Analysis on Early Dropouts in Engineering Careers

Ignacio Manes*, Tomas Lubertino, Jorge Anca, Karen Roberts and Hernán Merlino

Universidad de Buenos Aires, Buenos Aires, Argentina

Abstract

Low graduation rates, careers that extend longer than normal and university dropouts are a reality in Universities in Argentina. The objective of this work is to analyze, from the perspective of data sciences, the early dropouts in the field of the Faculty of Engineering of the University of Buenos Aires, in the context of the careers of Computer Engineering and Bachelor of Systems Analysis.

Keywords

University Dropouts, Higher Education Dropouts, Computer Science Careers, Data Science

1. Introduction

The present work takes as its main focus the careers of Computer Engineering and Bachelor of Systems Analysis that are taught at the Faculty of Engineering of the University of Buenos Aires, which historically register high levels of student dropouts that cause low graduation rates and an increase in the average time to degree.

2. State of the art

This section summarizes the work of some academic research projects which try to explain the phenomena of higher education students' dropout from different models.

Some explain these phenomena based on two sociological theories: "The student integration model" [1, 2] where the integration of the student into the academic world directly affects the determination of whether or not to drop out of school, another is the "Student attrition model" [3] that gives relevance to factors external to the educational institution.

According to [4] the relevance given to the variables that try to explain the phenomenon of dropout and retention, whether family, individual or institutional, it addresses different dimensions of analysis: Psychological, Economic, Sociological, Organizational and Interaction.

In a report of the Argentine Ministry of Education [5], which included a total of 21 unified terminals of the discipline according to CONFEDI (Federal Council of Deans of Engineering), focusing on the Informatics/Systems career of public institutions, it was possible to carry out

ICAIW 2022: Workshops at the 5th International Conference on Applied Informatics 2022, October 27–29, 2022, Arequipa, Peru

*Corresponding author

imanes@fi.uba.ar (I. Manes); tlubertino@fi.uba.ar (T. Lubertino); janca@fi.uba.ar (J. Anca); kroberts@fi.uba.ar (K. Roberts); hmerlino@fi.uba.ar (H. Merlino)

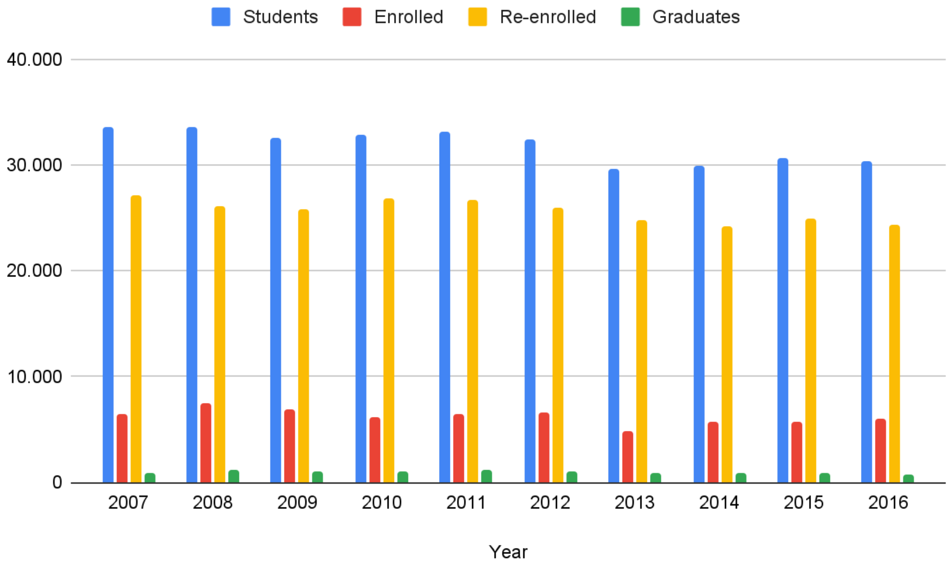


Figure 1: Students, Newly Enrolled, Re-enrolled, and Graduates of Computer Science/Systems Engineering per year.

a series of graphs showing the evolution of the number of students between 2007 and 2016 (Figure 1).

Figure 2 presents the evolution of engineering degree graduates is observed. It is evident to note how the total number of students in the program (Figure 1), as well as the number of

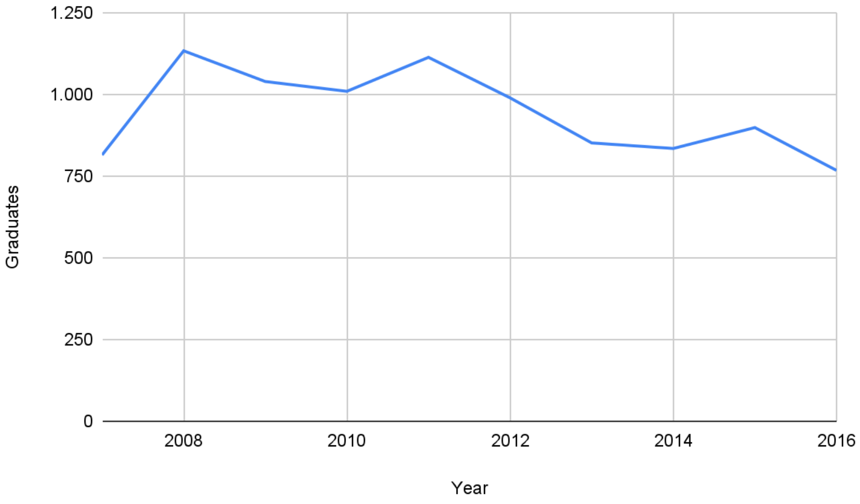


Figure 2: Computer Science/Systems Engineering graduates over the years.

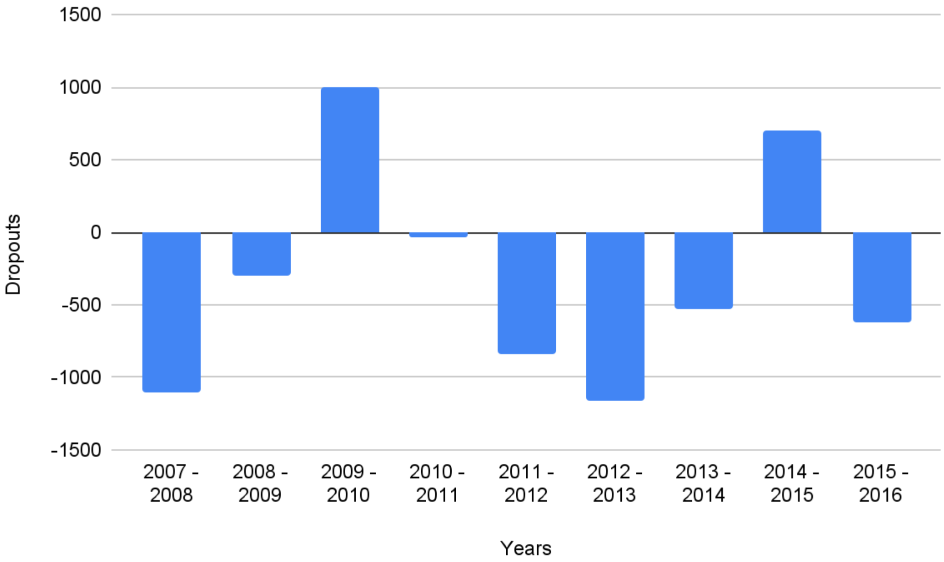


Figure 3: Dropout of Computer Science/Systems Engineering students year by year.

graduates (Figure 2) decreased over the years. Where in addition, the percentage of the latter was always very low in relation to the total (average of less than 1000 graduates per year).

Finally, based on the number of students re-enrolled in this period, it was grouped by consecutive years to see the difference between them (Figure 3) and thus calculate how many students allegedly dropped out of the degree. For example, in 2007 there were a total of 27,179 re-enrolled students, while in 2008 the total was 26,079, so the difference was 1,100 fewer students.

The goal of this work is to detect the early dropout in the Systems Analysis and Computer Engineering degree courses. We will try to obtain the behavior patterns of the myRPL.ar platform database[1] through the use of information exploitation processes.

3. Analysis

An automatic data analysis was carried out from the database of the myrpl.ar platform, using Python[2], Jupyter[3] notebooks and the Pandas Profiling library[4]. We concluded that the tables that were not included in the query that builds the dataset, didn't have relevant information to carry out the early dropout analysis. In this way, the following tables were chosen:

- *activities:* Provides information on the activities carried out by the students.

[1] https://myrpl.ar/
[2] https://www.python.org/
[3] https://jupyter.org/
[4] https://github.com/ydataai/pandas-profiling

- *activitiy_submissions:* Relates each of the activities with the submissions made by the students.
- *course_users:* Provides information about all the courses in which the student is or was enrolled.

From the discarded tables, the rpl_files_report table is included, which has the code of each of the student submissions. We run a linter to obtain a score for each of the student's submissions, but this option was discarded since the activitiy_submissions table contains a status of the submissions (failure, build_error, success, runtime_error, time_out) which was according to the teacher's criteria.

A dataset was built from the information obtained that contains data about the student, the semester of the subject, the code of the subject, and all the deliveries for the different tasks with their respective status. Within the first EDA (Exploratory Data Analysis) carried out, it was found that the data obtained belonged to 10 different careers.

To complement the analysis made with Pandas Profiling, the D-tale library[5] was used to analyze the predictive power of each of the variables and see which models the library recommended. To this analysis, the application of several clustering algorithms was added to find a common pattern among the data that could mark a tendency to leave one or several subjects. Using the PyCaret[6] Automated Machine Learning library, the following models were run:

- kmeans
- meanshift
- sc
- hclust
- birch

[5] https://pypi.org/project/dtale/
[6] https://pycaret.org/

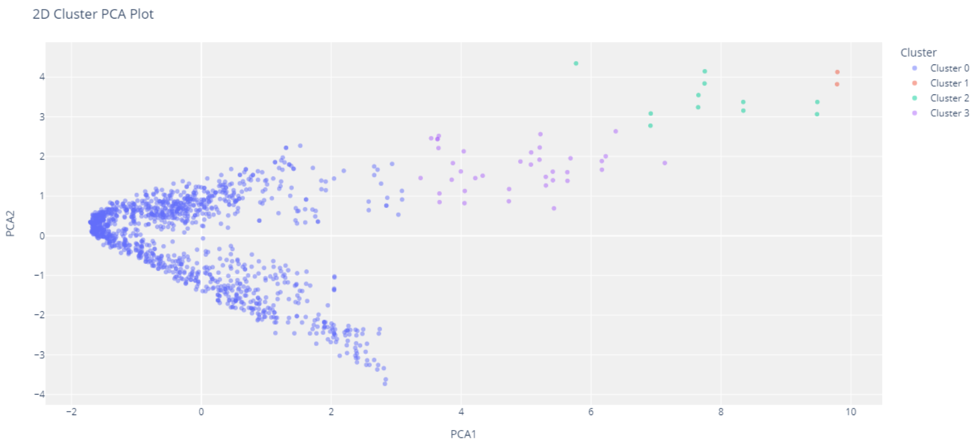


Figure 4: 2D Spectral Clustering PCA.

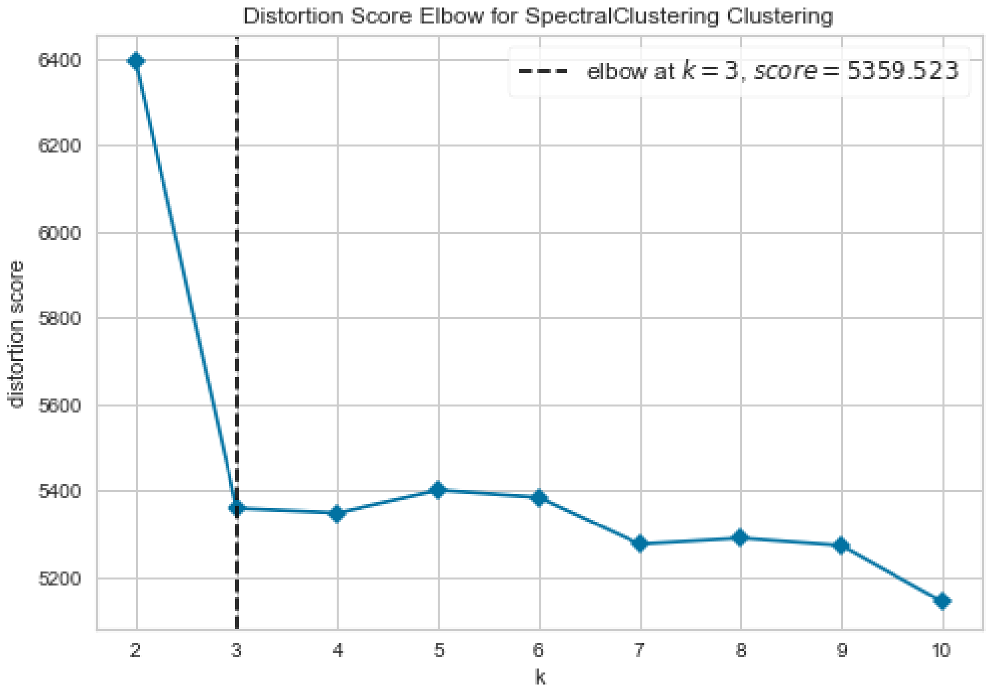


Figure 5: Distortion Score Elbow for the Spectral Clustering Model.

The model that had the best performance was the Spectral Clustering model with a much higher silhouette score than the other models, resulting in the following clustering (Figure 4)

As it can be seen in Elbow (Figure 5) the ideal number of clusters is 3. By analyzing the groups formed, did not notice any pattern that shows desertion by the students.

After this stage, an anomaly detection model was applied to emulate the dropout variable that could not be obtained. To address this issue, we used the PyCaret model for anomaly detection. Three different types of models were tested:

- knn
- iforest
- svm

This last model was the only one that agglomerated a group of students in almost a single point in space, in relation to the dimension reduction variables used by PyCaret.

As can be seen in the 3D TSNE (Figure 6), the anomalous points are clustered in a single region.

The same thing happens in the uMAP (Figure 7), showing the same result for two dimensions.

Two classification models were applied to the resulting dataset, each one from a different library. The first one was the PyCaret classifier where the random forest was the most performant model (Figure 8).

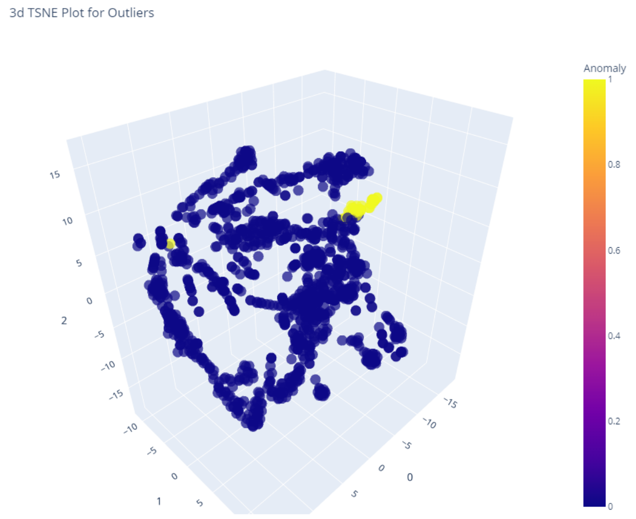


Figure 6: 3D TSNE for Outliers (svm).

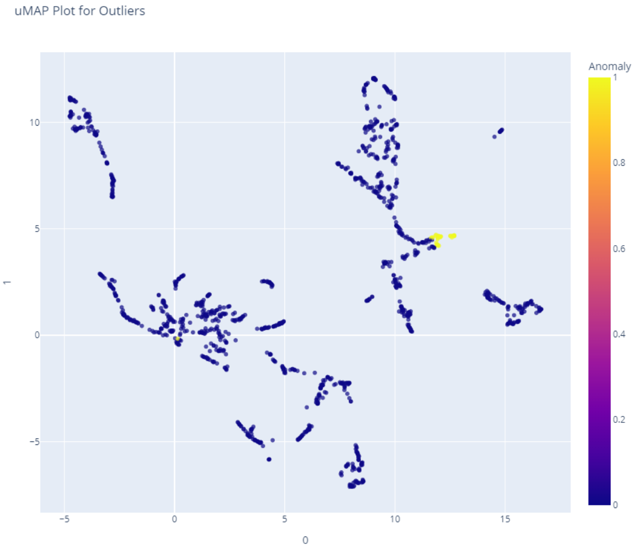
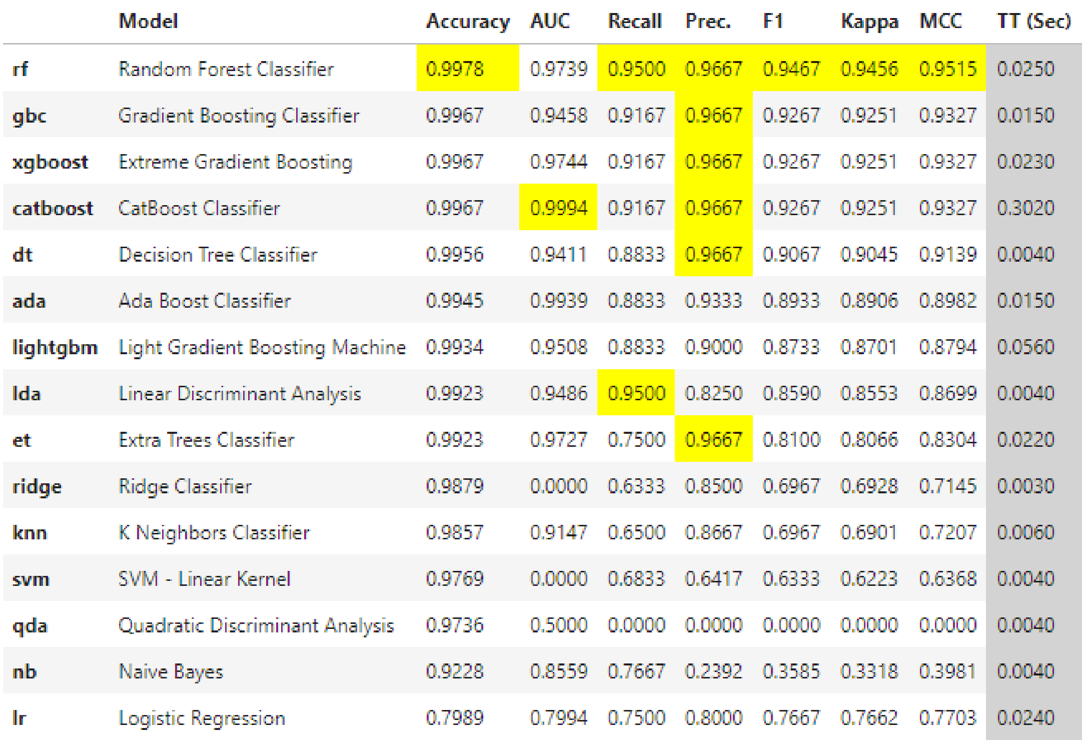


Figure 7: uMAP for Outliers (svm).

Similarly, the confusion matrix (Figure 9). shows the results for this unbalanced dataset, with only one false positive.

On the other hand, it can be seen (Figure 10) that the FAILURE variable is extremely important for the prediction.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9978	0.9739	0.9500	0.9667	0.9467	0.9456	0.9515	0.0250
gbc	Gradient Boosting Classifier	0.9967	0.9458	0.9167	0.9667	0.9267	0.9251	0.9327	0.0150
xgboost	Extreme Gradient Boosting	0.9967	0.9744	0.9167	0.9667	0.9267	0.9251	0.9327	0.0230
catboost	CatBoost Classifier	0.9967	0.9994	0.9167	0.9667	0.9267	0.9251	0.9327	0.3020
dt	Decision Tree Classifier	0.9956	0.9411	0.8833	0.9667	0.9067	0.9045	0.9139	0.0040
ada	Ada Boost Classifier	0.9945	0.9939	0.8833	0.9333	0.8933	0.8906	0.8982	0.0150
lightgbm	Light Gradient Boosting Machine	0.9934	0.9508	0.8833	0.9000	0.8733	0.8701	0.8794	0.0560
lda	Linear Discriminant Analysis	0.9923	0.9486	0.9500	0.8250	0.8590	0.8553	0.8699	0.0040
et	Extra Trees Classifier	0.9923	0.9727	0.7500	0.9667	0.8100	0.8066	0.8304	0.0220
ridge	Ridge Classifier	0.9879	0.0000	0.6333	0.8500	0.6967	0.6928	0.7145	0.0030
knn	K Neighbors Classifier	0.9857	0.9147	0.6500	0.8667	0.6967	0.6901	0.7207	0.0060
svm	SVM - Linear Kernel	0.9769	0.0000	0.6833	0.6417	0.6333	0.6223	0.6368	0.0040
qda	Quadratic Discriminant Analysis	0.9736	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0040
nb	Naive Bayes	0.9228	0.8559	0.7667	0.2392	0.3585	0.3318	0.3981	0.0040
lr	Logistic Regression	0.7989	0.7994	0.7500	0.8000	0.7667	0.7662	0.7703	0.0240

Figure 8: Analysis of models from the PyCaret library applied to the synthetic variable.

The other model used was generated with the Teapot library[7]. It produced a pipeline that includes a logistic regression, and like PyCaret, a random forest. The metrics Figure 11) show excellent results, just like the first model.

All the notebooks used by this paper can be found on github [6].

Finally, a dashboard[8] was developed using Next.js[9] for the frontend and Vercel[10] for cloud deployment, in order to give visibility to the data and reflect the problems mentioned above. It shows the data table used for the present work together with some graphs that helps to see the distribution of students and the types of events generated by the course.

4. Conclusions

As a conclusion of the process carried out, it can be seen how data science helps to answer the research question if it is possible to detect student dropout, which implies low graduation rates. Here, hints have been found about the existence of indicators such as the FAILURE characteristic, which represents the number of times that the student fails in the assignment and

[7] http://epistasislab.github.io/tpot/

[8] https://dashboard-tp-profesional.vercel.app/

[9] https://nextjs.org/

[10] https://vercel.com/

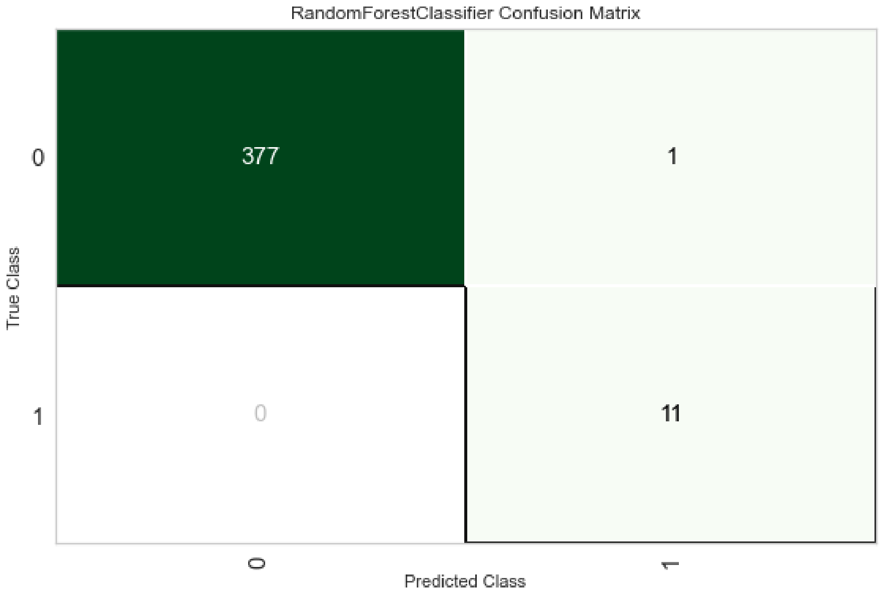


Figure 9: Confusion matrix

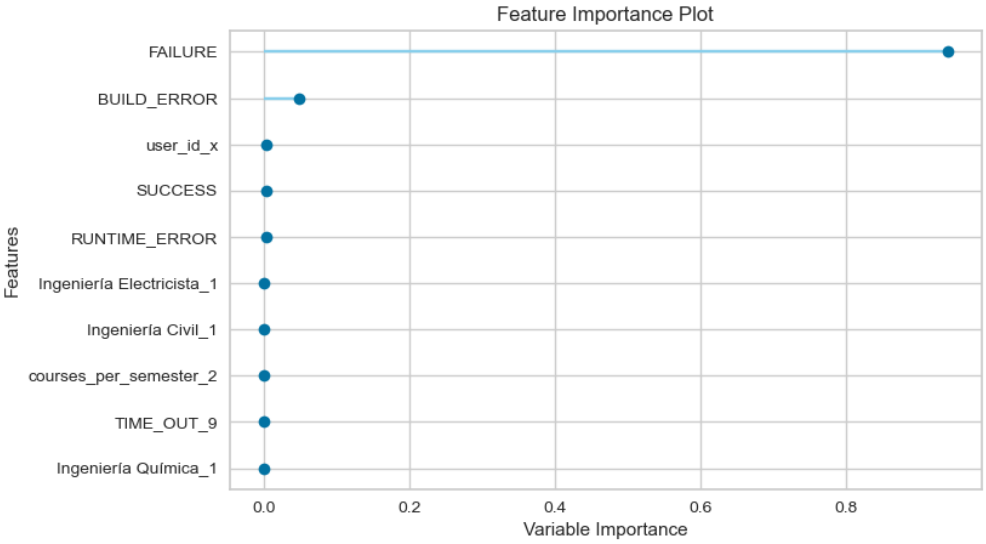
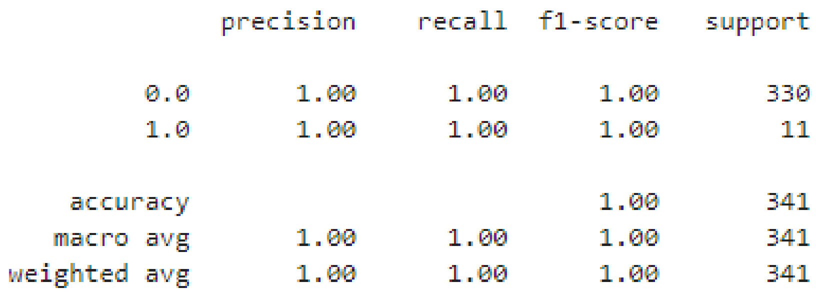


Figure 10: Variable's importance for the pycaret classification model

would represent a good predictor in this first instance of the ongoing investigation, allowing teachers to take actions in early stages to avoid students dropout. The ongoing research allows us to start detecting alternative characteristics that are good predictors of university dropout. This line of research will continue in order to find new predictive characteristics.



```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00       330
         1.0       1.00      1.00      1.00        11

    accuracy                           1.00       341
   macro avg       1.00      1.00      1.00       341
weighted avg       1.00      1.00      1.00       341
```

Figure 11: Teapot Ranking Model Metrics

5. Future work

As next steps to obtain a model that reflects the reality of the students, the following datasets should be included:

- Data from the SIU Guaraní[11] to obtain the dropout variable to be predicted.
- Data of all the careers of the University, to extend the analysis carried out in computer science subjects to the subjects of all the careers.
- Student satisfaction statistics.

Having acquired the new data will require iterating on the model made to confirm that it fits the newly added variables. Models of early dropout of the University could be implemented, as well as performance and dropout by subject, making it easier to reinforce the monitoring of students who are at risk of any of the situations mentioned previously.

6. Acknowledgments

The authors of this work acknowledge the Faculty of Engineering, University of Buenos Aires. This work is framed within the PIDAE "Intelligent systems in university management".

References

[1] W. G. Spady, Dropouts from higher education: An interdisciplinary review and synthesis, Interchange 1 (1970) 64–85.

[2] V. Tinto, Dropout from higher education: A theoretical synthesis of recent research, Review of educational research 45 (1975) 89–125.

[3] J. P. Bean, Student attrition, intentions, and confidence: Interaction effects in a path model, Research in higher education 17 (1982) 291–320.

[4] J. Braxton, Reworking the student departure puzzle, Vanderbilt University Press, 2000.

[11] https://guaraniautogestion.fi.uba.ar/g3w/acceso

[5] Ministerio de Educacion de Argentina, Informe especial: Estudiantes, nuevos inscriptos, reinscriptos y egresados de ingeniería, 2007-2016.
[6] I. Manes, Desgranamiento temprano en las carreras de ingeniería, 2022.