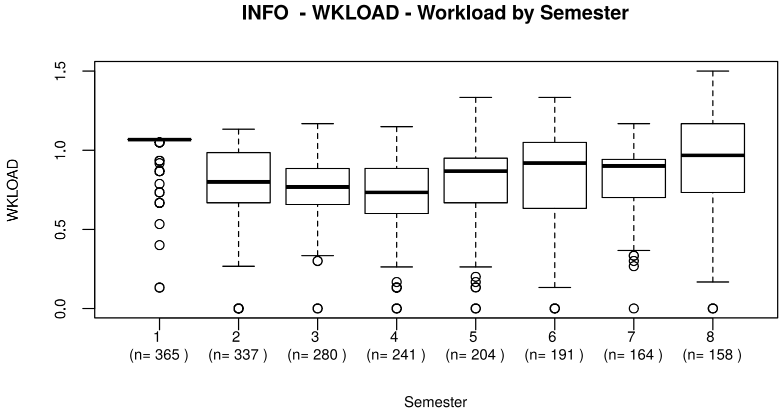

Fig. B.4: Variability in the samples for the academic workload (WKLOAD), throughout the 8 semesters.

Table of Fig. B.4: Academic workload (WKLOAD)

Semester	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	1.036	0.108	0.131	1.067	1.067	1.067	1.067
2	337	0.770	0.226	0.000	0.667	0.800	0.984	1.133
3	280	0.754	0.196	0.000	0.656	0.767	0.875	1.167
4	241	0.713	0.251	0.000	0.600	0.733	0.885	1.148
5	204	0.797	0.253	0.000	0.667	0.867	0.942	1.333
6	191	0.832	0.316	0.000	0.633	0.918	1.049	1.333
7	164	0.827	0.203	0.000	0.700	0.900	0.938	1.167
8	158	0.926	0.289	0.000	0.738	0.967	1.167	1.500

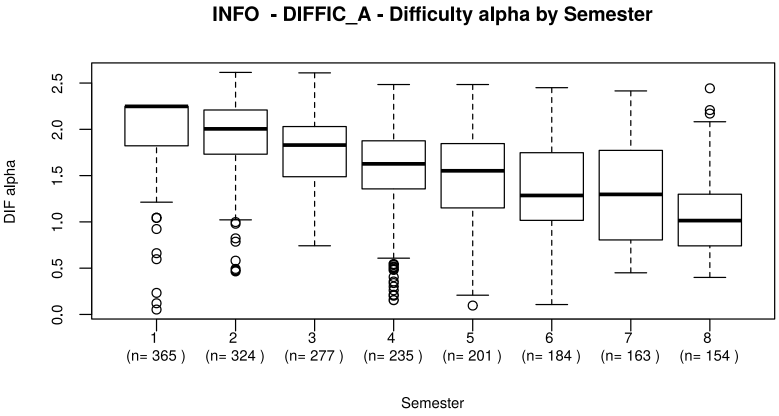

Fig. B.5: Variability in the samples for the alpha difficulty (DIFFIC_A), throughout the 8 semesters.

Table of Fig. B.5: Difficulty (alpha) (DIFFIC_A)

Semester	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	2.009	0.331	0.053	1.822	2.248	2.248	2.248
2	324	1.914	0.386	0.465	1.737	2.005	2.209	2.615
3	277	1.741	0.340	0.742	1.488	1.830	2.030	2.610
4	235	1.552	0.490	0.155	1.357	1.627	1.876	2.484
5	201	1.490	0.472	0.096	1.151	1.552	1.846	2.484
6	184	1.336	0.487	0.107	1.018	1.285	1.747	2.450
7	163	1.307	0.509	0.450	0.805	1.297	1.772	2.415
8	154	1.068	0.436	0.400	0.747	1.014	1.299	2.443

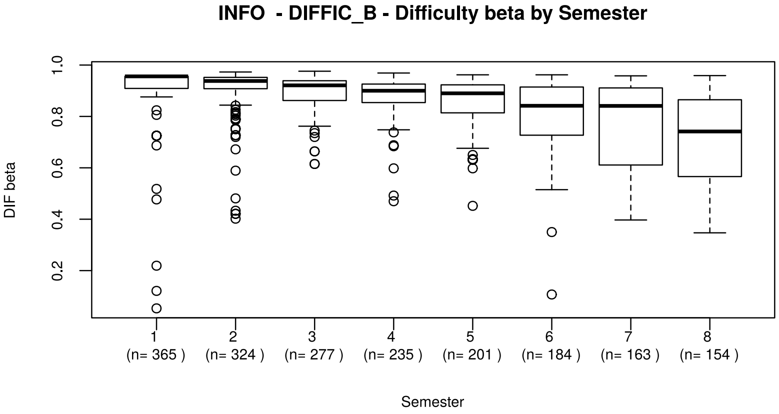

Fig. B.6: Variability in samples for beta difficulty (DIFFIC_B), over the 8 semesters.

Table of Fig. B.6: Difficulty (beta) (DIFFIC_B)

Semester	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	0.922	0.089	0.053	0.909	0.956	0.956	0.956
2	324	0.913	0.084	0.402	0.908	0.938	0.952	0.973
3	277	0.897	0.058	0.615	0.862	0.921	0.939	0.976
4	235	0.882	0.067	0.470	0.854	0.900	0.926	0.969
5	201	0.853	0.096	0.452	0.814	0.890	0.923	0.962
6	184	0.812	0.121	0.107	0.727	0.841	0.914	0.962
7	163	0.773	0.156	0.397	0.611	0.841	0.911	0.958
8	154	0.709	0.175	0.347	0.567	0.742	0.865	0.959

Table of Fig. B.7: Disparity in the semester (DISPAR)

Semester	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	0.000	0.000	0	0	0	0	0
2	337	0.110	0.112	0.000	0.000	0.143	0.143	0.429
3	280	0.132	0.106	0.000	0.000	0.143	0.143	0.429
4	241	0.190	0.136	0.000	0.143	0.143	0.286	0.571
5	204	0.196	0.124	0.000	0.143	0.143	0.286	0.571
6	191	0.237	0.161	0.000	0.143	0.286	0.286	0.714
7	164	0.257	0.150	0.000	0.143	0.286	0.286	0.857
8	158	0.302	0.189	0.000	0.143	0.286	0.429	0.714

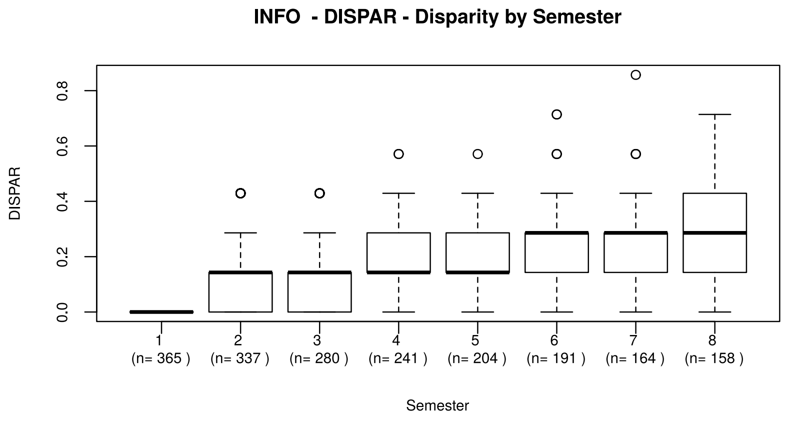

Fig. B.7: Variability in the samples for the disparity in the semester (DISPAR), throughout the 8 semesters.

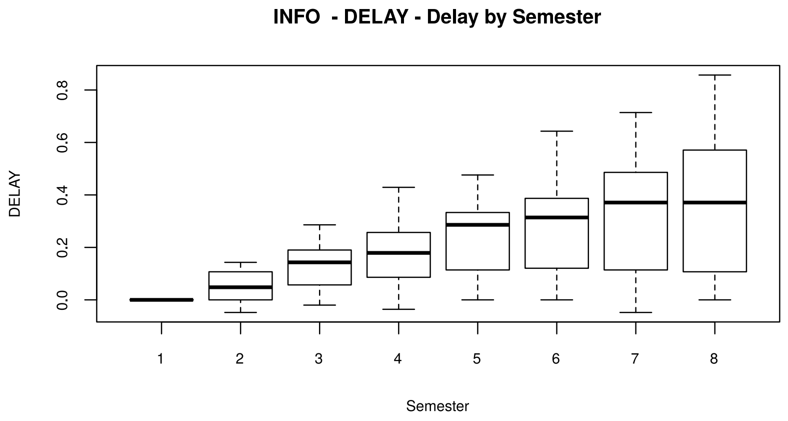

Fig. B.8: Variability in the samples for the delay related to the curriculum (DELAY), throughout the 8 semesters.

Table of Fig. B.8: Delay related to curriculum (DELAY)

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	0.000	0.000	0	0	0	0	0
2	337	0.052	0.055	−0.048	0.000	0.048	0.107	0.143
3	280	0.131	0.085	−0.020	0.057	0.143	0.190	0.286
4	241	0.173	0.105	−0.036	0.086	0.179	0.257	0.429
5	204	0.232	0.134	0.000	0.114	0.286	0.333	0.476
6	191	0.275	0.165	0.000	0.120	0.314	0.387	0.643
7	164	0.316	0.214	−0.048	0.114	0.371	0.486	0.714
8	158	0.362	0.234	0.000	0.107	0.371	0.571	0.857

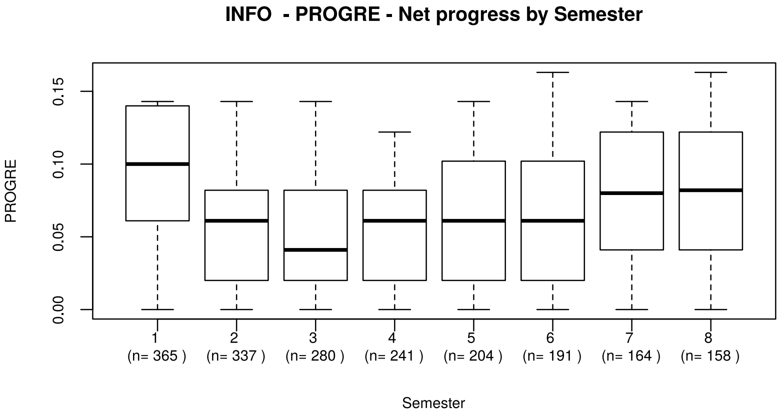

Fig. B.9: Variability in the samples for the curricular progress (PROGRE), throughout the 8 semesters.

Table of Fig. B.9: Curricular progress (PROGRE)

Semester	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
1	365	0.094	0.041	0.000	0.061	0.100	0.140	0.143
2	337	0.055	0.037	0.000	0.020	0.061	0.082	0.143
3	280	0.053	0.037	0.000	0.020	0.041	0.082	0.143
4	241	0.053	0.037	0.000	0.020	0.061	0.082	0.122
5	204	0.062	0.039	0.000	0.020	0.061	0.102	0.143
6	191	0.065	0.047	0.000	0.020	0.061	0.102	0.163
7	164	0.074	0.046	0.000	0.041	0.080	0.122	0.143
8	158	0.082	0.046	0.000	0.041	0.082	0.122	0.163

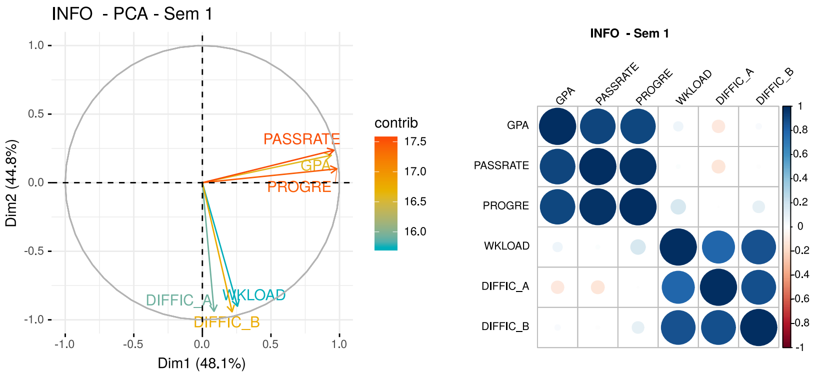

Fig. B.11: Biplot and correlogram for semester 1.

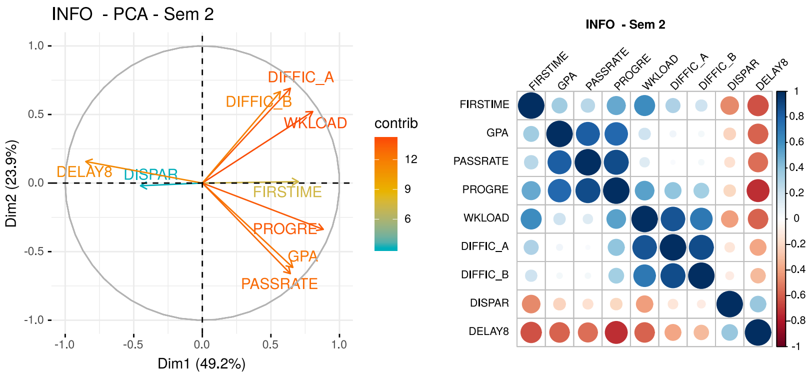

Fig. B.13: Biplot and correlogram for semester 2.

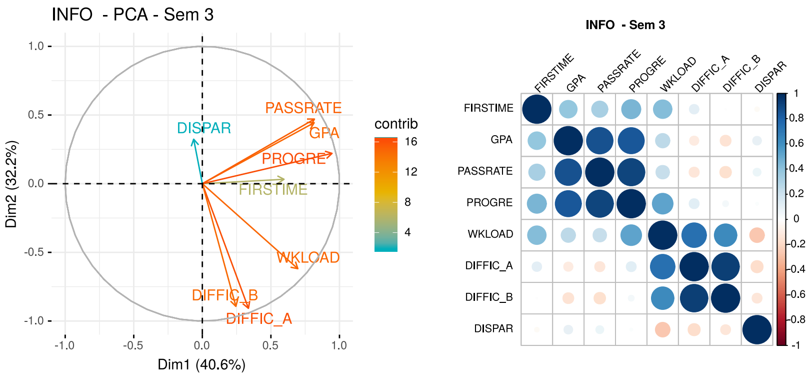

Fig. B.15: Biplot and correlogram for semester 3.

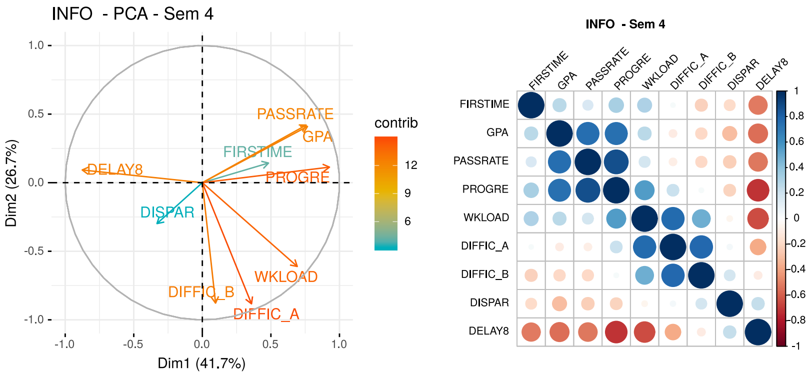

Fig. B.17: Biplot and correlogram for semester 4.

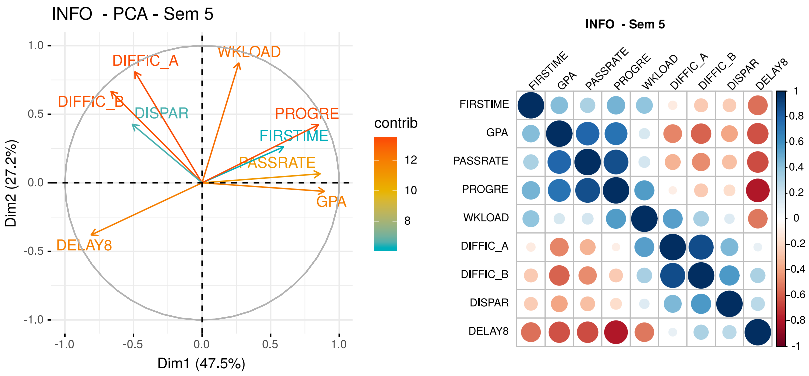

Fig. B.19: Biplot and correlogram for semester 5.

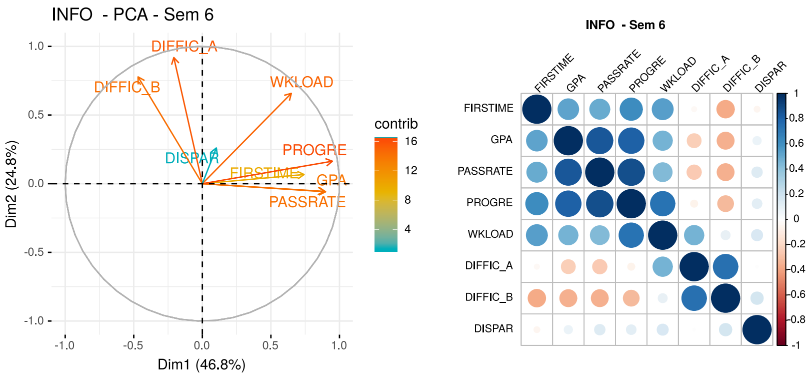

Fig. B.21: Biplot and correlogram for semester 6.

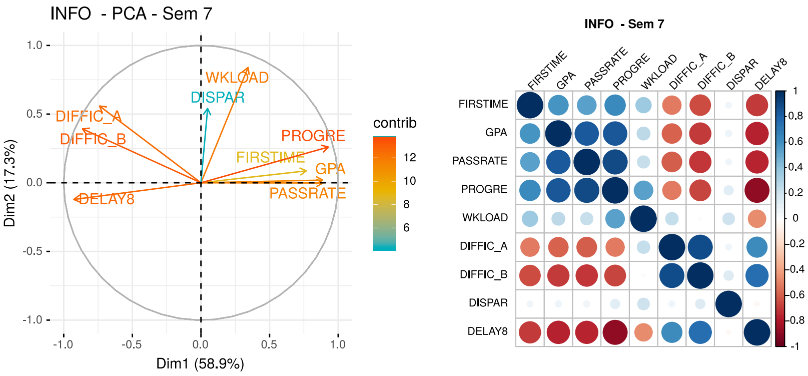

Fig. B.23: Biplot and correlogram for semester 7.

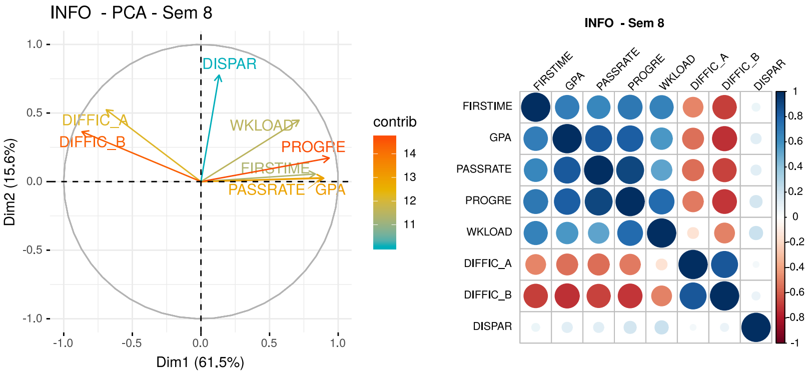

Fig. B.25: Biplot and correlogram for semester 8.