Revista Colombiana de Estadística - Applied Statistics
July 2024, volume 47, issue 2, pp. 475–497
http://doi.org/10.15446/rce.v47n2.112870

Analysis of Academic Data to Group Students According to Their Academic Risk

Agrupamiento de información académica para identificar estudiantes según su riesgo académico

Kevin Andrés Leal Pérez[a], Liliana López-Kleine[b]

Department of Statistics, Universidad Nacional de Colombia, Bogotá, Colombia

Abstract

The Consillium Academica initiative, spearheaded by the academic vice-deanship of the Faculty of Sciences at Universidad Nacional de Colombia in Bogotá, is based on a comprehensive clustering analysis of undergraduate students. This study leverages data spanning from the 2012-1S to 2022-2S academic terms to semi-automatically identify a group of students consistently exhibiting academic underperformance each semester with a potential high risk of academic dropout. The methodology employed in this initiative serves as a proactive measure to identify and support students at risk, to improve the effectiveness of the intervention strategies of the tutor-teacher program, facilitating direct contact between mentors and identified students to provide personalized guidance and academic advisement. This article presents the methodology, key findings, and implications of the Consillium Academica initiative, shedding light on its potential to fortify academic support systems and contribute to the overall success and retention of undergraduate students.

Key words: K-means; Fuzzy C-means; Bootstrap clustering; Academic triage.

Resumen

La iniciativa Consillium Academica, liderada por la vicedecanatura académica de la Facultad de Ciencias en la Universidad Nacional de Colombia-sede Bogotá, lleva a cabo un exhaustivo análisis de agrupación de estudiantes de pregrado. Este estudio aprovecha los datos de los periodos académicos 2012-1S a 2022-2S para identificar de forma semiautomática una cohorte de estudiantes que muestran sistemáticamente un bajo rendimiento

[a] B.Sc. E-mail: klealp@unal.edu.co
[b] Ph.D. E-mail: llopezk@unal.edu.co

académico cada semestre, indicativo de un mayor riesgo de expulsión o abandono académico. La metodología empleada en esta iniciativa sirve como medida proactiva para identificar y apoyar a los estudiantes en situación de riesgo, con el objetivo de mejorar la eficacia de las estrategias de intervención del programa de profesores tutores, el cual, desempeña un papel fundamental en este proceso, facilitando el contacto directo entre los tutores y los estudiantes identificados para proporcionarles orientación personalizada y asesoramiento académico. Este artículo presenta la metodología, los principales resultados y las implicaciones de la iniciativa Consillium Academica, arrojando luz sobre su potencial para fortalecer los sistemas de apoyo académico y contribuir al éxito general y la retención de los estudiantes universitarios.

***Palabras clave*:** K-medias; Fuzzy C-medias; Bootstrap clustering; Triage académico.

1. Introduction

The Consilium Academica project, initiated by the Academic Vice-Deanship of the Faculty of Sciences at the Universidad Nacional de Colombia in Bogotá, addresses the imperative need to group undergraduate students across diverse curricular programs and generate an early alert for students and tutors to avoid academic dropout. The primary objective is to cluster students of the faculty of science (Biology, Computer Science, Statistics, Pharmacy, Physics, Geology, Mathematics, and Chemistry) based on academic data spanning from the first semester of 2012 to the first semester of 2023. The aim is to identify cohorts of active students exhibiting academic performance below the common standard, potentially placing them at risk of academic dropout, given an academic average below 3.0 on a grading scale of 0.0 to 5.0.

Recognizing the critical nature of academic advise and support by teachers in the role of tutors, once students at academic risk are identified, the project proposes a systematic communication of their situation to their designated tutors. These tutors, as stipulated by the Academic Council's Agreement 028 of 2010, play a pivotal role in providing tailored advice on academic aspects, thereby contributing to the students' training process and academic trajectory. The envisaged proactive involvement of tutors includes contacting their respective students, identifying challenges faced by students, and offering guidance and support, whether the difficulties are academic, economic, personal, or familial. In cases where challenges extend beyond academic concerns, students are encouraged to seek assistance from welfare offices.

The Consillium Academica project is designed for sustained effectiveness, intending to run at the beginning of each semester to incorporate newly admitted students and maintain continuous follow-up with at-risk mentees. The core of this initiative lies in the development of a semi-automatic application, facilitating ease of use for future teachers and students involved in generating alerts. To achieve this, the statistical procedures implemented to obtain the resulting groups were Principal Component Analysis, K-Means and Fuzzy C-Means. Additionally, Bootstrap Clustering was used to evaluate the stability of the generated clusters, while the ongoing monitoring of at-risk students and their respective tutors was executed through dedicated surveys tailored for students and teachers, respectively.

1.1. Preventing University Dropout

Early identification of at-risk students is crucial in the academic environment to provide adequate support and reduce dropout rates, which have significant personal, social, and institutional consequences (Ulriksen et al., 2010). The heightened concern regarding university dropout is driven by both financial and human costs, including wasted personal resources, time, and money, coupled with negative emotions of inadequacy and self-doubt (Ulriksen et al., 2010). From a university perspective, dropout negatively impacts economics and conflicts with pedagogical goals of fostering high completion rates within a reasonable timeframe (Sarra et al., 2019). To address this, data mining techniques such as prediction, clustering, relationship mining, and discovery with models are employed in education (Nithya et al., 2016). This paper primarily focuses on clustering techniques to identify students at risk of abandoning university due to academic performance through the analysis and grouping of their academic histories. Clustering techniques are powerful tools for organizing datasets, in this case, academic histories, into groups where elements within a group are more similar to each other than to elements in other groups. However, their implementation requires several additional analyses and statistical techniques to evaluate and improve performance, as the clustering process is highly dependent on the dataset (Hennig, 2007).

Numerous previous studies have employed the techniques mentioned above for identifying at-risk students. Regarding clustering implementation, Sarra et al. (2019) work introduces a Bayesian Profile Regression (BPR) model, a technique that combines classification and prediction. Initially, the clustering model detects subgroups of student profiles using categorical variables related to academic performance, along with indicators of motivation and academic resilience. Simultaneously, the regression model examines the association of these subgroups with the target variable of interest, in this case, academic dropout. In the Colombian context, the work of Roldán Jiménez (2021), conducted by researchers at Pontificia Universidad Javeriana, focuses on developing various methods to predict the number of at-risk students per semester. The study identifies patterns and transitions in the academic performance of undergraduate students using models involving neural networks, logistic regression, and model based on population with Monte-Carlo simulations, accordingly to different academic stages defined in the education model from this institution.

In contrast to the aforementioned studies, this work exclusively employs clustering techniques that take as input quantitative variables constructed to measure academic performance for the entire university population enrolled in undergraduate programs at the Faculty of Sciences since 2012. To implement models like the one presented by Sarra et al. (2019), measuring the motivation and academic resilience of all students would be required. This, in turn, implies the willingness of all students to respond to the necessary questions to obtain those indicators. On the other hand, applying the methodology presented by Roldán Jiménez (2021) proves to be more complex due to the necessity of training or estimating various models that function in different stages of the student's university journey, stages that are already well-defined. In the case of students at the Universidad Nacional de Colombia, there is no equivalent definition for these stages. Instead, the goal is to establish an initial characterization of the student population in the Faculty of Sciences. This approach allows for a broader and more inclusive analysis, focusing on the unique academic landscape of the university and laying the foundation for a comprehensive understanding of academic risk factors.

2. Materials and Methods

This section delineates the methodology employed in constructing and executing the Consillium Academica system, designed with the primary aim of generating reports for students within the Faculty of Sciences who are deemed academically at risk. Subsequent subsections detail the essential procedures, beginning with the reception of data sources, following with the implementation of the clustering technique, then the corresponding analysis and characterization of the results as well as the performance evaluation of the system and finally extending to the monitoring of interactions between tutors and students at risk (see Figure 1). It is noteworthy that the entire system development transpired within the R statistical software framework.

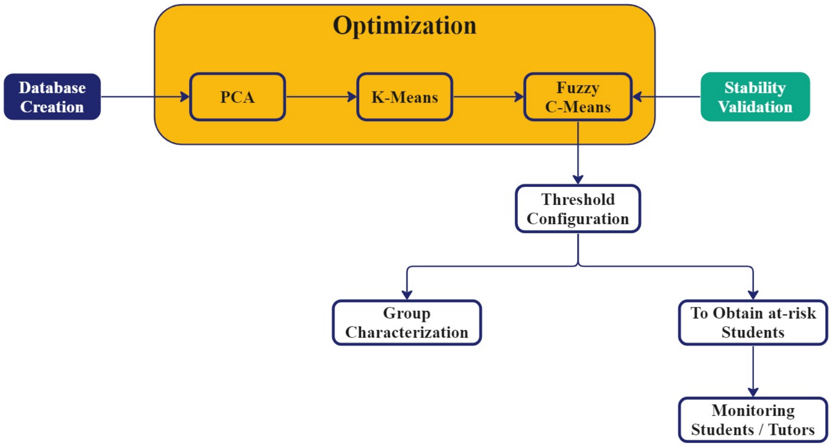


Figure 1: Diagram of the procedures involved in the development of Consillium Academica

2.1. Database Preparation

The source of information corresponds to the databases provided by the Registration and Enrollment Division at the Bogotá Campus, which contain the grades obtained in each course attended by each undergraduate student of the Faculty of

Sciences for all semesters from the first semester in 2012 to the second semester in 2022. These databases present relevant information such as the name, code, typology and number of credits of the class, as well as the unique identification code for the academic history of the students together with the numerical and alphabetical grade obtained. In total these databases have 16 columns and require preprocessing before being used for the development of the project since they present problems of duplicate records and missing data which resulted in a total of 313.064 useful registers.

After cleaning the database, the project focused on grouping student's academic histories because it contains all the information on a student's academic performance in each of the classes taken for a given curricular program of the faculty, for this reason, the total number of students will not match the total records of academic histories since we can find occurrences like double degrees or changes in career paths between the mentioned curricular programs during the available time window leading to the creation of a new identifier for the same student. Therefore, a second database was constructed using the entries for each course, with each new statistical individual representing a student and its history for each specific curricular program. Moreover, the window (number of semesters backwards to be included) stability analysis showed that including or not an atypical semester does not affect mainly the clustering process. We have treated the pandemic years as the other years after preprocessing (eliminating the missing and aberrant entries).

In this new base, the following seven quantitative variables are constructed: cumulative weighted arithmetic average (PAPA, for its Spanish acronym), total credits enrolled, number of enrollments in which the student is registered, total credits passed, percentage of progress in the program, average percentage of progress per semester and number of courses lost. It is important to clarify that the seven variables for each entry of the database have only one value representing this measure at the actual time point. The data set is not longitudinal as all the history of academic entries is resumed at the present time point in the seven numeric variables, in other words, every time that this methodology is executed the matrix of analysis X of size $(N, 7)$ is constructed to execute the clustering techniques, where N is equal to the number of academic histories from students who entered the university during the semesters under consideration.

2.2. Implementation of Clustering Technique

Since the new base of academic histories has seven quantitative variables, before testing the performance of any clustering technique, the principal component analysis (PCA) method is used to reduce the dimensionality of the variable space, by means of the *factoextra* and *FactoMineR* libraries. In data sets with many variables, it is common to use dimensionality reduction techniques to simplify the analysis. On the other hand, an additional advantage offered by this method is that by the very nature of how the dimensions are reduced, the data are reorganized in the new plane, locating individuals that are similar to each other close to each other on the map, which is beneficial at the time of clustering (Pardo, 2020).

Once the PCA is performed, the coordinates of the academic histories in the first three principal axes are used as input for the clustering techniques. Let $Z_{(n,3)}$ be the matrix of these coordinates, the first step is to apply the K-Means (KM) method with the R *kmeans* basis function. It is a partition technique that seeks to group the data into K clusters. This technique starts from the initial choice of K centroids, $\mathbf{C} = \{\mathbf{c_1}, \mathbf{c_2}, \dots, \mathbf{c_k}\}$ and assigns each point $\mathbf{z}_i$ to the nearest centroid. This assignment is performed by minimizing the Euclidean distance:

$$\arg\min_{k} \left\|\mathbf{z}_i - \mathbf{c}_k\right\|^2 \tag{1}$$

Subsequently, the centroids are recalculated based on the newly assigned points:

$$\mathbf{c}_k = \frac{1}{|S_k|} \sum_{\mathbf{z}_i \in S_k} \mathbf{z}_i \tag{2}$$

Where S_k is the set of points assigned to centroid $\mathbf{c_k}$. This process repeats iteratively until the centroids do not vary significantly in their coordinates or until a specified number of iterations are completed. An important challenge of this method is the choice of K. The elbow method provides a solution by plotting the sum of squared errors (SSE) against K. The SSE is given by:

$$\text{SSE} = \sum_{k=1}^{K} \sum_{\mathbf{z}_i \in S_k} \left\|\mathbf{z}_i - \mathbf{c}_k\right\|^2 \tag{3}$$

As K increases, SSE decreases, but at a slower rate. The point where this decrease slows down significantly indicates the optimal K. To know more about K-Means clustering, chapter 7 of Pardo (2020) makes a deep explanation of the algorithm. It is important to note that, as mentioned in Chapter 5 of Vendramin et al. (2015), the result of this algorithm corresponds to a rigid partition, i.e., each individual belongs to one and only one of the groups.

A rigid partition means that if some students in need of attention are on the other side of the *borderline boundary* of a group characterized by containing at-risk students, then it would be impossible to identify them and therefore they would not receive the necessary support. For this reason, the Fuzzy C-Means (FCM) clustering method was resorted, which unlike KM, the final result of this type of techniques is a fuzzy partition matrix from the data into a certain number K of clusters, such that $U = [u_{ij}]_{NxK}$ where elements u_{ij} represents the value of the membership of the ith individual to the jth fuzzy group, which can be interpreted as a probability, because this value is constrained with $u_{ij} \in [0,1]$ and:

$$\sum_{j=1}^{K} u_{ij} = 1, 1 \leq i \leq N \tag{4}$$

Most of clustering algorithms minimize the function

$$J = \sum_{i=1}^{N} \sum_{j=1}^{k} \left(u_{ij}\right)^m D_{ij} \tag{5}$$

the m parameter, referred to as the *fuzzification* exponent, governs the degree of group diffusion and the smoothness of point membership assignment to the groups, it is usually $m = 2$ (Vendramin et al., 2015) and it was the default value for the execution of this project. The closer the value is to 1, the more rigid the K-Means-like partition is obtained, and the higher the value of m, the more uniform the assignment of the probabilities among the groups. D_{ij} is the distance between the ith individual and the jth cluster, and J is a measure of intra-cluster dissimilarity. FCM algorithm provides a fuzzy partition matrix U by minimizing (5) with

$$D_{ij} = \|\mathbf{z}_i - \mathbf{d}_j\|_{\mathbf{A}}^2 = (\mathbf{z}_i - \mathbf{d}_j)^T \, \mathbf{A} \, (\mathbf{z}_i - \mathbf{d}_j) \,, \tag{6}$$

being any squared inner-product distance norm between the ith individual and the jth cluster centroid, $\mathbf{d}_j$. The norm-inducing matrix $\mathbf{A}$, a positive-definite$(p \times p)$ matrix (with p equal to the amount of variables presented in the database, in this case 3), defines the shape of the clusters, however in most of occasions the researcher does not know this shape so it is usually set as the identity matrix $\mathbf{I_{pxp}}$ and D_{ij} becomes the squared Euclidean distance. For more information on the FCM technique look at Vendramin et al. (2015).

Thus, the problem is solved by manipulating the group membership thresholds. If the threshold is close to 1, it is expected a smaller number of students within the clusters, while if the threshold is close to zero then the coverage of a specific cluster over the students closer to the center will increase. The method is applied using the *fcm* function from the *ppclust* library. Since the results obtained with KM already offered a good interpretation of the student's academic history groups, then the final centroids obtained with KM are used as initialization points of the groups in FCM, i.e., $\mathbf{c}_j^{[final]} = \mathbf{d}_j^{[0]}$ for all $1 \leq j \leq K$ and by using the squared Euclidean distance for both algorithms similar results are obtained in the partition of the histories. Increasing the value of the parameter m leaded to more dissimilar groups, which was not the objective.

Then, with the latter method, it has been gained the advantage of the manipulation of group membership values, however, the execution time of this method is considerably longer than the first one, Therefore, it is necessary to optimize the time by reducing the number of records that are actually necessary to obtain a grouping that is similar to the one produced when considering the total information in the database, consequently, a comparation of the results when constructing the database of academic histories with the information available in the semesters of the 10, 8, 6 and 4 years prior to the current semester is made.

2.3. Grouping Stability Validation

After performing a clustering exercise on a particular database, it is important to evaluate the robustness of the solution obtained, which can help to ensure the reliability and validity of the analysis and interpretations made. Bootstrap Clustering, whose procedure is documented in Hennig (2007), is a statistical method by which it is possible to study the stability of the groups obtained with the FCM technique, a characteristic that is highly dependent on the data set. This analysis is based on resampling with replacement and on the comparison of groups obtained in different samples with the calculation of the Jaccard index, which takes values between 0 and 1.

Henning's bootstrap method provides a systematic way to assess cluster stability by generating multiple bootstrap samples and performing clustering on each sample. By comparing the clustering results across different samples, one can gauge the stability of the identified clusters. By repeatedly resampling the data, it captures the variability in cluster assignments, which may arise due to sampling variability or noise in the data. This leads to more reliable estimates of cluster stability compared to single-sample methods. Moreover, the method is flexible and interpretable as it provides insights into the consistency of clustering patterns across different subsets of the data.

The *clusterboot* function of the *fpc* package was used to carry out the validation analysis of the clustering stability. This function performs bootstraps with replacement of the same size as the original base, and in each of the bootstraps it runs first KM and then FCM to finally calculate the Jaccard similarity index for all the groups in the original clustering against the corresponding most similar groups in the bootstrap. This procedure will allow us to obtain the distribution of the Jaccard index for each group, together with its respective mean and thus determine whether these are stable (index close to 1), or on the contrary are unstable and are affected by atypical data and/or random variation (index close to 0).

2.4. Interpretation of Clustering

The Fuzzy methodology allows for flexibility in adjusting the threshold for group membership. In characterizing groups, only academic histories with a probability of belonging to their respective class equal to or greater than 0.5 are considered. Any record falling below this threshold for any group is deemed noise and remains unclassified. The chosen threshold of 0.5 strikes a balance, allowing for clearer differences between groups in the analysis of quantitative variables characterizing academic performance. Simultaneously, the threshold considers that the proportion of students treated as noise should not be excessively large, capped at 30% of the total data. It is important to mentioned that just by increasing the parameter m to three, the value of the probabilities for each individual were more equally distributed among the groups hampering this interpretation, since there were less registers that reach the threshold of 0.5 for a specific group.

Once this filter is applied, proceed with the interpretation and understanding of the resulting groups through descriptive analysis of the original variables and where the objective is to establish the Academic Triage that will define an order of attention to the students at risk depending on the urgency that is evidenced in advising them according to their performance. Since it is desired to have the scope to offer counseling to all students who require it, the threshold of membership is moved to a lower level, such as 0.2, so that the range of coverage of the groups that represent the risk triage is extended and both students and tutors receive due warning.

2.5. Evaluation and Socialization of the System

The implementation of the Consillium Academica system culminates with the application of a survey for both students and tutors after a maximum of three weeks after the alert was sent. This survey fulfills three specific purposes: first, it allows to follow up with students and teachers to know if the meetings suggested by the alert were carried out and to evaluate the tutoring program. Secondly, it allows finding possible differences between the values of the academic performance variables that were calculated and the real values that students have in their official academic history. Lastly, feedback is received from those involved, allowing us to know their opinions, ideas and complaints regarding the alert received, and in this way to propose any type of improvement that contributes to the development and growth of the system.

3. Results

In this section the results for each of the steps included in the methodology is presented. It is important to mention that, as it will be explained in the subsection on the optimization of the algorithm execution contained in the section on the clustering technique, all the results shown below were those obtained using only the information in the databases from the first semester of 2017 to the first semester of 2023, since it was evidenced that although we are working with the most actual information of the courses taken by the students, only 5 years of registers are necessary to obtain similar groups that the ones obtained with 10 years of information.

3.1. Database Preparation

The initial database, sourced from the record office, required meticulous refinement. Duplicate records of classes taken by students across different semesters were eliminated. Additionally, records of master's students were filtered out, focusing exclusively on undergraduate students. A significant number of rows lacking numerical grades, attributed to academic exchange courses, language classes and others without reported grades, were also removed, as a grade is essential for constructing a comprehensive academic history.

Appendix A shows some graphs of the descriptive analysis performed on the databases. Figure A1(a) illustrates program-wise grade distributions, revealing a consistent pattern across curricular programs. Approximately 50% of grades fall within the 3.5 to 4.5 range, with a median around 4.0. Figure A1(b) presents box plots of grades categorized by subject typology, highlighting differences. Degree work, elective subjects, and optional disciplinary subjects exhibit higher distributions compared to other typologies. Notably, a descriptive analysis unveiled the absence of records for courses failed by undergraduate students from the first semester of 2020 to the first semester of 2022, as a response to guidelines implemented during the COVID-19 pandemic to ensure the student community's continuity.

After calculating all seven numerical variables that define the academic history of a student, a refined database was obtained. Figure A2 displays pertinent box plots, emphasizing differences in P.A.P.A. averages by curricular program (Figure A2(a)), and average semester progress percentages (Figure A2(b)). The latter reveals that up to 75% of students across programs exhibit progress below 10%, indicating a slower pace than the planned 10-semester completion. Additionally, median failed classes per student were notably low, with the biology program showing a median of zero failures, while other programs averaged one failed class. Although, most students had minimal failures, some outliers demonstrated up to 17 failed classes during the analyzed period.

We have treated the pandemic years as the other years after preprocessing (eliminating the missing and aberrant entries). It is possible to observe in the descriptive analysis that semester mean and the number of approved courses was affected due to the policies defined at Universidad Nacional de Colombia during pandemic. Nevertheless, for the seven variables created for the clustering the input was used as it was after preprocessing.

3.2. Implementation of Clustering Technique

3.2.1. Dimensionality Reduction

It was decided that the data set would be reduced to 3 dimensions considering that in the first 3 principal axes retain approximately 95% of the variance and therefore of the complete information of the database. Figure 2 displays the first factorial plane by the first two principal axes of the PCA, capturing 89% of the variance in the data set. Additionally, it incorporates the projection of the correlations of the original variables onto these two new components of the factorial plane.

The first axis in Figure 2 is interpreted as students' career advancement, highly correlated with variables such as the number of enrollments (SEMESTERS), total credits enrolled (CREDIT_T), approved credits (CREDIT_A), and the percentage of career advancement (ADVANCE_T). Conversely, the second axis is associated with academic performance, correlated with variables like P.A.P.A, the percentage of average progress per semester (ADVANCE_AV), and inversely related to the number of classes reproved (LOST), then, this axis represents students' academic success. The third axis adds information on variability in the number of courses failed and average progress per semester. In Figure 2, moving from left to right along the first axis reflects the progression of students starting their studies, while from bottom to top signifies the academic performance, with the best grades at the top and the worst grades, along with multiple failed courses, at the bottom.

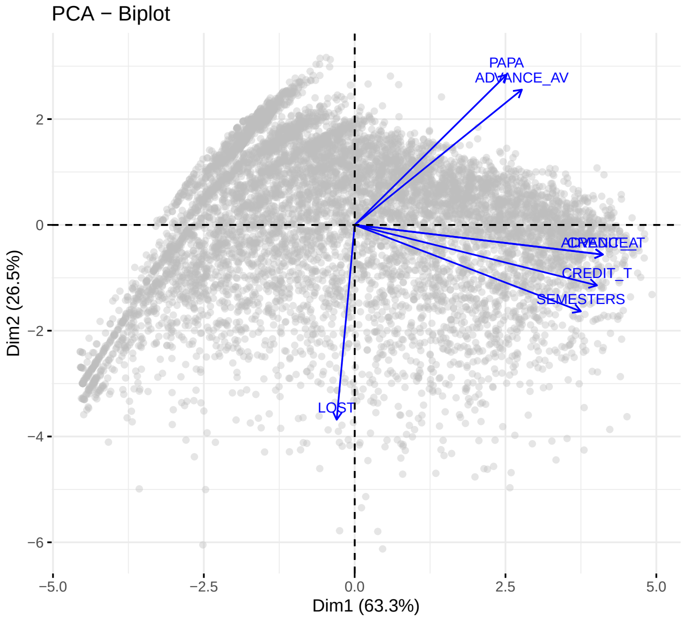


Figure 2: Factorial plane of the first two axes of the principal component analysis

3.2.2. Combination of Clustering Algorithms

The distribution of the data set in the first factorial plane (Figure 2) shows a very dense and cohesive point cloud, making it initially impossible to differentiate subsets of latent data. For this reason, the implementation of KM and FCM is appropriate, because these methods are able to identify clusters that have similar densities even though they are highly concentrated among them. The academic historie's coordinates on the principal axes of the PCA are employed for the KM technique. Utilizing the elbow method, it is determined that the optimal number of clusters is 6, as this is the point where squared errors begin to slowly reduce. The choice of 6 clusters is based on the observation that conducting descriptive analyses within these groups reveals clearer differences for the 7 numerical variables among the sets of academic histories belonging to each class.

Given the high concentration of the entire dataset, the KM method produces artificial boundaries between resulting groups. As previously mentioned, this can pose a problem where individuals needing identification for tutor advice might be located just on the other side of these boundaries. Consequently, they could be classified within a group that is not considered in the tutor search. In such cases, the KM method lacks the capability to identify these individuals effectively. In contrast, the FCM algorithm provides a solution by assigning probabilities of membership to each group for every history. This enables users to choose a threshold of membership to a certain group, offering greater flexibility in manipulating the boundaries between groups.

Since it is desired to maintain the interpretations of the classes obtained with KM, then the coordinates of the centroids of the groups are used as initialization points for the search of groups with the FCM algorithm and since the information of only 6 years of records is being considered, the execution time is around one and a half minutes. Figure 3 shows the grouping obtained with the application of this method on the first factorial plane. The group in which each academic history is assigned corresponds to the one with the highest probability of belonging.

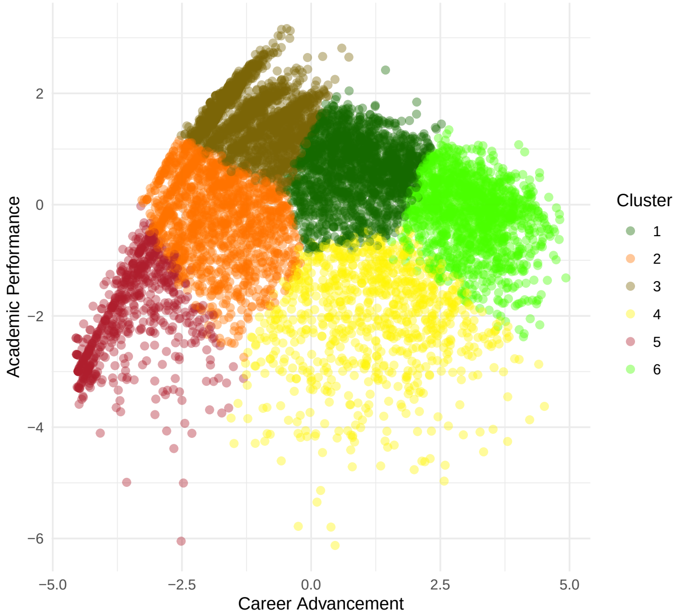


Figure 3: Fuzzy C-Means results with 6 groups on PCA coordinates

3.2.3. Clustering Optimization

With the choice of the Fuzzy C-Means technique as the most appropriate for the database, the possible optimization of the procedure to be followed to obtain the final partition in each semester was evaluated. This, taking into account that when using the total number of records obtained from the students enrolled in the last 11 years, the execution of the algorithm took more than 8 minutes to obtain the results.

To address this concern, the entire methodology was reiterated, with each iteration involving a reduction of two years in the scope of time for extracting information from the original database. The results were assessed for academic histories created from the first semester of 2013, 2015, 2017, and 2019 until the second semester of 2022. These timeframes represented data from 20, 16, 12, and 8 semesters, respectively. Additionally, the amount of academic histories for each timeframe was 11.269, 9.508, 8.404, and 6.806, respectively.

Despite the varying number of semesters considered, the first 3 axes obtained from dimension reduction consistently retained approximately 94% of the variance from the database, maintaining the explained variance proportion in different time frames indicated the correlation structure among variables is stable. As expected, a proportional reduction in clustering execution time was observed with a decrease in the scope of time. For instance, transitioning from a 10-year to an 8-year window resulted in a 6-minute reduction, followed by a 1-minute reduction for the shift from 8 to 6 years. Finally, for the database with only 4 years of records, the clustering time was reduced to 30 seconds.

Figure 4 compares the FCM results for 10 and 6 years. First, it is noteworthy that the shape of the point cloud in the PCA is very similar between the two factorial planes. It should be mentioned that there were more noticeable differences in the point cloud for the 4-year base, since many records disappeared in the lower right quadrant, considering this sector is characteristic of students who have made great progress in their careers but have low academic performance, Therefore, they require a higher number of enrollments to graduate, and by reducing the number of semesters for which data are collected, the maximum values of the progress variables for these students are greatly limited.

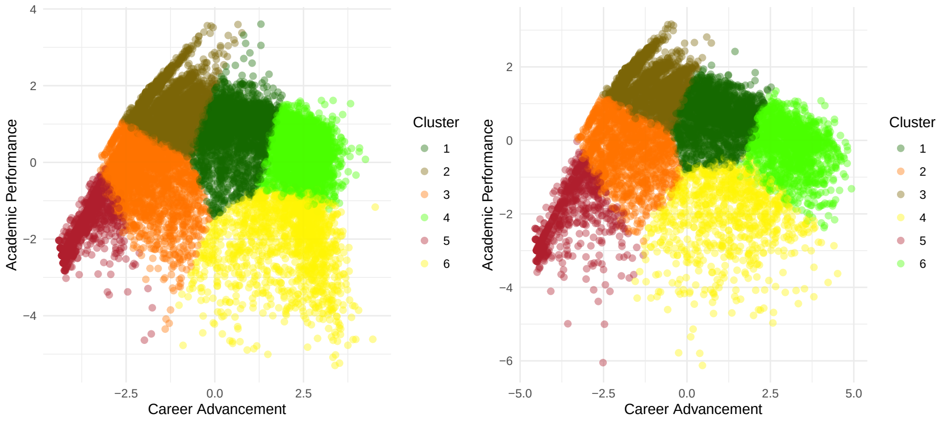


Figure 4: FCM results with 10 years of recordings(left) and with 6 years of recordings (right)

In addition, with the help of subfigures (a) and (b) it can be seen how the distribution of the two datasets in the first factorial plane are very similar, then the shape and location of the resulting groups by the FCM method do not suffer significant changes, and the interpretation given to each of them either. All these findings suggested the feasibility of selecting a smaller number of records without significantly altering clustering outcomes. It was also necessary to consider that if the 4-year base was chosen, then a low quantity of students would have finished their professional careers in this time, after all most students take longer than planned to complete their studies. Thus, it is more useful to reduce the extraction of information to 6 years, so that the range coincides with the period that most undergraduate students need to finish their degree studies (between 10 and 13 semesters).

3.3. Clustering Stability Validation

The following results correspond to the Bootstrap Clustering analysis performed on the 6 groups determined before and with the execution of 100 resamplings in a time of approximately 3 hours. Figure 5 shows the histograms corresponding to the Jaccard index values obtained in the 100 resamplings for each group, where it is observed that all the distributions are located between 0.85 and 1. In addition, the vertical line showed in each histogram represents the average of the index in that class, so it is observed that in all cases the average of the index is higher than 0.95, which indicates a high stability in the conformation of the groups.

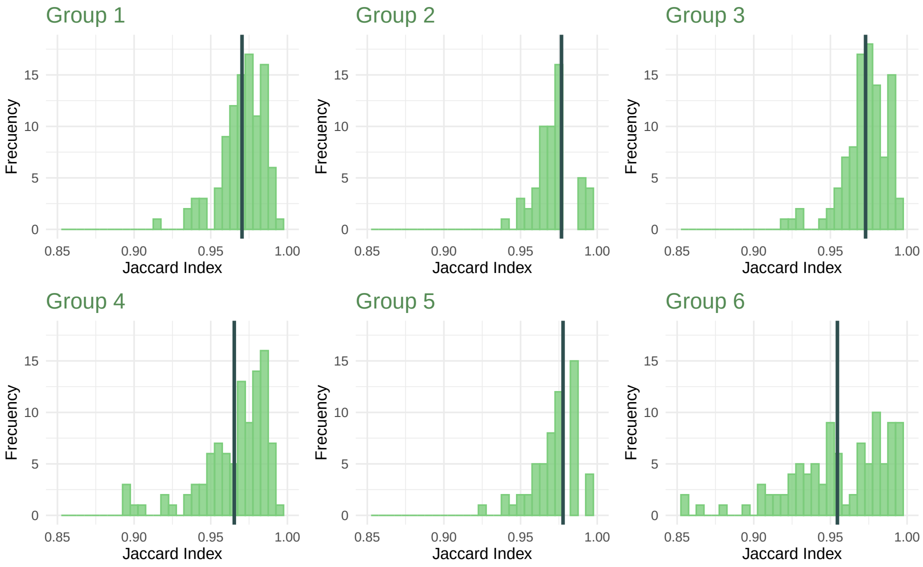


FIGURE 5: Bootstrap Clustering results for the 6-year base with 100 resamples.

All the distributions show great stability in the execution of the Fuzzy C-Means method on this database and it is not evident that the result of the partition can be influenced by atypical data or variability due to the random points of initialization of the algorithm, this means that the results obtained from the Consillium Academica methodology are solid, accurate and guarantee validity.

3.4. Interpretation of Clustering

3.4.1. Academic Triage

Only academic histories surpassing a probability threshold of 0.5 for belonging to any group are chosen, while those falling below this threshold are not classified and are regarded as noise in this instance. Figure 6 presents the clustering result taking into account this adjustment, the gray dots correspond precisely to the noise data which are labeled as cluster 0, comprising 26.24% of the total records. As expected, from the way FCM finds the clusters the academic histories closest to the centroids are the ones with the highest probabilities of belonging to the clusters.

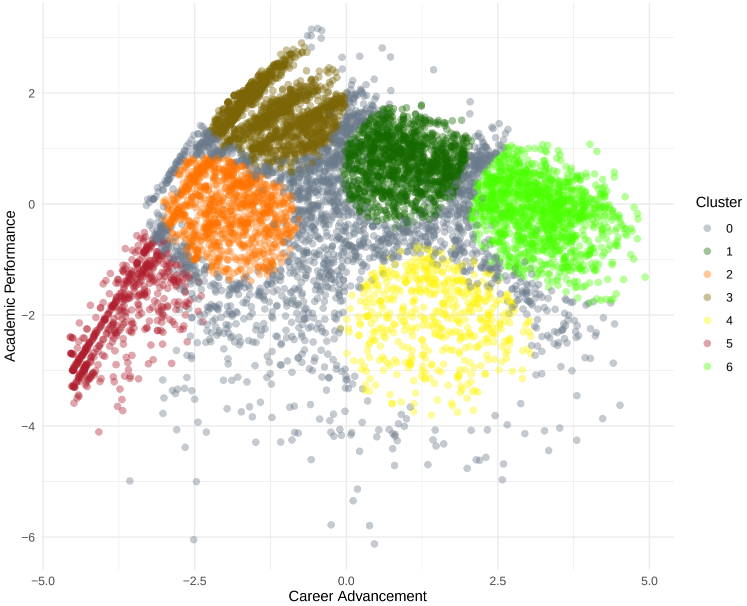


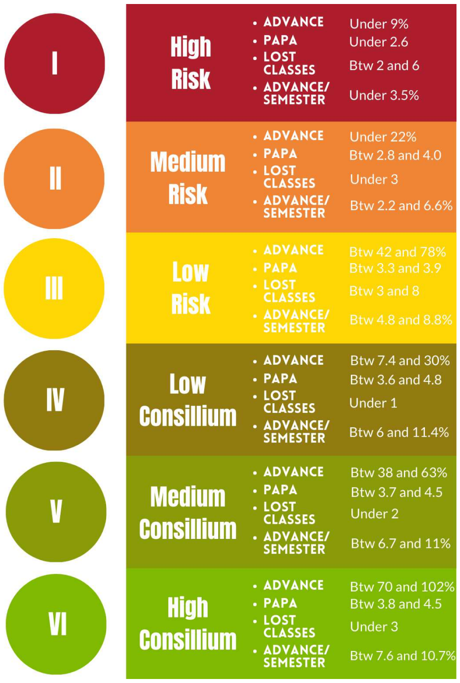
FIGURE 6: Clustering result when modifying the membership threshold to 0.5

Finally, the characterization of the groups allows the creation of the Academic Triage that determines the urgency of attention required by each active undergraduate student in the Faculty of Sciences. The triage is composed of 6 levels, the first 3 correspond to students at risk of dropping out the curricular program or expulsion from the university, while the next 3 correspond to students who are performing well and are called Consillium groups (Guidance in Latin). Figure 7 shows the resulting Academic Triage scheme, with the respective names of each level and the values of the 5th and 95th quantiles for the variables of advance in the career, P.A.P.A., number of failed classes and the percentage of progress per semester average. In case it is desired to match the triage levels with the groups shown in the Figure 6, the order of the clusters is presented in the Table 1.

TABLE 1: Matching the triage levels with the labels of the clusters

Risk groups	Consillium groups
I - cluster 5	IV - cluster 3
II - cluster 2	V - cluster 1
III - cluster 4	VI - cluster 6

In summary, the triage levels are determined based on the student's academic performance, considering P.A.P.A. and the number of failed subjects, along with variables related to progress in the curricular program. Triage I comprises active students needing urgent attention, mostly consisting of those who entered the university but failed many subjects, leading to expulsion due to a P.A.P.A. below 3.0. Triage II is the next in order of attention, it includes students starting their studies with low academic averages, relatively slow progress, and few failed subjects. Triage III identifies students at academic risk who have made slow but significant progress, maintaining an average grade close to 3.0 and failing several subjects.



I	High Risk	• ADVANCE • PAPA • LOST CLASSES • ADVANCE/ SEMESTER	Under 9% Under 2.6 Btw 2 and 6 Under 3.5%
II	Medium Risk	• ADVANCE • PAPA • LOST CLASSES • ADVANCE/ SEMESTER	Under 22% Btw 2.8 and 4.0 Under 3 Btw 2.2 and 6.6%
III	Low Risk	• ADVANCE • PAPA • LOST CLASSES • ADVANCE/ SEMESTER	Btw 42 and 78% Btw 3.3 and 3.9 Btw 3 and 8 Btw 4.8 and 8.8%
IV	Low Consillium	• ADVANCE • PAPA • LOST CLASSES • ADVANCE/ SEMESTER	Btw 7.4 and 30% Btw 3.6 and 4.8 Under 1 Btw 6 and 11.4%
V	Medium Consillium	• ADVANCE • PAPA • LOST CLASSES • ADVANCE/ SEMESTER	Btw 38 and 63% Btw 3.7 and 4.5 Under 2 Btw 6.7 and 11%
VI	High Consillium	• ADVANCE • PAPA • LOST CLASSES • ADVANCE/ SEMESTER	Btw 70 and 102% Btw 3.8 and 4.5 Under 3 Btw 7.6 and 10.7%

FIGURE 7: Academic Triage for the order of attention to undergraduate students of the Faculty of Sciences

The triages named Consillium are interpreted as the counterpart of the first 3 already defined, these groups contain students who have a very good academic performance, since they have failed few or no subjects, in general, the P.A.P.A. is above 3.6 and their progress per semester is within the expected. The difference between the three triages lies in the number of semesters completed and the progress in the respective curricular program, i.e., Triage IV contains students who are just beginning their studies, Triage V has students with progress percentages around 50% and, finally, Triage VI represents students who are close to completing their studies or who have already finished them.

Finally, students who are located in Triage I, II and III are contacted by e-mail, together with their respective tutors to encourage meetings between them to guide and motivate these students to remain in their respective curricular program. However, for Triage II and III additional filters are made on the students who receive the mail, for Triage II only those with a P.A.P.A. lower than 3.4 are selected and for Triage III the same condition is taken into account, but additionally that their percentage of progress in the career is less than 80%.

3.5. Evaluation and Socialization of the System

Three weeks after the mailing was sent to students and tutor teachers, a survey was shared with them to follow up on the suggested meetings and receive feedback from them. A different survey for students and for tutors was proposed and the most relevant results obtained with these are presented in the following sections.

3.5.1. Students Survey

The survey, disseminated to 605 students who received email alerts from Consillium Academica, garnered 100 responses, resulting in a response rate of 16.52%. With 22 questions, the questionnaire featured both general inquiries and tailored questions contingent on whether the student had met with their tutor.

First, it is necessary to corroborate the accuracy of academic history information derived solely from class records. Students were queried about the congruence between the P.A.P.A. reported by the system and the data in the Academic Information System (SIA, for its Spanish acronym). The results, shown in Figure A3(a) in Appendix B, indicate that, for 24% of the students, their averages did not coincide, practically a quarter of those surveyed. Upon further inquiry, these students were asked for their official P.A.P.A. It was observed that, for some of them, differences of up to one full grade unit existed. Upon reviewing the individual records of the classes for these students, it was determined that, by the time the emails were sent, changes might have been made in the grades of reproved courses or courses canceled under a faculty council decision. Consequently, these updates might not have been included in the database used by the system.

Figure A3(b) illustrates a pie chart representing student communication with their respective tutors. Alarmingly, only 20% reported successful contact. Coupled with the fact that merely 37% of students knew their tutor's identity, it underscores deficiencies in the tutor program implementation. Despite these challenges, 74% of students found the alerts from Consillium Academica beneficial.

Notably, students who sought counseling expressed satisfaction, citing genuine interest and effective assistance from their tutors. Among these, 75% had sought help from other university resources. Conversely, students without tutor contact elucidated reasons for the absence of meetings (Figure A4). Primarily, tutors' lack of initiation or responsiveness ranked as the leading cause, followed by students' insufficient confidence to initiate contact. Strikingly, 91.3% of these students did not seek academic support from department or faculty welfare services.

3.5.2. Tutor Survey

The survey was distributed to 249 tutor professors who received email alerts from Consillium Academica, providing them with a list of students to be contacted. By the closing time, 38 responses were obtained, resulting in a response rate of 15.26%. The questionnaire comprised 13 questions, focusing on the professors' experiences as tutors and their perceptions of the system, considering the variable number of students assigned to each professor.

The pie chart in Figure A5(a) reveals that 89.5% of these teachers view the system's promotion of meetings between at-risk students and their respective tutor teachers as a positive initiative, showcasing a willingness to assist students. However, Figure A5(b) indicates that 26.3% of these teachers admit to feeling inadequately equipped with the necessary knowledge to fulfill the role of a tutor, emphasizing the need for training to enhance their effectiveness in this capacity.

Professors were further queried about the types of problems encountered when advising students. Results, presented in Figure A6, indicate that, according to their perceptions, personal situations ranked first, followed by economic situations, the student's study habits, and motivation issues. In related questions, some professors expressed attempts to contact students listed in their respective rosters but reported no response from them.

4. Discussion

This work shows the development of an application facilitating the clustering of undergraduate students within the Faculty of Sciences at Universidad Nacional de Colombia based on their academic performance. Utilizing Principal Component Analysis, K-Means, and Fuzzy C-Means techniques, the clustering process not only hierarchically identified students at risk but also distinguished groups with varying academic characteristics. The application, implemented through two Rmarkdown notebooks, allows for semi-automatic grouping of academic histories, providing instructions available on this Github.

The hierarchical clustering approach, along with Fuzzy C-Means manipulation capabilities, offered a nuanced classification. It not only prioritized attention to at-risk students across three distinct groups but also identified three other groups of students with good academic performance exhibiting diverse progress percentages in their curricular programs. Furthermore, the Bootstrap Clustering procedure validated the stability of the methodology, showcasing robust results despite to the input of new academic records and minor database changes.

Survey results indicated that both students and professors found the system-generated alerts beneficial, promoting valuable contact between students and their designated tutors. However, only 20% of surveyed students reported meeting with their tutors. The remaining respondents expressed reluctance, citing a lack of contact initiation or insufficient confidence. To address this, it is recommended that future iterations of the Consillium Academica system consolidate alert emails to both students and tutors in a single communication. This adjustment aims to enhance transparency and information accessibility, potentially fostering increased contact and collaboration between students and their tutors, aligning with the overarching goal of the tutoring program as outlined in the Academic Council's Agreement 028 of 2010. Additionally, it is important to work with the recent and updated grades from previous semesters.

Despite previous work on analysis of academic data to identify students at risk such as Sarra et al. (2019) and Roldán Jiménez (2021), this application is novel for the Universidad Nacional de Colombia and has proved to be useful and easily applicable. The implementation process and feedback every semester is necessary to make adjustments and update the application.

Acknowledgements

We extend our sincere appreciation to Maryory Rojas, head of the Registration Division Bogota Headquarters, for providing the invaluable data sources. Additionally, we are grateful for the support and collaboration of the Faculty of Sciences at the Universidad Nacional de Colombia. Special thanks go to Professor Luis Ochoa for his initiative and creativity in spearheading this project. Furthermore, we would like to express our gratitude to all the professors and students for their valuable feedback and contributions throughout this endeavor.

[Received: February 2024 — Accepted: June 2024]

References

Hennig, C. (2007), 'Cluster-wise assessment of cluster stability', *Computational Statistics and Data Analysis* **52**(1), 258–271.

Nithya, P., Umamaheswari, B. & Umadevi, A. (2016), 'A survey on educational data mining in field of education', *International Journal of Advanced Research in Computer Engineering & Technology* **5**(1), 69–78.

Pardo, C. (2020), *Estadística descriptiva multivariada.*

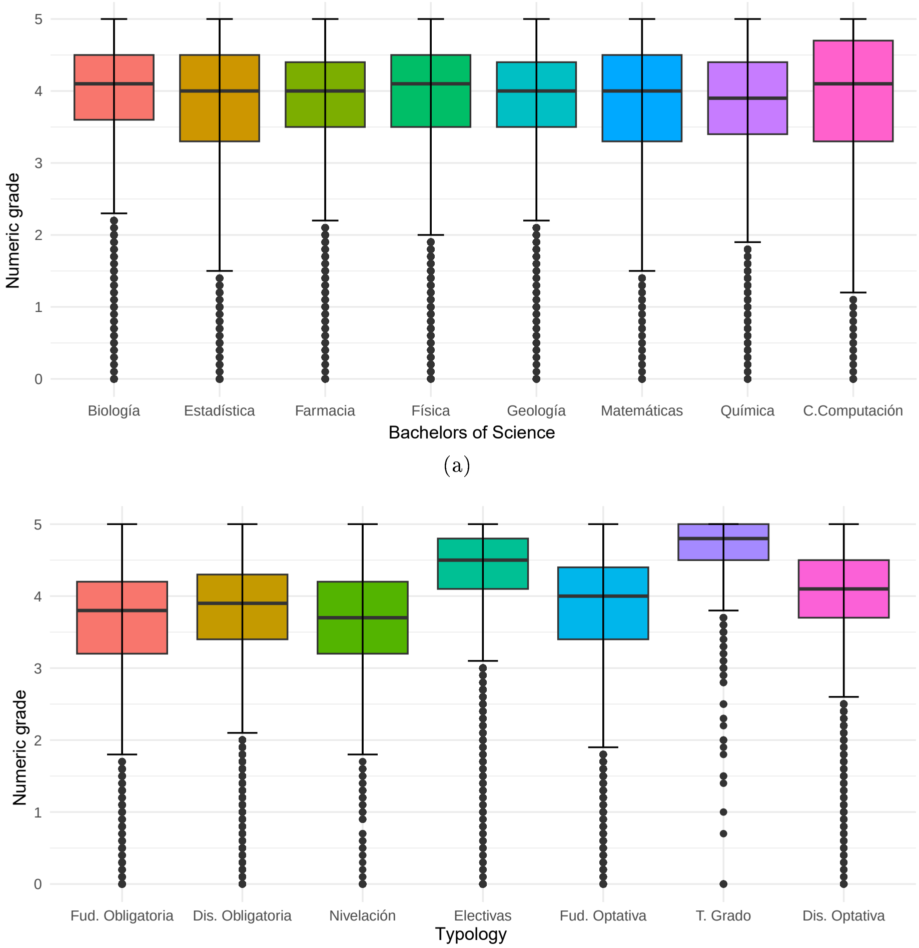
Roldán Jiménez, L. S. (2021), 'Machine learning to predict student academic risk in engineering', *Encuentro Internacional de Educación en Ingeniería* . https://acofipapers.org/index.php/eiei/article/view/1579

Sarra, A., Fontanella, L. & Zio, S. D. (2019), 'Identifying students at risk of academic failure within the educational data mining framework', *Social Indicators Research* **146**(1/2), pp. 41–60. https://www.jstor.org/stable/48704856

Ulriksen, L., Madsen, L. & Holmegaard, H. (2010), 'What do we know about explanations for drop out/opt out among young people from stm higher education programmes?', *Studies in Science Education* **46**, 209–244.

Vendramin, L., Naldi, M. C. & Campello, R. J. G. B. (2015), *Fuzzy Clustering Algorithms and Validity Indices for Distributed Data*, Springer International Publishing, Cham, pp. 147–192.

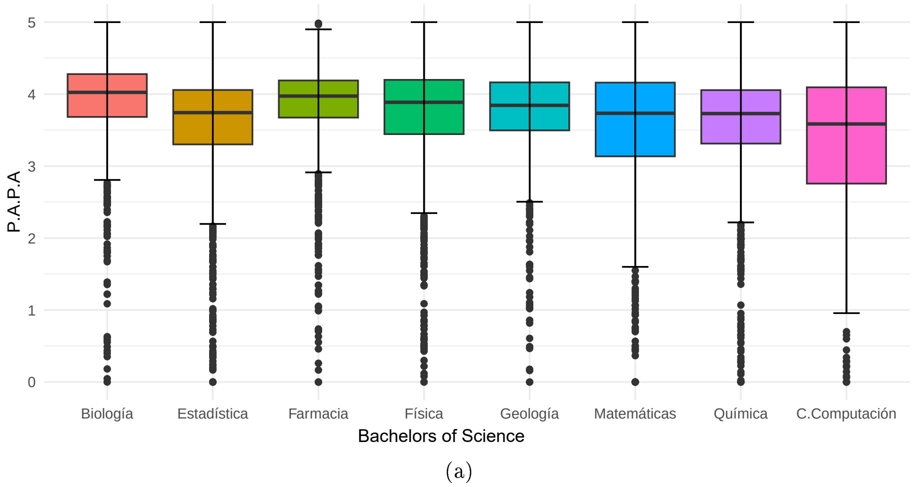
Appendix A. Descriptive Analysis



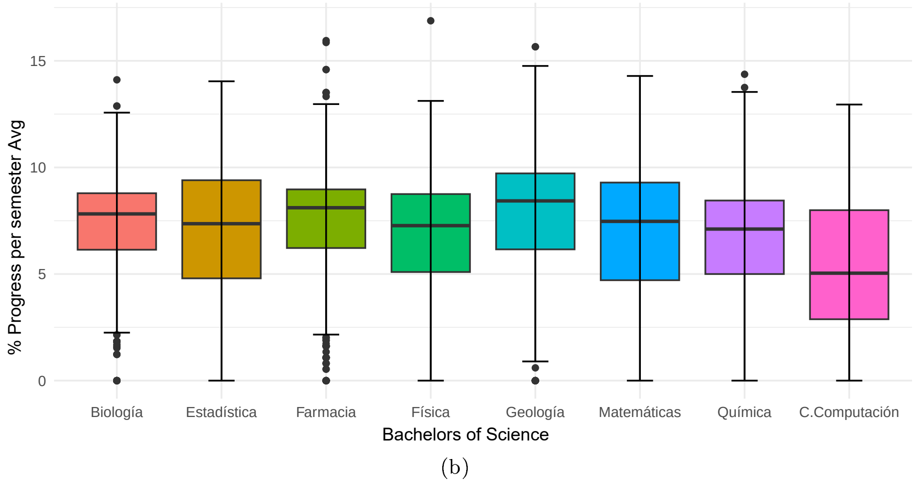
(a)

(b)

FIGURE A1: Box plots for the numerical grades obtained in each course by Bachelor of Science (a) and course's typology (b)



(a)



(b)

Figure A2: Box plots for some variables present in the academic history database by Bachelor of Science

Appendix B. Survey's Results

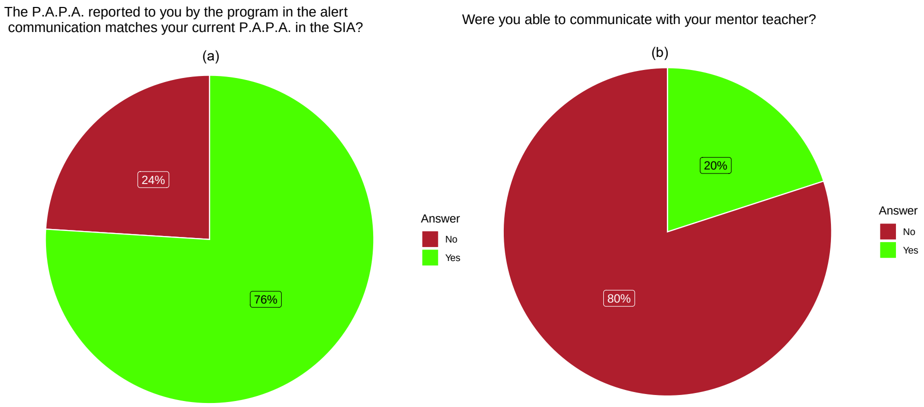


FIGURE A3: Pie charts for some of the general questions made to students

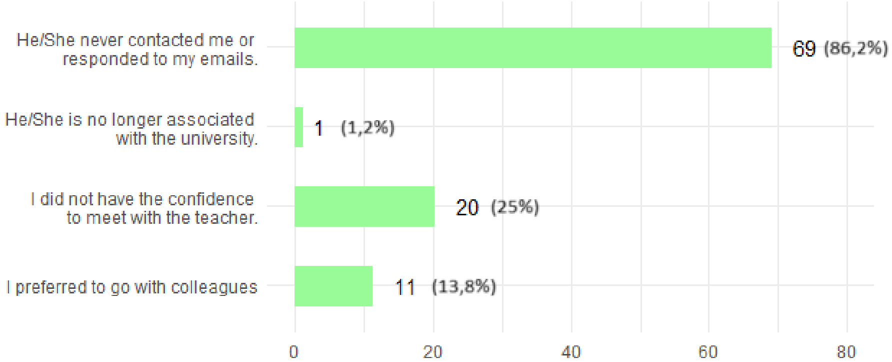


FIGURE A4: Result for the question: What was the reason for not communicating with your tutor?

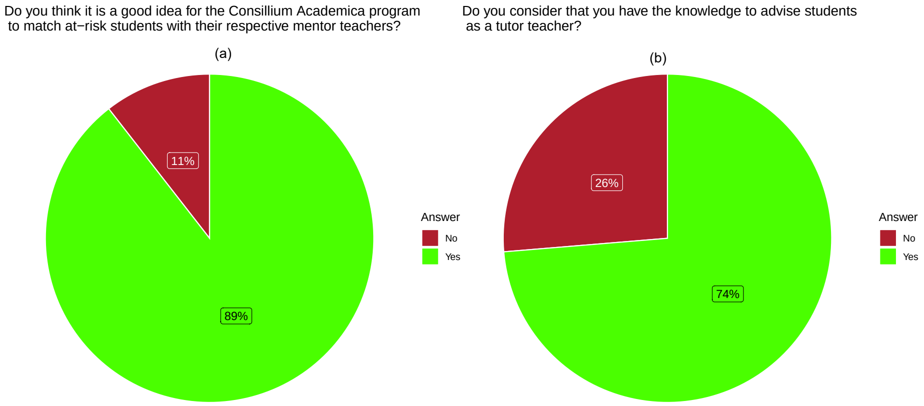


Figure A5: Pie charts for some of the questions asked to the tutors

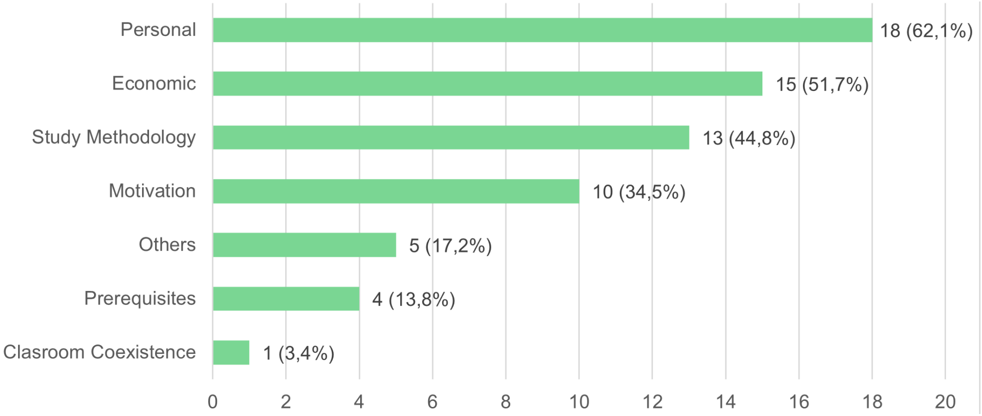


Figure A6: Result for the question: Considering the advisories given to all the tutored students, in general, what type of problems were evidenced that were affecting the academic performance of the tutored students at risk