Classification and prediction model to profile student success variables

*Note: Sub-titles are not captured in Xplore and should not be used

Rodrigo Rodriguez de Luna
Tec de Monterrey
Monterrey, Mexico
A01384318@tec.mx

Jackeline Conant Rubalcava
Tec de Monterrey
Monterrey, México
A01280544@tec.mx

Oscar Ariel Ortega Franco
Tec de Monterrey
Monterrey, México
A00833051@tec.mx

Luis Alberto Portilla López
Tec de Monterrey
Monterrey, Mexico
A00829935@tec.mx

***Abstract*—**A classification and prediction model trained to differentiate student profiles. This model takes in various significant variables to classify and predict the different types of student profiles**.**

In this case, the most relevant variables are analyzed to support decision-making regarding the classification of students in the Líderes del Mañana program.

***Keywords*—**classification, prediction, graduated, artificial intelligence**.**

Introduction (*Heading 1*)

la introduction y abstrac se competaran al completar las demas seciones del paper

Artificial intelligence has created a significant impact on the education students receive today, from the automation of daily activities to grading exams and helping students improve their efficiency. Turning into a controversial topic, AI has been considered as having a high risk on education by the European Union, since many point to the risks and deficiencies it has. However, UNESCO explains that not everything is negative about the use of AI, since it has the potential to offer new opportunities for improving the education system by using it as a tool to complement learning. The organization also provides ideas on how artificial intelligence can be regulated to give each student more autonomy, increasing possibilities and better adapting to their learning needs.(UNESCO,2024) [1]

Machine learning, as a subset of artificial intelligence, has great potential to be adapted and regulated for the improvement of educational strategies. In this context, data is processed and learned by algorithms, which can be either supervised or unsupervised. The article on the use of machine learning for strategic decision-making in higher educational institutions discusses how ML has demonstrated a strong ability to recognize patterns in data and predict outcomes. Its main strength lies in creating models that can be used to support decision-making processes.(2019)[2]

Based on this point of view, this research has the purpose of using machine learning algorithms to improve the strategies and accentuate the social and educational impact of “Leaders of Tomorrow’s” undergraduate students program from the institution Tecnológico de Monterrey.

Within this framework, the research will create a supervised AI classification model driven by the data of the Institute for the Future of Education (IFE) with the objective of predicting the demographic and geographic areas where students are less likely to graduate from the undergraduate student program “Leaders of Tomorrow’s”. Generating a better understanding of the given data and developing a clear path to improve the graduation rate of the students.

The paper will include a thorough exploration of the dataset, which contains students' sociodemographic, admission, academic, and student life information. This will be done through data understanding, data preparation, and statistical analysis, leading to the creation of a classification model and a final performance evaluation.

I. Data understanding

A. Initial Data

The complete dataset to be used has already been provided, and formatted by the organization Future of Education (IFE) data of the students from the “Leaders of Tomorrow’s” program.

B. Data Description

The dataset is preprocessed and standardized, consisting of 45 variables that include socio demographic information (e.g., gender, age region), academic performance, admission details (e.g., admission type, origin school), and participation in student development programs.

C. Explore Data

The provided Dataset contains 45 columns and 22,718 rows. Having 1,796 registered students on the undergraduate student program ‘Leaders of Tomorrow’s’.

- The dataset shows longitudinal data for each student across multiple terms.
- From the 1,796 students, 51% (924) are male and 49% (872) are female.
- The first 5 most common majors the students take are: IMT being the 11%(198), IIT being the 10%(189), NEG being the 9%(160), IBQ being the 8%(139) and IIS being the 6%(106) . These 5 majors are 44% of the students in total.

- The dataset shows that there are no students from 18 and below who graduated.
- The dataset shows a trend of more male students graduating than female students.
- From the 1,796 students, 53%(959) have been registered as graduates, while 47% (837) of students didn't graduate or have not graduated yet.

For the data exploration, we selected the 959 graduated students and 872 not graduated students to analyze the data that causes the students to graduate the program

- From the other data, the most graduated by default are the students that live in urban areas, while the other percentages are minimal. This can also aid for future usage, where depending on the student area, the smaller or bigger possibility of graduating.

D. Quality Data

There don't seem to be any null values in key variables like student.id, status_academic_desc, isGraduated. Repetition of student ID's across rows is expected due to multiple-term records. We will check for duplicates and inconsistencies during the data preparation phase.

II. Data Preparation

A. Select Data

After exploring the data, the most important aspects are the basic information of a registered student, including their demographic information, their gpa and aptitude scores, and the binary value that determines if the student graduated or not. The variables are directly related to the problem statement and will help model distinguish patterns associated with academic success or dropout.

The selected data includes:

a) Sociodemographic variables: age, gender, zone type region, first-generation status.
b) Academic variables: origin school, test scores, gpa, full-time equivalence, cohort year, major.
c) Participation and engagement indicators: student activity levels, social project scope, international and cultural experience.

B. Clean Data

Before modeling, several actions were performed to be able to use the data for the analysis.:

a) Standardizing values: The "Does not apply values" that exist in half the columns, need to be replaced by a value or being taken out of the set of data to use in the modeling part.
b) Handling nulls: The columns with missing data such as mainCampus.region_code, student.fte, student.term_gpa, and student.term_gpa_program were carefully handled.
   Depending on the importance of each column:
   - For gpa-related variables missing values were imputed using the median or mean gpa based on similar groups.
   - For categorical values like mainCampus.region_code, mode imputation or logical inference from related variables was applied.
c) Duplicated records: Although multiple records per student exist due to longitudinal data, the dataset was filtered to either retain only the latest term per student or to aggregate values depending on modeling strategy.

C. Construct Data

To improve the predictive power of the model, new derived features were created:

- gpa difference: A new variable was constructed to measure the difference between the term gpa and the program gpa.
- zone classification: Simplified classification for zone_type into "Urban" vs "Non-Urban".
- engagement index: Summation of different student_admission_cv.*_level columns to create a unified score indicating overall extracurricular involvement.
- social project impact: Binary indicator derived from student_admission_socialProject.scope to indicate whether the student's project had a national impact.

The constructed variables helped reduce dimensionality while preserving relevant information for classification.

D. Construct Data

For modeling purposes, the dataset was formatted as follows:

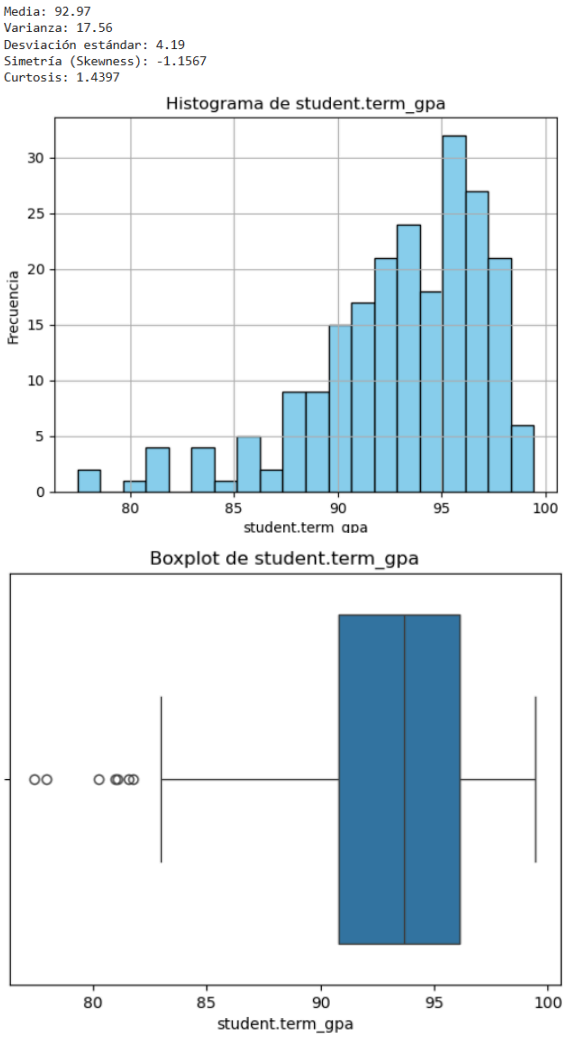
a) Encoding: Categorical variables such as gender_desc, zone_type, and student_admission_test.type_desc were encoded using one-hot encoding or label encoding based on the model requirements.
b) Scaling: Continuous numerical values were normalized or standardized to improve performance in models sensitive to feature scale, such as logistic regression or SVM.
c) Final structure: The final modeling dataset includes one row per student, with all relevant features and a binary label. It was split into training and testing sets, stratified to preserve the graduation rate distribution.

III. Data Statistical Analysis

For this section,we selected two key quantitative variables from the dataset: student.term_gpa (academic GPA per term) and student_admission_test.score (admission test score). Since the dataset included multiple records per student corresponding to different academic terms, the data was grouped by student.id, averaging the GPA and test score to obtain a single representative observation per student. This avoided inflating the sample or skewing statistics due to repeated entries.

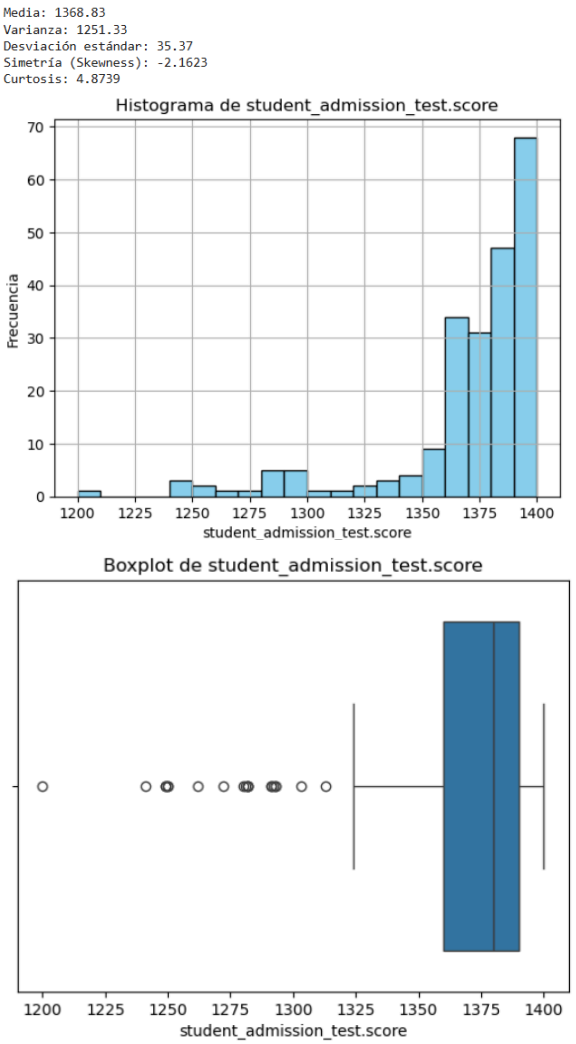
Subsequently, a descriptive statistical analysis was conducted on both variables. For student.term_gpa, the mean was 92.97 with a standard deviation of 4.19. The

distribution showed slight negative skewness (skewness = -1.16) and a kurtosis of 1.44, indicating a concentration of higher values and a flatter shape than a normal distribution.



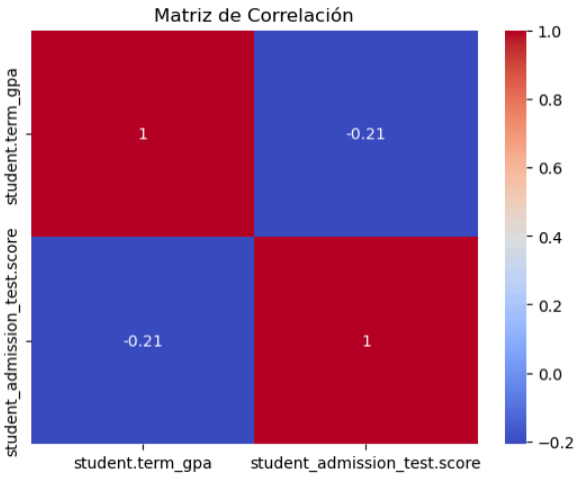
Hypothesis: The distribution of GPA by academic period for students in the Leaders of Tomorrow program is characterized by a high mean (92.97), suggesting generally strong academic performance among students. The low variance (17.56) and standard deviation (4.19) indicate that most GPAs are concentrated near the mean. The moderate negative symmetry (-1.16) suggests a slight accumulation of high scores and a leftward tail, while the low kurtosis (1.44) indicates a more flattened than normal distribution. With these data, it is hypothesized that student.term_gpa does not follow a normal distribution and exhibits a slight concentration of students with outstanding performances.

For student_admission_test.score, the mean was 1368.83 with a standard deviation of 35.37. This variable had more pronounced negative skewness (skewness = -2.16).



Hypothesis: The distribution of admission scores of students in the Leaders of Tomorrow program has a high mean (1368.83), indicating a high level of academic entry. However, the pronounced negative symmetry (-2.16) suggests a strong accumulation of students with scores close to the maximum, and a leftward tail with few cases of low scores. The high kurtosis (4.87) indicates a sharper distribution with heavier tails than a normal one. Despite the low variance (1251.33) and standard deviation (35.37), it is hypothesized that student_admission_test.score does not follow a normal distribution, but a left-skewed distribution, possibly fitted to an inverted lognormal or a truncated gamma.

Correlation Matrix:



After the two key quantitative variables selected for the statistical analysis, we moved to other numerical variables, to test the correlation between them and their possible future usage inside the machine learning model. The variables chosen were student.originSchool.gpa(highschool gpa) and student.fte(the amount of time the students spent in the Tecnologico de Monterrey). These were selected to determine if their previous school performance and the time they spend on the school premises have any correlation to our case.

In this case, for student.originSchool.gpa, the mean value was 96.35 with a standard deviation of 2.75, having a very slight negative skewness (skewness = -0.5756) and a negative kurtosis of -0.6433, this shows that the distribution of data is more flat than normal, having more high values on the right side of the tail.

Hypothesis: The distributions of the highschool gpa scores of the students of Leaders of Tomorrow program has a high mean value of 96.35, this shows that the minimum score to access the program is of 90, following a normal distribution and a standard deviation of 2.75 that depicts that most of the students that access the program are between a range of 93 and 98 gpa in highschool. Since the gpa score has a high standard, the kurtosis represents a negative value of -0.64, this proving that there's not a possibility of having outliers or very extreme values as a result (only values between 90 and 100), while the skewness demonstrates that most of the highschool gpa scores are on the right of the tail, which can be seen on the histogram and in the mean value result and the standard deviation.

In this case, for student.fte, we have a mean value of 0.84 with a standard deviation of 0.11, exhibiting a skewness of 0.21 and a negative kurtosis of -0.8.

Hypothesis: The distribution of academic hours at Tecnológico de Monterrey is slightly lower (0.84) compared to the expected value of 1 for a regular semester. With a variance of 0.01 and a standard deviation of 0.11, the typical course load for Leaders of Tomorrow Program students tends to fall below the standard load, although it remains close to the mean. A kurtosis of -0.8 suggests that the distribution has light tails, indicating fewer extreme values than a normal distribution.

A. Normality Tests and Distribution Fitting

To verify whether the variables followed a normal distribution, the Shapiro-Wilk test was applied. In both cases, the results were conclusive: p-value = 0.0000, leading to the rejection of the null hypothesis of normality. This justified the need to fit the variables to alternative probability distributions.

Five common distributions were tested: exponential, gamma, Weibull, lognormal, and Pareto. The best fit was determined by comparing their negative log-likelihood values. In both variables, the gamma distribution provided the best fit, as it had the lowest log-likelihood value. These findings were also supported by overlaying probability density function (PDF) curves on the histograms.

B. Outlier Detection

Finally, an outlier analysis was conducted using the interquartile range (IQR) method. For student.term_gpa, 7 outliers were identified:

```
=== Outliers en student.term_gpa ===
Se encontraron 7 valores atípicos
34    77.976000
56    77.449750
65    81.011000
95    81.767500
96    80.245667
Name: student.term_gpa, dtype: float64
```

while for student_admission_test.score, 20 outliers were found:

```
=== Outliers en student_admission_test.score ===
Se encontraron 20 valores atípicos
7     1282.0
9     1292.0
22    1282.0
30    1282.0
36    1280.0
Name: student_admission_test.score, dtype: float64
```

These extreme values were listed and could represent exceptional cases.

MODEL *(Heading 5)*

Several machine learning models were trained to determine which one could most accurately classify whether a student in the "Leaders of Tomorrow" program is likely to graduate. The models used were Random Forest Classifier,Extreme gradient boosting (XGBoost), and Logistic Regression. Various experiments were conducted to identify the model with the lowest error rate and the best performance in classifying the students. Among them, the XGBoost model achieved the best results.

The XGBoost classification model was trained with the XGBoost library, with a number of 100 trees or boosting rounds of the model(n_estimator = 100) and the objecting of predicting a binary logistic classification of the encoded dataset, we also modified the train and testing set for this case using a 60/40 ratio. The model trained successfully, and its accuracy on the test set was approximately 97%. However, when we examined the classification report, we

discovered that the model predicted only one class almost perfectly, while it failed to identify 22% of students in the test set that graduated. The precision, recall, and F1-score for the positive class (graduates) were all below 0.81, while the negative class(not graduated) was on a 0.99. This outcome pointed to a great class imbalance in the dataset, where only around 78% of students were labeled as graduated. This model counted with an enormous quantity of boolean columns(caused by the get_dummies encoding).

C. General Conclusion

This statistical analysis provided a clearer and more robust understanding of the academic performance and admission test results of the program's students. By cleaning and aggregating data per student, potential bias from repeated records was avoided. The descriptive metrics, distribution fitting, and outlier detection together offer a strong foundation for predictive modeling as well as insight into the diversity of student profiles within the dataset.

---

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I. TABLE TYPE STYLES

Table Head	Table Column Head		
	Table column subhead	*Subhead*	*Subhead*
copy	More table copy[a]		

[a] Sample of a Table footnote. (*Table footnote*)

Fig. 1. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

ACKNOWLEDGMENT *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1]UNESCO. (2024, mayo 17). El uso de la IA en la educación: decidir el futuro que queremos. UNESCO. https://www.unesco.org/es/articles/el-uso-de-la-ia-en-la-educacion-decidir-el-futuro-que-queremos

[2] Y. Nieto, V. Gacía-Díaz, C. Montenegro, C. C. González and R. González Crespo, "Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions," in IEEE Access, vol. 7, pp. 75007-75017, 2019, doi: 10.1109/ACCESS.2019.2919343.