




A data analytics approach for university competitiveness: the QS world university rankings

Ana Carmen Estrada-Real¹ · Francisco J. Cantu-Ortiz² 

Received: 29 January 2021 / Accepted: 14 June 2022 / Published online: 9 July 2022
© The Author(s) 2022

Abstract

In recent years, higher education has felt pressured to prepare its graduates for the highly competitive international market due to globalization. Thus, many institutions have turned to position themselves well in university rankings as a way to attract the best academic and student talent from all over the world. Our work presents a predictive model for measuring university performance in the QS world university rankings (QS-WUR). We used a ten-year dataset to build models with statistical and machine learning algorithms contained in the library Caret of the RStudio software tool, to forecast global university position in QS-WUR. With these tools, we designed a methodology to predict the university partners' Final Scores based on their historical performance, achieving errors in the range of one or two points out of 100. The modelling may be a useful aid for university officers to develop strategies for improving institutional processes to attract the best students, faculty, and funding, enhance international collaboration and outlook, and foster international university prestige.

Keywords Data science · Predictive modelling · University rankings · Machine learning · Statistics · Educational innovation · Higher education

1 Introduction

At the beginning of the twenty-first century, the concept of world-class universities was introduced. World-class universities have characteristics that differentiate and distinguish them, including recognized graduates, cutting-edge research, and international information and technology transfer. Higher education institutions compete and collaborate to attract the most talented students, the best academics, and international funding [1].

Worldwide, datasets collected by private institutions were used to create display tables, and numerical values were assigned to these institutions' attributes to analyse performance at a global level. These included Nobel prizes,

citations, and publications in *Nature*. The data got ordered and weighted, and World University Rankings were created by ranking institutions. Currently, these rankings are not only used by universities; they are also used by students seeking access to the best education and by academicians and professors wanting to start their careers. These private institutions want to collaborate with research centres and governments to compose education budget assignments [2].

By November 15, 2015, about 40% of the world's population had access to the internet, an indicator of increasingly rapid information creation and trafficking by individuals, institutions, governments, and universities, responding to events worldwide every day. Scientific information has also been part of this growth. The numbers of publications, books, patents, reviews, access to articles and citations, and the large bases of authors and documents that need to be found in Bibliometric databases have increased exponentially. An example of database size is Clarivate Analytics Web of Science (WoS): in 2014, it contained 50,000 schoolbooks, 12,000 research journals, 160,000 conference proceedings, 90 million records, and 1 billion appointments, with 65 million added every year.

✉ Francisco J. Cantu-Ortiz

Ana Carmen Estrada-Real
fcantu@tec.mx; ana.real@exatec.tec.mx

¹ School of Science and Engineering, Tecnológico de Monterrey, Ave. Lago de Guadalupe Km 3.5, 52926 Cd. López Mateos, Mexico State, Mexico

² School of Science and Engineering, Tecnológico de Monterrey, Ave. Eugenio Garza Sada 2501, 64849 Monterrey, NL, Mexico

The phenomenon of university competitiveness caused by the appearance of international rankings is sociological. Higher education institutions (HEIs) are grouped and evaluated worldwide to benefit students, parents, university officers, and other stakeholders [3]. The result of university comparisons is usually given in tables that allow the best universities to be displayed hierarchically, using quantitative measures. These results are published annually by ranking organizations, and the media coverage of these rankings intensifies public interest, praise, and critique. In fact, in 2007, the OECD published a document describing how rankings influence HEI: (1) 50% of respondents used rankings for publicity, (2) 70% wanted to be in the top 10% nationally, (3) 71% wanted to be in the top 25% internationally, (4) over 50% had a formal process of reviewing results, (5) and 68% used rankings as a strategic tool for management and academic improvement [4]. Chavez et al. analysed the importance of having research professors with excellent teaching abilities to attract good students [5]. Cantu et al. elaborates about the importance of watching artificial intelligence technology trends and its impact in higher education [6]. Maria Yudkevich discussed the analogy between university rankings and the Olympic games [7]. Grewal et al. of Leigh University analysed the indicator-manipulation strategies carried out by different universities in the United States to obtain benefits [8].

Not only are universities and students paying attention to rankings, but governments and organizations continuously evaluate HEIs to plan their education resources. In a similar study, the OECD analysed the top 300 universities in the QS-WUR published in 2018, looking at the type of university, the funding they get, the number of students, and their ranking position. They found that 84% of the universities in the top 300 are public, which means that public funding is critical for them to achieve excellence in most cases. Also, the top 300 North American universities have double the average budget, and the European universities quadruple the average. Furthermore, the top 100 universities in the ranking have double the budgets of those in the 101–200 band [9]. Another way universities attract funding is by enrolling international students. The globalization of higher education is a reality, and university rankings play an important role in students' choices. It has been proved that a university with a high position in the rankings has 24% more chances of being chosen by a high performing student [10].

The higher education system is inevitably being marketed. In general, the education system participants are more aware of their roles in a business model, and they obtain benefits from the institution while receiving support and acquiring skills [11]. Universities are employing market strategies, trying to be more effective, studying their competitors, and knowing the indicators of their strengths and weaknesses. Education administrations have experienced

declining government support and rising costs, which has increased competition to attract potential donors, talented students, and qualified academicians [12].

The research question we address in this work is as follows: Is it possible to provide university administrators and educators a framework to diagnose its current teaching and research performance quantitatively, know its perceptions abroad, and predict future achievement using data-analytics modelling for a specific ranking methodology? We address this research question in the following sections in the context of the QS-WUR methodology.

2 Rankings and the QS WUR methodology

The publishing of rankings is defined as the practice of listing universities in an ordered list based on performance indicators. The results of comparing universities performance are usually displayed in tables, which allow the best universities in the world to be hierarchically quantified. These results are published periodically by the ranking organizations and intensive media coverage provides information for students, parents, government, and funding agencies. A highly dynamic and evolving process underlies elaboration of institutional rankings in a periodic basis.

For this work, we chose to use the QS ranking while being aware that there are two other major university world rankings: the Academic Ranking of World Universities (ARWU) and Times Higher Education (THE). The reason for choosing QS is the availability of the data in their website. Should someone intend to analyse either ARWU or THE data, it would be necessary to contact the company and acquire data. Instead, data for QS WUR are publicly available in their Intelligence Unit. QS strives to identify gaps and seek further data and methodological refinement to improve the accuracy of its rankings and other regional and specialized tables. QS is committed to world academic institutions, transparency, continued accuracy, and relevance, which continue to be a powerful tool for stakeholders. Data acquisition teams are responsible for validating the rankings' data, including domestic performance, survey performance, geographical balancing, and managing requests by universities directly to be added to the ranking lists. Many considerations are weighed to correctly present the development of educational institutions worldwide in the best way possible.

QS has identified four main pillars that contribute to a world-class university. These are (1) research, (2) teaching, (3) employability, and (4) internationalization. QS Ranking has been published since 2004, and besides the four pillars, it defines six indicators that have a numerical value from 1 to 100. Each indicator is given a weight. The six indicators are added to yield an overall score with a value between 1 and 100. The institutions are then listed in descending order, the

Table 1 QS World University Rankings dataset (head)

Year	Rank	Institution	AcRep	EmRep	StuFac	CitpFac	IntFac	IntStu	Overall
2019	1	MIT	100	100	100	99.8	100	95.5	100
2019	2	Stanford University	100	100	100	99	99.8	70.5	98.6
2019	3	Harvard University	100	100	99.3	99.8	92.1	75.7	98.5
2019	4	Caltech	98.7	81.2	100	100	96.8	90.3	97.2
2019	5	University Of Oxford	100	100	100	83	99.6	98.8	96.8

Table 2 QS World University Rankings methodology

Indicator	Acronyms	Weight (%)
Academic reputation	AcRep	40
Employer reputation	EmRep	10
Teaching: students per faculty	StuFac	20
Research: citations per faculty	CitpFac	20
International faculty	IntFac	5
International students	IntStu	5

one with the highest score occupying the first position in the ranking. QS-WUR publishes a database with the indicators in Score and Rank for each university and the Total Score and Global Rank. The QS-WUR dataset was downloaded from the QS Intelligence Unit site [13]. In Table 1 the first five universities ranked in 2019 with their corresponding scores for each indicator can be observed.

Table 2 displays the indicators and weights used by the QS-WUR methodology. Academic Reputation, the metric with the highest weight (40%), is based on a survey currently responded to by approximately 100,000 academicians worldwide. Employer Reputation, weighted 10%, comes from another survey responded to by about 45,000 employers. Students per Faculty is a ratio with a 20% weight that aims to represent an institution's teaching quality, favouring those with the lowest number of students per faculty member. Citations per Faculty is a ratio weighted 20% that measures research quality based on the number of citations received by faculty members' publications in the last five years, as reported in Elsevier's Scopus database. International Faculty (5%) is the proportion of professors from other countries. It represents the international strength of an institution measured by its capacity to attract staff worldwide. International Student (5%) is the proportion of students from other countries that reflects the institutional capacity to attract students from around the world and is an indicator of the institutional international brand's strength.

A predictive model would be of great help to universities that want to improve their position. A prediction exercise has already been carried out using THE rankings with data from

2011 to 2016, doing a regression analysis using these years of evolving indicators [13]. Another work was elaborated specifically for universities in Japan with data from THE, predicting its ranking position from university size and internationalization [14]. Finally, a study of English institutions hypothesized that an evaluating system based on citations would be highly correlated to one based on peer evaluation. They proved that citations could make good indicators for evaluation systems and predictions for ranking institutions without using surveys [15]. Scientific data can be analysed appropriately to evaluate and design indicators for diverse academic structure behaviours like the effects of funding on the quality of knowledge produced, and the role of gender in graduate programs [16].

Top universities are distinguished by the quality of the research and technological outcomes of their faculty and students enrolled in academic programs. These results are used to build a prestige that attracts new students, funding from government and agencies, and a worldwide reputation. An example of this behaviour is the study conducted by Lanchao et al. to analyse how top universities collaborate among themselves based on their research outcomes, using the THE university ranking [17].

In general, universities and institutions are affected by the lack of a methodology for analysing the ranking indicators, where useful information could be extracted on indicators the competitors use to improve their positions. Also, no model is known to analyse the dynamics of the universities participating in the rankings, how many universities enter and exit the rankings each year, and how many places a university can rise or fall depending on its performance in the previous year. The novelty of this work consists in giving an in-depth analysis of the information in the ranking indicators that can give universities a way to appear in elite academic institutions, know their competitors better, and advise them about strengths and weaknesses to watch. The purpose of this study is to provide stakeholders a methodology and tools that may allow them to design and deploy strategies to improve overall institutional performance and international perception.

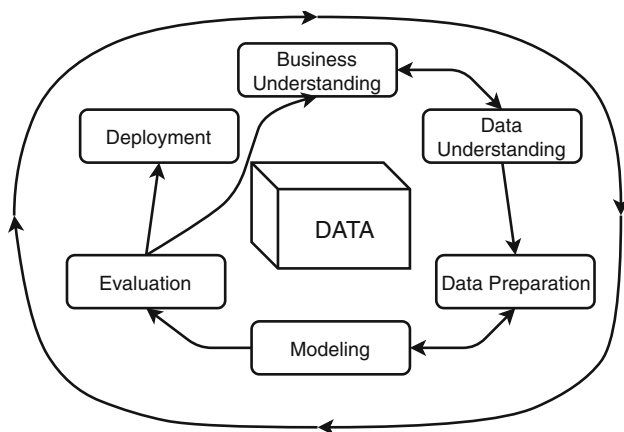


Fig. 1 CRISP Methodology Diagram

3 Methodology and data

To answer our research question, we followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, a data science international standard used by decision-makers in business settings [18]. CRISP-DM is composed of six phases, as shown in Fig. 1: (1) Business understanding, (2) Data understanding, (3) Data Preparation, (4) Modelling, (5) Evaluation, and (6) Deployment, which are explained, applied, and discussed using the QS-WUR dataset in the following sections.

This section describes the first three steps from the CRISP-DM methodology, starting with business understanding and working with the data in the data understanding and data preparation phases, using the QS-WUR dataset.

3.1 Business Understanding

We define the problem's objective in this phase and set the business goals by defining quantitative targets. In our case, we expect to understand and predict university performance in world university rankings by applying data science algorithms and using the QS-WUR methodology to advise university administration about strategies and actions to improve internal academic processes that eventually would be reflected in institutional performance as measured by rankings. After building predictive models, checking the test data, cleaning it, and doing exploratory data analysis, we obtained a deeper understanding of QS-WUR methodology and universities' internal processes to implement academic improvements. We also acquired a better understanding of how universities gather the data required by the ranking methodology, submit their institutional data to the QS website, and elaborate a media communication strategy once the rankings are published every year.

Table 3 Clustering universities by rank ranges

Group	Rank range
A10	[1, 10]
A50	[1, 50]
A100	[1, 100]
A200	[1, 200]
A101	[101, 200]
A201	[201, 317]
All	All

3.2 Data Understanding and Data Preparation

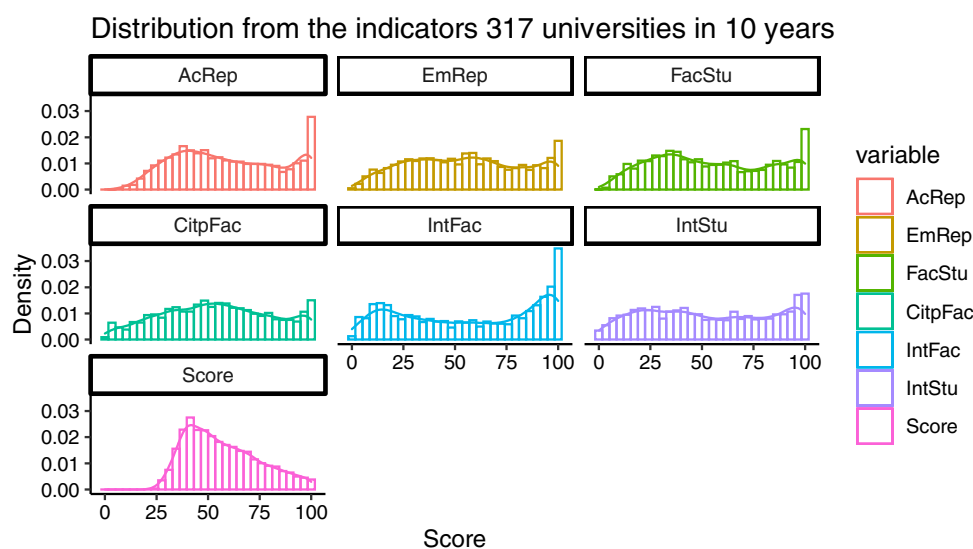
These two phases are carried out in parallel to collect the data to build the model. We formed a data mining table with rows and columns representing records and attributes of various types. We did data cleaning, filling in missing data, formatting operations on rows and columns, and data integration. Then, we interweaved data preparation with exploratory data analysis (EDA) using descriptive statistics to understand the nature of the dataset and its relationship to the business goal. We prepared the data in a longitudinal format so that the ten-year data appeared in one data mining table with the year 2020 at the top and 2011 at the bottom. We selected consistent universities for the analysis, i.e., a total of 317 universities that had appeared in the ranking for ten consecutive years. We did name validation of all the universities. This cleaning process was carried out using the R software. Some universities were still missing values after doing the name cleaning process; these were imputed using K-Nearest Neighbour (KNN) Imputation, particularly 3NN. We carried out an exploratory analysis, including some visualizations, such as indicator histograms and probability density. Then, we performed a clustering exercise, dividing the dataset into groups according to rank ranges, as shown in Table 3.

We chose the groups of particular interest to stakeholders. Top-Ten universities is the upper group, then the Top-50, the Top-100, and the Top-200. With these groups, we carried out statistical measures to understand the differences between them and determine the most critical indicators for a university's focus when designing a plan to grow and improve its academic performance.

Table 4 shows the first exercise with these groups, extracting the maximum and minimum scores achieved by universities in 2020. Suppose a university in a position below the 100 ranking intends to enter the Top 100. In that case, it can evaluate the minimum score obtained by universities in the group in the previous year of the ranking. For Academic Reputation, a minimum score of 40.10 is required, and the minimum overall score is 59.90.

Table 4 Maximum and minimum scores achieved by universities in each group by the indicator in the 2020 ranking

	Group	AcRep	EmRep	FacStu	CitpFac	IntFac	IntStu	Score
Max	A10	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Min	A10	97.80	81.20	85.00	72.10	70.20	62.20	92.00
Max	A50	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Min	A50	71.20	51.80	19.80	24.00	11.10	10.10	74.20
Max	A100	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Min	A100	40.10	22.70	11.80	2.40	6.90	3.60	59.90
Max	A200	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Min	A200	19.90	7.40	5.40	2.40	3.30	1.50	44.00
Max	A101	90.90	97.20	100.00	97.10	100.00	99.10	59.50
Min	A101	19.90	7.40	5.40	3.80	3.30	1.50	44.00
Max	A201	72.90	71.60	100.00	95.40	100.00	100.00	43.50
Min	A201	7.80	5.00	3.40	1.90	1.90	1.00	24.20
Max	data	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Min	data	7.80	5.00	3.40	1.90	1.90	1.00	24.20

**Fig. 2** Histogram and probabilistic score distributions of the six indicators and the overall score

We plotted histograms and the density line above them to see the indicators' distribution (Fig. 2). The X-axis represents the score, and each bin is ten units wide, so it is easy to see the distributions of scores that are more frequent for universities. For example, most universities have Academic Reputation scores in the range [40, 50]. For Faculty Student, the more frequent range of scores achieved was [30, 40]. For the overall score, there are no universities below 23, as most of them are in the range [40, 50].

We also carried out a normality test to know if our data followed a Gaussian distribution. The test was Shapiro–Wilk, and it was performed on the six indicators and the overall score. We found that none of them met the normality assumptions, as indicated by the p -values for $p < 2.2e-16$ for all six

scores. The non-normality test justified using Spearman's correlation instead of the Pearson correlation, which requires that the data be normally distributed. Also, Spearman's correlation assumes that the data must be ordinal, monotonic, and independent. Table 5 shows the different levels of correlation between the indicators of each group and the total score. For the A10 group where the first ten universities are located, there is a very similar correlation, approximately 0.5, among the indicators of Academic Reputation, Employer Reputation, and Citations per Faculty. Citations and Employer Reputation have the highest correlation. Typically, the most recognized universities also display a high quality in research and their graduates' performance in working life. On the other hand, the A50 group has the strongest

Table 5 Spearman correlation coefficients for the six indicators related to the final score

Group	AcRep	EmRep	FacStu	CitpFac	IntFac	IntStu	Score
A10	0.56	0.53	0.32	0.46	0.07	− 0.09	1
A50	0.59	0.48	0.62	0.57	0.20	0.22	1
A100	0.82	0.65	0.51	0.40	0.24	0.31	1
A200	0.84	0.68	0.41	0.48	0.25	0.37	1
A101	0.45	0.25	0.20	0.27	0.02	0.12	1
A201	0.49	0.42	0.10	0.11	0.01	0.08	1
All	0.88	0.69	0.45	0.56	0.24	0.34	1

correlation between Faculty Student Ratio and International Faculty. These universities are highly focused on teaching and have an excellent international attraction.

In the following sections, we present the modeling, evaluation, and results as part of the CRISP-DM methodology.

4 Modeling and Evaluation

We present the models obtained by applying statistical and machine learning algorithms to various sets of universities. We split the dataset into training and test data applying a cross-validation approach. We carried out a Feature Selection exercise to evaluate the six indicators' validity using the Recursive Feature Elimination (RFE) algorithm of the R caret package, as exhibited in Fig. 3. RFE used different machine learning algorithms to help select features. RFE differs from filter-based feature selection methods that score each feature and select those features with the largest or smallest score up to a certain number. Thus, RFE is a wrapper-style feature selection algorithm that also uses filter-based feature selection up to a predefined maximum number.

This algorithm delivers a list with the most relevant features from highest to lowest. The algorithm recommended keeping the six indicators because it achieved the lowest Root Mean Square Error (RMSE).

Then we used the resulting training data to build a predictive model with supervised machine learning methods with categorical response variables that included Multiple Regression with Panel Data, Logistic Regression, Decision Trees and Random Forest, and Support Vector Machines. These algorithms are contained in the library Caret of the RStudio software tool and briefly explained the modelling section. The results obtained by the model were evaluated using test data and by interpreting several statistical measures like R², p-value, Accuracy, Sensitivity, Specificity, Confusion Matrix, Receiver Operational Characteristic (ROC), and Area Under the Curve (AUC). The rows in the Confusion Matrix represented model prediction, whereas the columns showed the real outcome. Positive outcomes corresponded to

Top 100 classification, while negative outcomes meant Top 200 assignment.

4.1 Multiple Regression and Panel Data

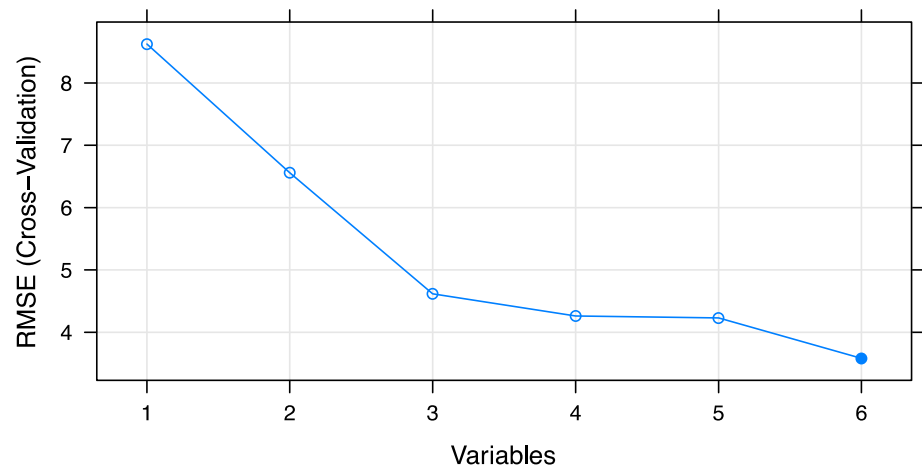
Multiple Regression fits a mathematical function to data such that the error between the predicted value and the real one is minimized. The function can be a polynomial of degree n higher than zero with unknown coefficients whose value is calculated by an optimization procedure such as Ordinary Least Squares (OLS). To build the model, we applied first Multiple Regression and then Panel Data. The training set consisted of 317 universities from year 2011 to 2019 while the test set was year 2020. The model found with Multiple Regression method over a single dataset with 10-year data with no temporal considerations is shown in Eq. (1).

$$\begin{aligned} \text{Score} = & -0.019 + 0.406 \times \text{AcRep} + 0.096 \times \text{EmRep} \\ & + 0.202 \times \text{FacStu} + 0.200 \times \text{CitpFac} \\ & + 0.049 \times \text{IntFac} + 0.054 \times \text{IntStu} \end{aligned} \quad (1)$$

Score the predicted value and the numbers are the coefficients found by the optimization procedure and is the mark achieved by an institution. Number -0.019 is the intercept, and the variables are each of the six indicators with values in the range of $[1, 100]$ multiplied by the coefficient to achieve the final score in the same range $[1, 100]$. This regression equation is used to predict the final score of any university that intends to be ranked by QS-WUR. If the university provides data for each of the six indicators with no temporal history, it is difficult to rank what has not been evaluated by QS; however, it is possible to approximate a score based on categorical characteristics like geography, size, and other attributes.

Interpretation of this model is as follows: notice that the intercept is negative, and the coefficient is very close to the real weight assumed by QS. To consider the effect of time, we used Panel Data with Fixed Effect instead of Random Effects after applying the Hausman test. This way, we obtained a more precise model with specific equation parameters for

Fig. 3 RMSE achieved by the Feature Selection algorithm with a different number of variables



each university and slightly better prediction, obtained by comparing the R-squared for multiple regression of 0.99906 with Panel Data, which was 0.99910, and the p -values of statistical significance.

Panel Data allows the designer to take advantage of trends and variations found over time. It takes entities called sections whose attributes change with time. It combines sectional analysis of entity features with longitudinal behavior of those features. In our case the entities are the universities, the attributes are the ranking parameters, and time goes from 2911 to 2020. There are several Panel Data methods that includes Fixed or Random effects which depends on the assumptions made about the type of variations on sections and time. We used the method *Fixed Effects* found in Caret library. After applying Panel Data, we obtained an equation that can make more accurate predictions compared to Multiple Regression. The model obtained with Panel Data is shown in Eq. (2). The equation has a coefficient for each university and for each year. The general form of the Panel Data equation is as follows:

$$\begin{aligned} \text{Score} = & 3.0013 + 0.3644 \times \text{AcRep} + 0.0969 \times \text{EmRep} \\ & + 0.1970 \times \text{FacStu} + 0.2039 \times \text{CitpFac} \\ & + 0.0487 \times \text{IntFac} + 0.0442 \times \text{IntStu} \end{aligned} \quad (2)$$

The coefficients for the variables have changed, moving slightly away from the values obtained by Multiple Regression. Nevertheless, we can show some of the resulting indicators by university. For example, for MIT, we got a coefficient of 2.1909. This means that this institution has a very positive impact on the final score; it has been in first place for over five years in the ranking. Stanford University has a similar coefficient, 2.0887, and Harvard with 2.1372; these universities are in the top three. As universities get farther from the first places in the ranking, their impact on the Final Score tends to be smaller. For example, the University

Table 6 Metrics with the performance of Multiple Regression and Panel Data on the test set

	RMSE	R-squared	MAE
Multiple Regression	0.9425	0.9990	0.8347
Panel data	0.5489	0.9991	0.3842

of St. Andrews is number 100 in the ranking with a coefficient of 0.7074; Keio University in place 200 has a coefficient of 0.4416; Ecole Des Ponts Paristech, in position 250, has a coefficient of -0.5546 , which has turned negative.

These coefficients are individual, so in trying to predict a specific university's score, each indicator's score is introduced to the equation. This coefficient is then added for each particular case, giving us a more accurate approximation of the final value.

Since we performed a multiple regression and a panel data model, we evaluated the performance and compared them to justify the one we choose for later experiments.

The metrics used are Root Mean Square Error (RMSE), in which a value of 0 represents a perfect fit; R-squared, which represents the proportional variance in the regression (when R-squared is closer to 1, the model has a better fit to data); and Mean Absolute Error (MAE).

In Table 6, we show the R-squared obtained for the test set. The closer to 1, the better the model behaves, so Panel Data is slightly better than multiple regression. For RMSE and MAE Panel Data obtained better performance compared to Multiple Regression. We know that the difference is minimal, but in a ranking as competitive as QS, decimals can define several places. That is why we seek the best performance. So, we decided to use the Panel Data model for the rest of the experiments.

Finally, for QS WUR, we used a non-linear regression algorithm because we found that linear regression was very optimistic in terms of prediction. Loess was used instead.

We trained the model with nine years of data and left the last one for testing/validation. We used a $span = 0.75$ and a $degree = 2$. Since this regression is applied individually, we did not get an equation. This regression was used for the three universities in explained in the deployment step: Tecnológico de Monterrey, the University of Texas at Austin, and Carnegie Mellon University.

4.2 Machine Learning Methods

We now present the use of machine learning methods employed to classify the universities included in the study. Classification algorithms predict the group to which each university belongs. We applied Logistic Regression, Support Vector Machines (SVM) and Random Forest in this experiment. SVM was applied with two variants, linear and radial method.

Logistic Regression is a binary classifier that uses a sigmoidal exponential function to split the search space in two regions or classes. SVM is a supervised machine learning method that builds a hyperplane in an n -dimensional space to classify elements of the dataset. It is also a binary classifier that maps training data to points in space to maximise the width of the gap between two categories. To build the hyperplane, SVC uses functions called kernels. Linear SVM uses a function to split linearly separable data into two classes by using a single linear hyperplane. Decision Tree is a classifier that uses a tree-structure where nodes represent the features of the dataset, branches represent the decision rules, and each leaf node represents a possible outcome. To select the best nodes in the tree attribute selection methods are used. The most popular ones are information gain measured by entropy calculation, and Gini index. We used information gain with entropy since it produced the best results. Random Forest is a classification method that uses a set of decision trees and select the best combination using random selection to avoid the problem of overfitting. Overfitting occurs when a model learns to well a dataset in such a way that it is unable to generalize to classify unseen data.

4.2.1 Top 200: Two Groups of 100

For the first classification exercise, we used the top 200 universities in the dataset and divided them into a training set of 70% and a test set of 30% of the universities. We divided the top 200 universities into two groups of 100. The algorithms were trained with the scores.

The categorization is fundamental because many universities aim to enter the Top 100 universities worldwide. This training makes it possible for universities to propose a combination of scores they plan to achieve in subsequent years by improving their resource management to see if their efforts are enough to enter the Top 100 group.

Table 7 Logistic regression confusion matrix (Accuracy 0.9948)

	Top 100	Top 200
Top 100	296	2
Top 200	1	288

Table 8 Accuracy and AUC for the four models in the test set

Model	Accuracy	AUC
Logistic regression	0.9948	0.995
SVM Linear	0.9812	0.959
SVM Radial	0.9608	0.903
Random forest	0.9948	0.995

Table 7 presents the confusion matrix of the first categorization exercise with the two groups using logistic regression. We can see that most universities are categorized correctly.

In general, all the algorithms performed well; however, Logistic Regression and Random Forest stand out, having the same accuracy of 0.9948, as displayed in Table 8.

Random Forest had the same accuracy and performance as Logistic Regression. All four algorithms did an excellent job classifying the two groups. This is positive for us because these groups are easy to identify by a machine learning approach. If a university in the top 200 wants to know if it will get into the top 100, that question can be answered.

With the ROC curves, we could visualize Specificity vs. Sensitivity. These measures are extracted from the confusion matrix of each model. Figure 4 shows that Logistic Regression and Random Forest have almost perfect classification performance when the universities are categorized into the two groups of Top 100 and Top 200.

4.2.2 Top 200: Ten Groups of Twenty

The second classification exercise was carried out with the same Top 200 universities. We divided them into ten groups of 20. We wanted to test if the algorithms could predict more precise positions. We had to leave out the logistic regression algorithm because it works for binary classes only, so we decided to try Decision Trees instead.

The performance decreased a lot compared to the two classes' exercise. Random Forest attained the best accuracy with 0.8979 in the test set, as shown in Table 9.

In Fig. 5, we have the ROC curves obtained evaluating the four algorithms used to classify the ten groups. We can see that Random Forest and SVM radial have the best performances.

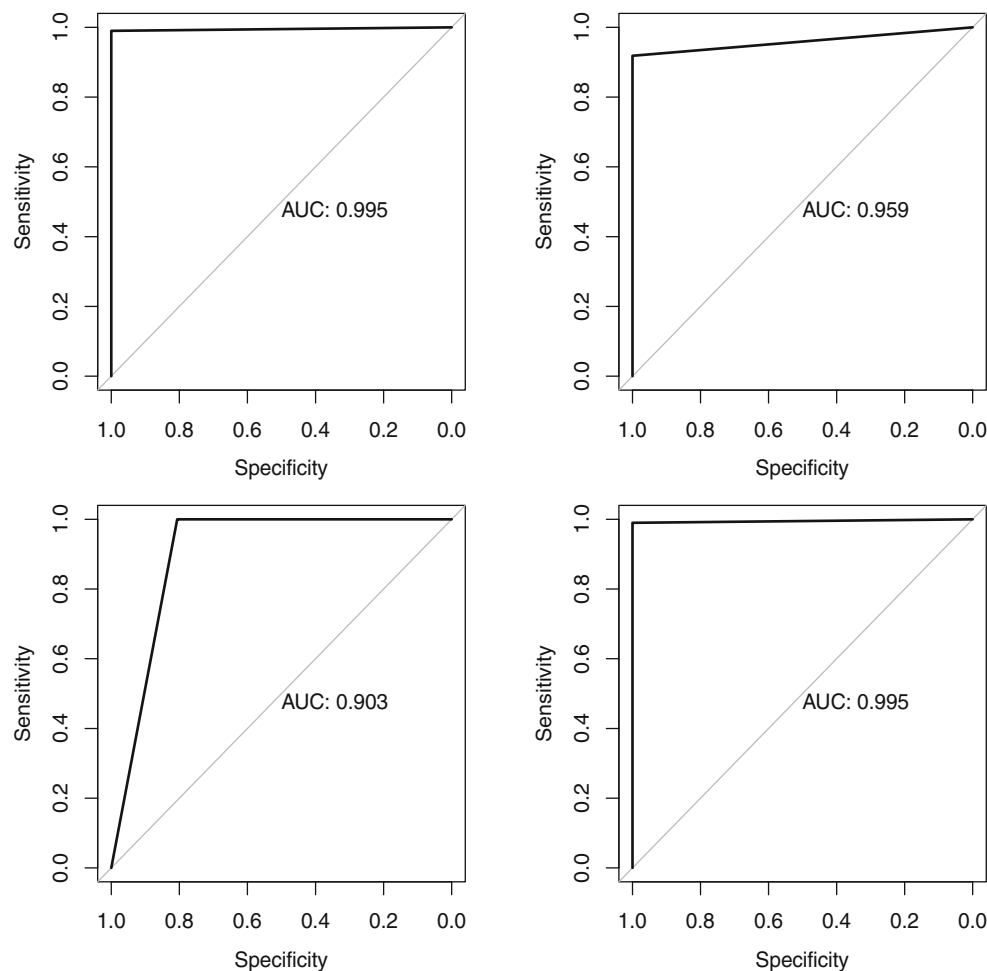


Fig. 4 ROC curves with AUC values for: **a** Logistic regression, **b** SVM linear, **c** SVM radial, and **d** Random forest

Table 9 Accuracy and AUC for the four models in the test set

Model	Accuracy	AUC
Decision trees	0.3622	0.833
SVM linear	0.3928	0.881
SVM radial	0.6734	0.964
Random forest	0.8979	0.994

4.2.3 Bayesian Modeling

In addition to the prediction exercises, we also wanted to find conditional structures in the data. These structures were modelled using Bayesian networks; the networks were learned directly from the data. We took the top 100 universities and trained a static network for each year, having a total of ten static networks trained.

Considering the seven indicators as random variables, we also performed Bayesian modelling that consisted of learning

a Bayesian network from the data and learning the conditional probability tables for each random variable. Then we were able to make a probabilistic inference about the ranking outcome for a particular institution.

In Fig. 6, we observe the ten structures learned from the data. The structure changes from year to year. We believe these reflect the evolution of the methodology with which QS calculates the Final Score. We kept the last one from the year 2020 to make inferences.

As part of the Bayesian-probability work, we decided to calculate the conditional probability tables. For the calculation of these tables, the variables needed to have categorical values. In this case, the indicators of the universities in the Top 100 of the year 2020 were taken, and two groups were created; namely, the indicators ranked 1–50 and 51–100. For this case, we took 1–50 as the true value and 51–100 as false. Next, we show the tables with the obtained probabilities:

$$P(\text{Score}, \text{AcRep}, \text{EmRep}, \text{CitpFac}, \text{FacStu}, \text{IntStu}, \text{IntFac}) \\ = \sigma_{\{\text{score}, \text{acrep}, \text{emrep}, \text{citpfac}, \text{facstu}, \text{intstu}, \text{intfac}\}}$$

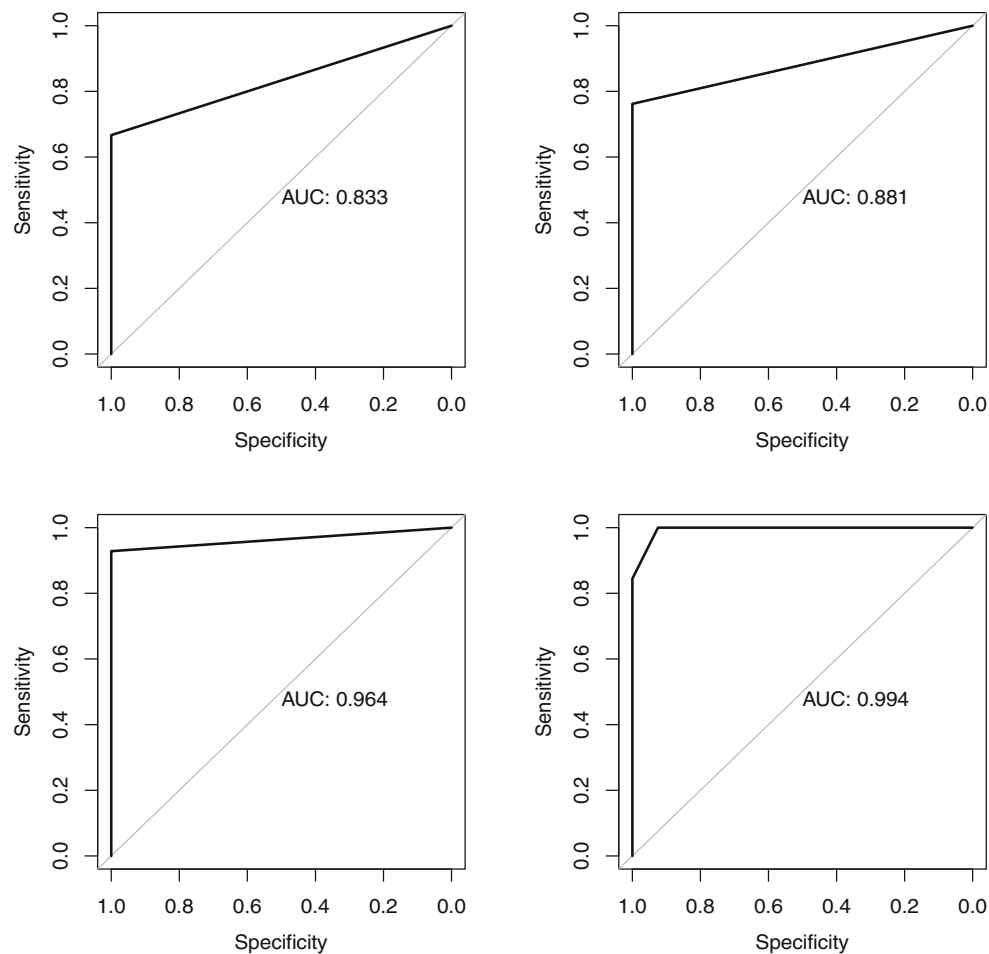
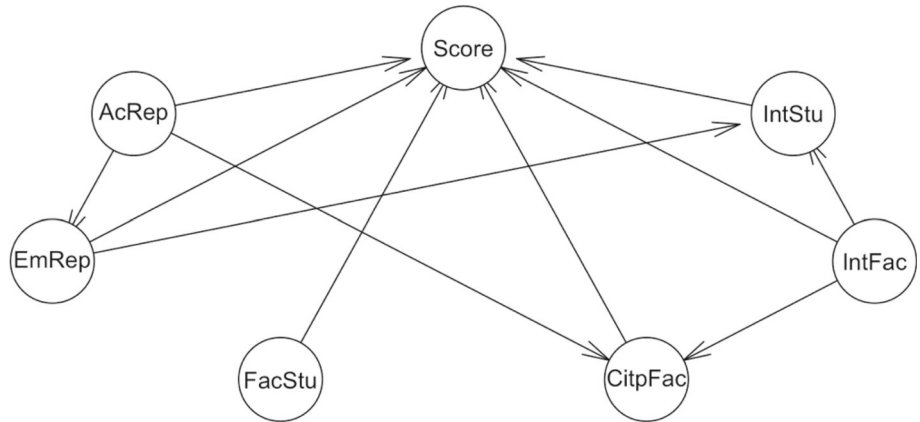


Fig. 5 ROC curves with AUC values for: **a** Decision Trees, **b** SVM linear, **c** SVM radial, and **d** Random forest

Fig. 6 Bayesian network learned from the 2020 data



$$\begin{aligned}
 &P(\text{Score}|\text{AcRep}, \text{EmRep}, \text{FacStu}, \text{CitpFac}, \text{IntStu}, \text{IntFac}) \\
 &\quad * P(\text{AcRep}) * P(\text{EmRep}|\text{AcRep}) * P(\text{FacStu}) \\
 &\quad * P(\text{CitpFac}|\text{AcRep}, \text{IntFac}) \\
 &\quad * P(\text{IntStu}|\text{EmRep}, \text{IntFac}) * P(\text{IntFac})
 \end{aligned} \quad (3)$$

The probability function is presented in Eq. (3).

For independent variables, such as Academic Reputation (AcRep), Faculty Student (FacStu), and International Faculty (IntFac), their probabilities were directly assigned. For the rest of the variables, we had to consult the conditional

Table 10 Conditional probability table for Academic Reputation

True	False
0.2808	0.7194

Table 11 Conditional probability table for Faculty-Student Ratio

True	False
0.3928	0.6072

Table 12 Conditional probability table for International Faculty

True	False
0.2959	0.7041

Table 13 Conditional probability table for Employer Reputation (dependent on Academic Reputation)

	EmRep	True	False
AcRep	True	0.6363	0.3637
	False	0.7588	0.2412

probability table (CitpFac) to compute the overall probability that a university belonged to the Top 50 or the 51–100 group.

With these Tables 10, 11, 12, 13, 14, 15, stakeholders should be able to calculate the probability of an institution belonging to the Top 50 or 50–100 groups, depending on the score of their indicators.

We wonder how good the results obtained are and how they could be improved. The standard procedure used to measure algorithm performance is by way of metrics. We have show that Random Forest and radial SVM models produced the best results. Random forest dealt with overfitting in decision trees whereas radial SVM may reflect nonlinear patterns present in the dataset. One way to improve

Table 14 Conditional probability table for Citations per Faculty, dependent on Academic Reputation and International Faculty

AcRep	IntFac	CitpFac	
		T	F
T	T	0.6153	0.3847
T	F	0.3572	0.6428
F	T	0.5333	0.4667
F	F	0.2812	0.7188

Table 15 Conditional probability table for International Students, dependent on Employer Reputation and International Faculty

EmRep	IntFac	IntStu	
		T	F
T	T	0.9444	0.0556
T	F	0.4705	0.5295
F	T	0.7250	0.2750
F	F	0.1954	0.8046

performance of machine learning methods is by applying parameter optimization algorithms with bagging and ensemble approaches applying bootstrap aggregation as well as gradient and boosting methods in classifiers like Bagging and XGBoost (Extreme Gradient Boosting).

5 Results

In this section, we explain the deployment of the CRISP-DM methodology we have been following. We present case studies and compare our results with the QS-WUR 2020.

This study was conducted before the 2020 ranking came out. So, we consider that this work is not affected by the final result of the ranking. We use it as a point of comparison to evaluate the models.

5.1 Tecnológico de Monterrey (Tec)

Equation (2) was used. The coefficient for Tecnológico de Monterrey is 0.2123. In the last year, Tec was in position 173; currently, it is in position 158.

In the linear regression, the International Faculty indicator has been systematically improving (Fig. 7 (left)). Thus, the prediction exceeds the value of 100, which is the highest score, so a maximum limit was set to 100.

In the Loess regression (Fig. 7 (right)), the Faculty-Student ratio and Citations per Faculty indicators tend to grow, while the others tend to decrease.

This contrast between the two models is noteworthy because observing the linear regression shows an optimistic scenario in all the indicators with a positive slope. Hence, the Loess regression provides a more conservative scenario that helps to contrast both results (Figs. 8, 9, 10, 11, 12, 13).

Table 16 shows that the two regressions and predictions for 2020 have a total score of 45.9 with Loess regression and 52 with linear regression.

In the 2020 ranking, there was a 48.5 overall score. If an average is calculated between the two predictions that were made, an overall score of 48.9 would be obtained. This

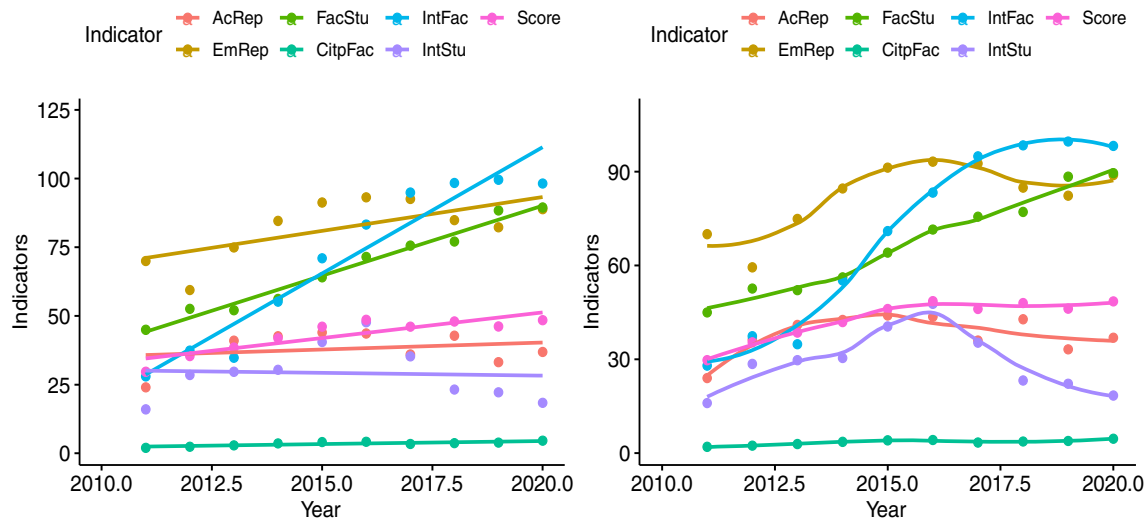


Fig. 7 Scatter plots of Tecnológico de Monterrey. Left: linear regression. Right: Loess regression

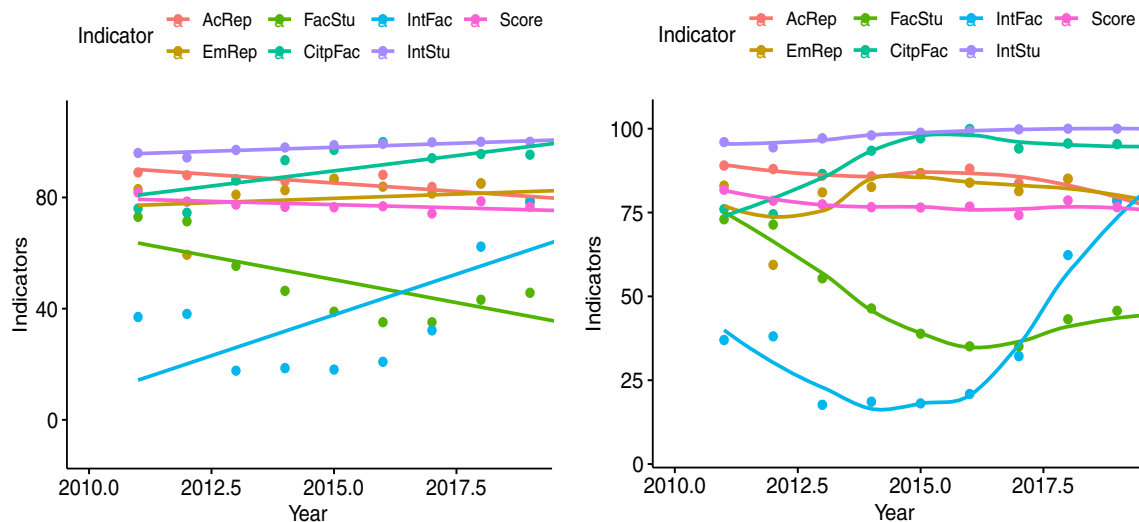


Fig. 8 Scatter plots of Carnegie Mellon University. Left: linear regression; Right: Loess regression

calculation was made as a means of comparison to know how far apart the values that were obtained were vs. the real one.

In general, linear regression gives a more optimistic prediction than Loess regression, which has a local weight for all indicators, giving a more pessimistic prediction.

5.2 Carnegie Mellon University (CMU)

The second case was Carnegie Mellon University, which currently has a rank of 48. Similar to the above, we first performed an analysis of the 2011–2019 indicators with linear and non-linear regressions.

In the linear regression, we see that international students have had consistently high scores. The Citations per faculty has a positive slope even though it has decreased in recent

years. Employer Reputation also has a positive slope, and International Faculty has improved significantly in recent years. We observe two indicators with a negative slope: Academic Reputation and Faculty-Student ratio, meaning that they have accepted many more students than professors. This is reflected in the Final Score, which can be seen with a negative slope as well.

The advantage of Loess is that it gives a local weight to each indicator; here, you see the indicators' changes better. We observe that in 2016, several indicators changed. Some began to improve, and others declined in performance. This may be due to a change in the methodology that was published in 2017. Faculty-Student ratio and International

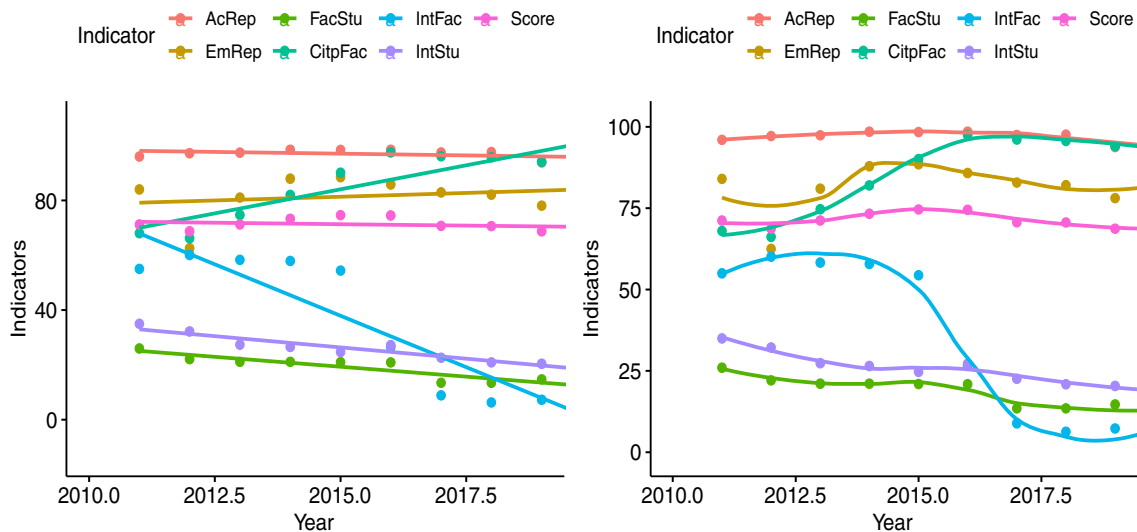


Fig. 9 Scatter plots of the University of Texas at Austin. Left: Linear regression; Right: Loess regression

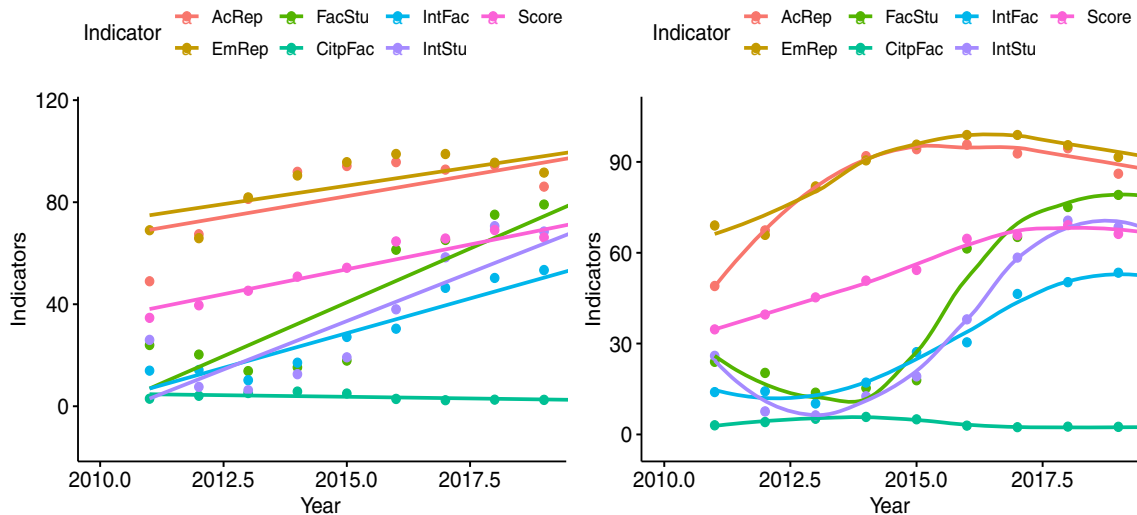


Fig. 10 Scatter plots of the University of Buenos Aires. Left: Linear regression; Right: Loess regression

Faculty similarly improved. Employer Reputation and Academic Reputation seem to be down in recent years. The Final Score dropped in the last year as well.

In Table 17, we see the indicator predictions. The linear regression favors Citations per Faculty and International Students, but it punishes International Faculty, obtaining a Final Score of 75. With the non-linear regression, International Faculty has benefited from its improvement in recent years; it performs better in this prediction, unlike Tecnológico de Monterrey, where the linear regression is optimistic.

5.3 University of Texas at Austin (UT Austin)

This university is currently ranked 65. We were interested in analyzing a range of institutions to validate our proposal. This university is in the top 100 but below the top 50 to which Carnegie Mellon belongs.

We first analyzed its performance in previous years. In the linear regression, only two indicators have a positive slope: Citations per Faculty and Employer Reputation. The rest have negative slopes. International Faculty drops a lot, as well as Faculty-Student ratio and International Student.

As in Carnegie Mellon, the 2016 methodological changes appear to have affected this university's performance in the ranking. With the non-linear calculation, we see the evolution

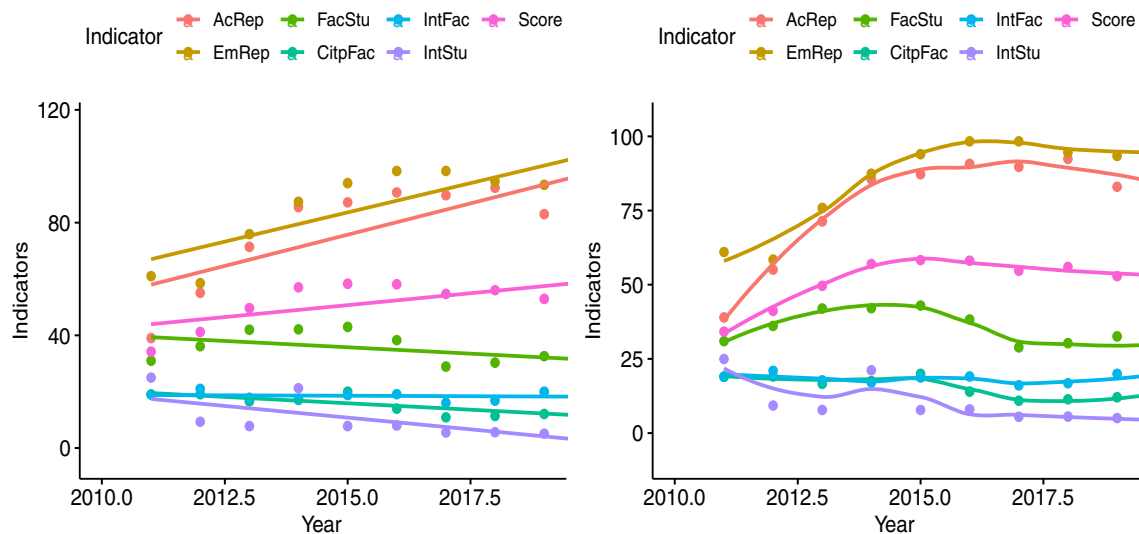


Fig. 11 Scatter plots of Pontificia Universidad Católica de Chile. Left: Linear regression; Right: Loess regression

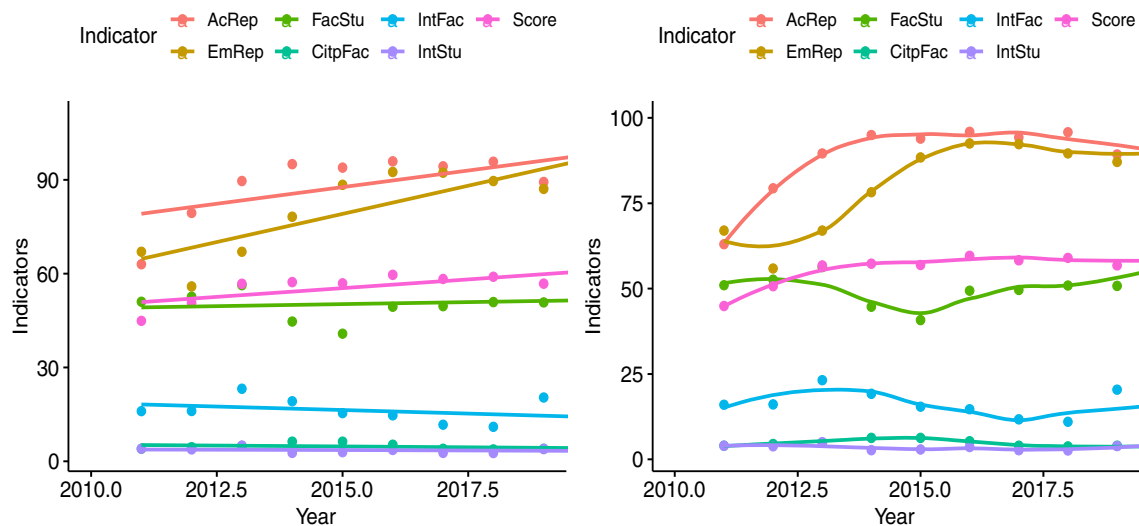


Fig. 12 Scatter plots of Universidad Nacional Autónoma de México. Left: Linear regression; Right: Loess regression

of the indicators. Faculty-Student ratio and International Faculty improved in the last year, while all the other indicators declined, including the Final Score.

Table 18 shows that the linear regression prediction benefited Citations per Faculty and predicted a very low (actually negative) rating for International Faculty. However, we assigned it a lower limit of zero. In the non-linear regression, we found a lower prediction because of Academic Reputation, the indicator with the highest weight, which decreased in recent years. Finally, comparing the average with the actual 2020 result, we had a difference of -1.67 in the Final Score.

5.4 University of Buenos Aires (UBA)

Currently, this university ranks 74. We were interested in analysing Latin American universities. Below is the evolution of UBA indicators over the nine years before 2020.

In the linear regression, we see that its strongest indicators are Academic Reputation and Employer Reputation. Faculty-Student ratio improved in recent years, and International Student and International Faculty also have a positive slope. The only indicator with a negative slope is Citations per Faculty.

With non-linear regression, we see that Academic Reputation and Employer Reputation have been decreasing despite having the best scores. International Students and Citations

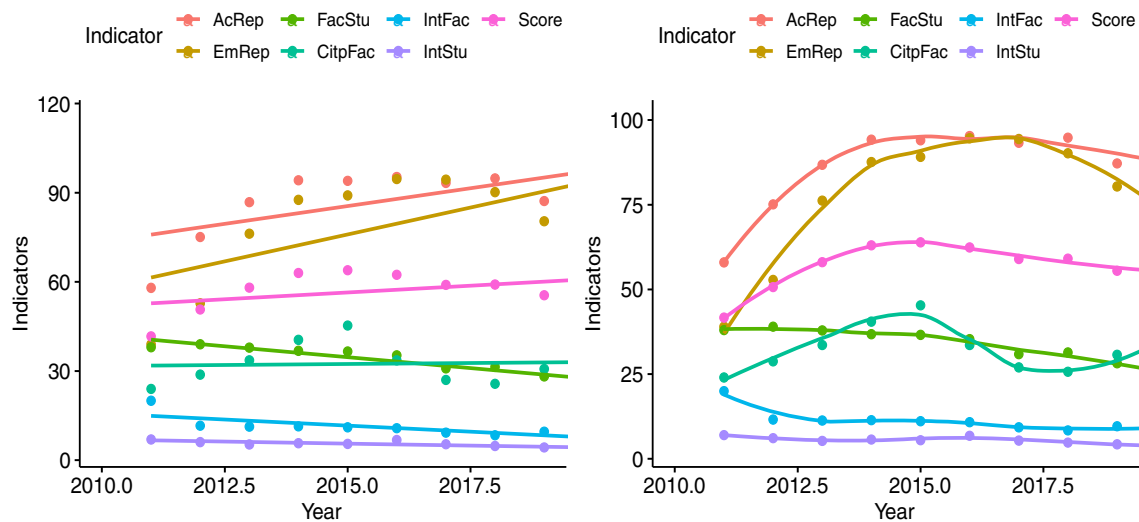


Fig. 13 Scatter plots of Universidade de São Paulo. Left: Linear regression; Right: Loess regression

Table 16 Prediction of indicators and the overall score for 2020 for Tecnológico de Monterrey

	AcRep	EmRep	FacStu	CitpFac	IntStu	IntFac	Score
Loess	28.94	75.08	98.11	4.52	16.96	95.34	45.90
Linear	42.11	95.61	90.52	4.45	33.51	100.00	52.00
Mean	35.52	85.34	94.31	4.48	25.23	97.67	48.95
Real	36.90	88.90	89.50	4.60	18.40	98.20	48.50
Difference	− 1.38	− 3.56	4.81	− 0.12	6.23	− 0.53	0.45

Table 17 Prediction of indicators and Overall Score for the year 2020 for Carnegie Mellon University

	AcRep	EmRep	FacStu	CitpFac	IntStu	IntFac	Score
Linear	81.21	85.51	28.88	100.00	100.00	58.34	75.00
Loess	73.02	75.92	53.76	97.16	99.79	100.00	77.36
Mean	77.115	80.715	41.32	98.58	99.895	79.17	76.18
Real	75.2	77.5	43.5	94.3	99.9	83.6	74.8
Difference	− 1.915	− 3.22	2.18	− 4.28	0.01	− 4.43	− 1.38

per Faculty have also decreased, but more subtly. International faculty and Faculty-Student ratio improved in 2019, which can help in predicting 2020. This is shown in Table 19.

In this case, we see a large difference between the linear prediction and that of the Loess model, more than 12 units. However, when calculating the average, we find that the real 2020 value difference is -1.54 in the Final Score. This is in the expected range of difference between our predictions and the real values of 2020.

5.5 Pontificia Universidad Católica de Chile (PUC Chile)

This university is currently ranked 127. It is part of our Latin American universities analysis. In the last ten years, it has risen from 331, improving more than 200 places.

In the Linear regression, there are excellent scores in the Academic Reputation and Employer Reputation indicators. However, all the other indicators have negative slopes.

In the Loess regression, we see behavior similar to that of the University of Buenos Aires. The Academic Reputation and Employer Reputation indicators have declined in recent years, but the other indicators have a positive slope showing an improvement from 2018 to 2019, except for International Students.

Table 18 Prediction of indicators and Overall Score for the year 2020 for the University Of Texas at Austin

	AcRep	EmRep	FacStu	CitpFac	IntStu	IntFacS	Score
Linear	96.59	84.50	12.12	100.00	17.89	0.00	72.29
Loess	90.00	75.04	18.57	90.53	19.83	20.88	68.25
Mean	93.295	79.77	15.345	95.265	18.86	10.44	70.27
Real	94.2	83.3	12.1	93.4	18.6	6.9	68.6
Difference	0.905	3.53	− 3.145	− 1.865	− 0.26	− 3.54	− 1.67

Table 19 Prediction of indicators and Overall Score for the year 2020 for the University of Buenos Aires

	AcRep	EmRep	FacStu	CitpFac	IntStu	IntFac	Score
Linear	100.00	100.00	85.82	2.47	74.85	58.17	73.77
Loess	78.34	84.64	75.17	3.26	60.36	52.49	61.30
Mean	98.17	92.32	80.49	2.86	67.61	55.33	67.53
Real	87.2	91.3	77.4	2.4	64.7	50.7	66
Difference	− 1.97	− 1.02	− 3.09	− 0.46	− 2.90	− 4.63	− 1.54

Table 20 Prediction of indicators and Overall score for the year 2020 for Pontificia Universidad Católica de Chile

	AcRep	EmRep	FacStu	CitpFac	IntStu	IntFac	Score
Linear	100.00	100.00	32.62	10.16	1.55	17.58	59.02
Loess	73.46	88.93	40.38	15.87	5.40	24.82	51.27
Mean	86.75	94.46	36.5	13.02	3.48	21.2	55.15
Real	85.2	95.5	28.6	13.6	4.2	19.4	53.4
Difference	− 1.53	1.035	− 7.9	0.58	0.73	− 1.8	− 1.75

Table 20 shows 8-point difference between the linear and non-linear regressions. We see that their Faculty-Student indicator worsened in the last year, down 4 points compared to 2019 and showing a difference -7.9 over our average. Finally, the difference between the average and the Final Score was -1.75 .

5.6 Universidad Nacional Autónoma de México (UNAM)

The next Latin American university is UNAM, currently in 103rd place, about to enter the Top 100; it has been rising steadily since 2011 from 222nd place.

In the linear regression, we observe two indicators with a positive slope and four with a negative slope. Their Citations per Faculty and International Students indicators are the lowest by far; the Academic Reputation and Employer Reputation scores are in the 90 s.

The non-linear regression shows different behaviour. Academic Reputation and Employer Reputation have fallen in recent years. This is a phenomenon that we have seen in several universities due to a change in methodology. International Faculty improved in the last year.

Table 21 shows that the most considerable difference was in the Faculty-Student ratio indicator, improving about 7

points between 2019 and 2020. International Faculty dropped more than 6 points in the last year. However, our prediction regarding the final score had a difference of 1.09, in the expected range.

5.7 Universidade de São Paulo (USP)

This university is currently in 116th place and has generally been improving every year.

Linear regression shows consistent behaviour in which Academic Reputation and Employer Reputation improved much compared to the other four indicators. Citations per Faculty does not appear to show much improvement but does not have a negative slope; the other three indicators subtly worsened.

For the Loess regression, we note again that after 2016, Academic Reputation and Employer Reputation were penalized and declined over the years and also Faculty-Student ratio. There is an improvement in Citations per Faculty in the last year and a smaller one in International Faculty.

Table 22 shows that the difference in the average prediction and the real value is low except for Employer Reputation with a value of -9.68 . This is quite high for this indicator, representing 10% of the Final Score weight. Our methodology has the greatest difference between the actual Final Score

Table 21 Prediction of indicators and overall score for year 2020 for Universidad Nacional Autónoma de México

	AcRep	EmRep	FacStu	CitpFac	IntStu	IntFac	Score
Linear	100.00	100.00	32.62	10.16	1.55	17.58	59.02
Loess	73.46	88.93	40.38	15.87	5.40	24.82	51.27
Mean	86.75	94.46	36.5	13.02	3.48	21.2	55.15
Real	85.2	95.5	28.6	13.6	4.2	19.4	53.4
Difference	− 1.53	1.035	− 7.9	0.58	0.73	− 1.8	− 1.75

Table 22 Prediction of indicators and Overall score for the year 2020 for Universidade de São Paulo

	AcRep	EmRep	FacStu	CitpFac	IntStu	IntFac	Score
Linear	100.00	100.00	28.44	31.92	4.54	6.85	62.32
Loess	79.97	65.97	26.82	41.04	3.55	11.04	53.60
Mean	89.98	82.98	27.63	36.48	4.05	8.95	57.96
Real	88.3	73.3	25.2	35.2	8.9	3.7	55.5
Difference	− 1.69	− 9.68	− 2.43	− 1.28	− 0.35	− 0.04	− 2.46

and the predicted average, − 2.46, more than two units for this university. The Employer Reputation indicator fell 7 points compared to 2019.

and design position strategies accordingly. Then we present a prediction model to forecast values and ranks for the various indicators and the overall score. We now discuss the results of the models fitted by machine learning algorithms.

6 Summary

The resulting model is wrapped in a framework used to predict unseen data in real-world operations. Our study predicted and evaluated Tecnológico de Monterrey's ranking performance, which has adopted a strategic Tec21 Educational Model, and other international universities using various strategies.

First, we studied the indicators of each university from 2011 to 2019. We carried out a linear regression to predict 2020 and a Loess (non-linear) regression also. Then with the Panel Data model, we predicted the Final Scores. We got two for each university; the means of the two are reported in Table 23.

Our predictions differed from the actual scores in the range of 3 units. We must be clear that in very close competitions like university rankings, decimals are crucial in the determinations of the final ranking. However, we believe that we have developed a transparent methodology for stakeholders to follow to get an excellent idea of their performance in future years.

7 Discussion

We presented a classification approach to define bands in which an institution can be sorted out using exploratory data analysis techniques. This analysis provides administrators a glimpse about how a university is doing with respect to other

7.1 Panel Data

One of the most relevant results was the validation of the Panel Data model. This method works because we have sectional data for a set of universities and longitudinal or temporal data for rankings from 2011 to 2020. Panel Data usually performs well when these conditions are satisfied, which was the case in our study. We applied cross-validation technique to prevent the presence of overfitting in the model. The predictive error in the two trained datasets, QS WUR and QS BSC, was very low in the different years. From this, we see that the model learns concerning the methodology and the individuals.

We believe that the Panel Data model fulfills the function of recognizing individual influences, compensating for each university's and city's historical performance with each indicator's weight.

After we carried out the training with machine learning models, they were trained to obtain categorical predictions of the universities' position and cities in the ranking.

For the QS WUR exercise, where the dataset was divided into two groups of 100 universities, the four algorithms achieved an accuracy above 0.95. This result encourages us that categorizing the universities in Top 100 and Top 200 groupings helps extract information about those in the best places in the ranking. Looking at the ten groups of 20, we see that only Random Forest achieved an accuracy above

Table 23 Summary of the universities' Final Score 2020 prediction

	Real 2020	Δ Linear	Δ Loess	Δ Mean	Current rank
Tec	48.50	3.5	– 2.6	– 0.45	158
CMU	74.80	0.2	2.56	– 1.38	48
UT Austin	68.60	3.69	– 0.35	– 1.67	65
UBA	66.00	7.77	– 4.4	– 1.54	74
PUC Chile	53.40	5.62	– 2.13	– 1.75	127
UNAM	58.8	2.37	– 4.56	1.09	103
USP	55.50	6.82	– 1.9	– 2.46	116

0.8. Thus, if we wanted to categorize a university more precisely, we would be wrong with a probability of 20% using the best-trained algorithm.

Regarding the Bayesian networks' training, we decided to stay with the most current and static structure (trained with data from 2020). We built the table of the relationship between nodes. We believe that this result is significant for those in charge of creating business strategies for universities because one can see the indicators' internal structure. Suppose a university invests in increasing its ranking, acquiring the best research equipment, and attracting international students. In that case, they will indirectly and positively impact their six indicators.

7.2 Model Evaluation

In the deployment section, we carried out seven experiments, each calculating the prediction of 2020 scores for different universities. We performed a linear regression, then a Loess regression, and subsequently calculated the average between them. This value was the one that was compared with the actual value published in QS WUR. Based on the differences, we evaluated our methodology.

Of the seven cases, the smallest difference in Final Score was observed for the Tecnológico de Monterrey and the largest difference for the Universidade De São Paulo. In five cases, our prediction was above the real value; thus, we can say that our model tends to be optimistic regarding a university's final score.

However, large differences can be found between the prediction and the real value of some indicators for the universities we tested. We believe that a methodological change announced by QS in 2017 might have strongly affected all universities. Significantly, Academic Reputation and Employer Reputation indicators fell for all the universities in 2017 and the following years.

Of the seven universities that were studied, five were Latin American, and we observed various trends. The indicator of Citations per Faculty member is usually the lowest, unlike the two American universities studied. The only exception was the Universidade de São Paulo, which presented scores

between 25 and 50 compared to 2–5 of its Latin American competitors; even so, it is below the University of Buenos Aires and UNAM in the ranking.

Another important observation is the International Students indicator. For Carnegie Mellon, this is its highest indicator. For the other six universities, it is among the three lowest, despite being an indicator with a weight of 5%. We believe that this finding reflects the excellent performance of a university in the top 50 in the world ranking.

Finally, we believe that in addition to predicting the Final Score and the universities' indicators, the best value of this work is to find the trends that allow a university to plan for long-term improvement as an institution. We know that each university's mission varies according to its nature and community; ultimately, it will always strive to provide better services to its students.

We should clarify that a formal deployment has not been carried out; that is, our methodology has been applied by an educational institution. However, we have made some proposals to universities that are growing as a point of comparison.

7.3 Related Work

Some work related to the analysis of university rankings was developed by Masao Mori [19] from the Tokyo Institute of Technology. He presents a study and analysis of how the World University Rankings scores distribute, defining criteria and weights for the THE and QS university rankings. Collecting relative scores from 800 universities of the THE WUR and 400 universities of the QS rankings, he presented histograms showing how the weights given by the methodologies proposed by the rankings fairly represent the universities, depending on how crowded each score is and varying the weights. QS resulted in being the single mode, which means that the score value was the most popular among most of the universities ranked.

Another work analysing QS World University Rankings more mathematically and computationally was the cluster analysis of universities done by Kathiresan Gopal [20] from

the University of Putra Malaysia. Using multivariable statistical techniques, this researcher showed that universities' Euclidian distance is another effective way to rank them, compared to the 200 best-ranked universities from QS. Clustering universities by Euclidian distance lets them learn what scoring in rankings do, how universities' positions can be interpreted, and how this influences the best universities in the world to keep their positions.

Muzakir Hussain [21] created an algorithm to aggregate rankings. He found correlations between different rankings and helped students attain better results by providing a recommendation system that considered many popular global rankings to build a complete and objective ranking model. While universities were focusing on increasing the most important rankings scores, his team tried to extract valuable information from the rankings relevant to students' specific needs. He proposed the Shimura Preference Order Rank Aggregation (SPORA) algorithm to aggregate many rankings and develop a useful recommending tool efficiently.

Another work with rankings analysed the results and the way the data was obtained from the source. Chengkai Shi [22] introduced the Computer Science Academic Rankings System (CSAR), which aims to extract information from rankings successfully. Information extraction is crucial when doing research. Shi's system collects data from Google Scholar, DBLP, ACM Digital Library, and Microsoft Academic and collects papers and authors' information. Then they start working on relating the topics with authors and papers. They measure the contributions and, finally, rank the authors and organizations. This kind of tool is an alternative to WUR (World University Rankings).

Szentirmai did a complete analysis of university rankings [23], considering cultural and geographical circumstances. His study examined the Times Higher Education WUR, Academic Ranking of World Universities, and the QS WUR, which are the most popular rankings. The top 200 universities from these three methodologies were analysed, and it was found that the top 10 universities in these rankings coincided. The United States dominated all the rankings. The reason for this phenomenon is that most university rankings use mono-dimensional systems with indicators that discriminate only research-intensive institutions. This work concludes that Europe must develop a system for international university comparisons with broader criteria to strive to be more competitive.

Anika Tabassum of the Bangladesh University of Engineering and Technology has been working on predictive models for university rankings, presenting different methodologies that implement learning algorithms with a newly proposed list-wise approach [24]. Depending on the data set, the indicators are broken down, showing behaviour by region, research area, gender, year, and numbers of students. Ultimately, they decided that splitting by country gave the best

results. Then they separated the last year of the data set from the others to be used as a test set and verified that the algorithm correctly predicted the following year. In conclusion, they rated the prediction algorithm based on outlier detection as acceptable.

Regarding our work compared to others focused on the world university rankings, we believe that our work stands out because the data we use is publicly available. All universities and administrators can replicate our analysis. We created a methodology that makes it easy to understand how far they are from the position they seek in the rankings. They can track their future performance with the proposed model. A disadvantage is that we did not provide tools for students or parents. Some of the other works try to help students weight the rankings to understand the various universities' different benefits that interest them. On the other hand, we depended on data availability and other approaches to rankings to get data from web resources such as Google Scholar and ACM digital libraries that help construct independent rankings.

To compare our results, we could not find works directly related to predicting ranking and scoring with the QS database. However, we highlighted the importance of using time series in studies like [25] and the work with THE data mentioned above. We also demonstrated the importance of correlating the indicators of the four most influential world rankings: QS, ARWU, THE, and URAP [26]. To address our research question, we found that the QS WUR dataset provided valuable information for studying and predicting university performance and its evolution over time. The best prediction model we found was Panel Data for the Final Score and the Random Forest model for ranking group prediction. In the different groups, we found trends that would allow universities to locate their most vital indicators and calculate the distance of their indicators from the targeted group to create long-term strategies. We found everything from exploratory analysis tools to predictive, machine learning, and probabilistic tools.

We believe that decision-making will support a comprehensive strategy to measure an institution's current position and its ability to project future improvements. We believe that this work's most important product is providing a methodology for universities that want to improve their ranking position and understand their performance over time. They can use information already available by the same ranking company, apply it to particular cases, and obtain results with low margins of error in future years.

8 Conclusion

Higher education has evolved because of globalization, putting institutions in a position to be compared internationally. With this work, we believe that we simplify this

comparative process with efficient metrics that allow knowing precisely the differences between the current and targeted ranking places. In addition to statistical metrics, we used clustering to separate universities from the Top 100 and know the distance between them, considering Citations and Academic Reputation. Finally, Panel Data became useful to know each institution's influence in the Final Score, improving the QS ranking prediction.

The Panel Regression with Fixed Effects helps improve the model's accuracy and deepen the impact of different methodological changes on each university's final performance. Similarly, the correlation coefficients resolve doubts about the most opportune areas of improvement for universities that want to enter a higher-ranking group. Universities can see the different degrees of correlation between the indicators and what areas they need to focus on.

Regarding the possibility of classifying universities into a group based on their ranking, we believe Random Forest is the best classification algorithm. It considers the universities' historical movement in the rankings, allowing the projection of the university's future scores and evaluating them to see if they are sufficient to attain their goal.

The Bayesian networks as a probabilistic method allow us to discover the relationships between the indicators. In this case, the Citations per Faculty member influences the Faculty-Student ratio, International Faculty, and probably Academic Reputation. International Students influence International Faculty, making sense because a student studying for a postgraduate degree could be hired later by the same institution. Thus, indirect actions can be taken to improve a university's performance in the rankings in the long term.

We believe that Panel Data outperforms other prediction methods. It analyses time trends, customizes the model to specific institutions, and provides an interpretable model that university administrators and other stakeholders can use to plan academic improvements, enhance reputation, and make decisions about collaboration and exchanges. Grouping universities by ranking position allows planners to identify the characteristics of institutions positioned in the highest part of the ranking and understand what it takes to join higher groups.

As future work, we recommend an application in which students can interact with the indicators and know the different universities' strengths depending on their interests. Other future work proposals include applying these models to other known datasets of university rankings such as THE and ARWU; each has different methodologies and numbers of indicators. We believe that our work can be transferable to these rankings. It is also possible to enrich the study by gathering complementary information from other available sources that may influence the rankings and help universities monitor their performance. This database can be internal,

based on teacher recruitment, international student enrolment, and publications. This information could supplement the decisions to refine improvements that lead to predicting group positions. Another concept that we explored was the learning of dynamic Bayesian networks. These results were not concluded due to time limitations, but we think that our time-dependent database could be useful.

We have also seen that universities' needs continue to evolve, and their success in delivering the best education and achieving the highest rankings will continue to evolve. That is why this work will continue to develop, depending on where higher education continues to advance.

Acknowledgements The authors thank Tecnológico de Monterrey (Tuition scholarship A01748866), CONACYT-PNPC (Student Grant Number 713372), and CONACYT-SNI (Researcher Grant Number 9804) for the financial support to conduct this study. We also want to thank the Intelligent Systems and Machine Learning research groups for their support and the Research Intelligence Department at Tecnológico de Monterrey for preparing and providing ranking data to perform this study. The authors acknowledge the technical support of Writing Lab, Institute for the Future of Education, Tecnológico de Monterrey, Mexico, in the production of this work.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Rybiński, K., Wodecki, A.: Are university ranking and popularity related? An analysis of 500 universities in Google Trends and the QS ranking in 2012–2020. *J. Market. High. Educ.* (2022). <https://doi.org/10.1080/08841241.2022.2049952>
2. Cantu-Ortiz, F.J.: *Research Analytics: Boosting University Productivity and Competitiveness Through Scientometrics*. CRC Press, Boca Raton (2018)
3. Tuesta, E.F., Bolaños-Pizarro, M., Neves, D.P., Fernández, G., Axel-Berg, J.: Complex networks for benchmarking in global universities rankings. *Scientometrics* **125**(1), 405–4251 (2020). <https://doi.org/10.1007/s11192-020-03637-9>
4. Valérie, L.: How do rankings impact higher education? Editor of Institutional Management in Higher Education Programme, OECD. IMHE. <https://www.oecd.org/education/imhe/39802910.pdf>
5. Chavez, M.D., Ceballos, H.G., Cantu-Ortiz, F.J.: A data analytics approach to contrast the performance of teaching (only) vs. research professors. *Int. J. Interact. Des. Manuf.* **14**, 1577–1592 (2020). <https://doi.org/10.1007/s12008-020-00713-5>
6. Cantú-Ortiz, F.J., Galeano Sánchez, N., Garrido, L., Terashima, H., Brena, R.: An artificial intelligence educational strategy for the

- digital transformation. *Int. J. Interact. Des. Manuf.* **14**, 1195–1209 (2020). <https://doi.org/10.1007/s12008-020-00702-8>
7. Yudkevich, M.: Global university rankings as the Olympic games of higher education. *The Global Academic Rankings Game 1–11*, (2016). eBook ISBN: 9781315677170
 8. Grewal, R., Dearden, J.A., Lilien, G.L.: The university rankings game: Modeling the competition among universities for ranking. *Am. Stat.* **62**(3), 232–237 (2008). <https://doi.org/10.1198/000313008X332124>
 9. Benito, M., Gil, P., Romera, R.: Funding, is it key for standing out in the university rankings? *Scientometrics* **121**(2), 771–792 (2019). <https://doi.org/10.1007/s11192-019-03202-z>
 10. Marconi, G.: Rankings, accreditations, and international exchange students. *IZA J. Eur. Labor Stud.* **2**, 5 (2013). <https://doi.org/10.1186/2193-9012-2-5>
 11. Vetrova, I.F., Amerslanova, A.N., Yuretskaya, Y.S.: An overview of the main types of university control in the leading countries of the world. *Lect. Notes Netw. Syst.* **280**, 996–1004 (2021). https://doi.org/10.1007/978-3-030-80485-5_111
 12. Moed, H.F.: A critical comparative analysis of five world university rankings. *Scientometrics* **110**(2), 967–990 (2016). <https://doi.org/10.1007/s11192-016-2212-y>
 13. Quacquarelli Symonds, QS Intelligence Unit. QS World University Rankings, QS-WUR 2020. Retrieved June 30, (2020). <https://www.topuniversities.com/university-rankings/world-university-rankings/2020>
 14. Mcaleer, M., Nakamura, T., Watkins, C.: Size, internationalization, and university rankings: evaluating and predicting times higher education (THE) data for Japan. *Sustainability* **11**(5), 1366 (2019). <https://doi.org/10.3390/su11051366>
 15. Pride, D., Knoth, P.: Peer review and citation data in predicting university rankings, a large-scale analysis. *Digit. Libr. Open Knowl. Lect. Notes Comput. Sci* (2018). https://doi.org/10.1007/978-3-030-00066-0_17
 16. Schlögl, C.: European doctoral forum at the 14th international society of scientometrics and informetrics conference. *Bull. Am. Soc. Inf. Sci. Technol.* **40**(1), 17–18 (2013). <https://doi.org/10.1002/bult.2013.1720400106>
 17. Dobrota, M., Bulajic, M., Bornmann, L., Jeremic, V.: A new approach to the QS university ranking using the composite i-distance indicator: Uncertainty and sensitivity analyses. *J. Am. Soc. Inf. Sci.* **67**(1), 200–211 (2016). <https://doi.org/10.1002/asi.23355>
 18. Lanchobarrantes, B.S., Cantu-Ortiz, F.J.: Quantifying the publication preferences of leading research universities. *Scientometrics* **126**, 2269–2310 (2021). <https://doi.org/10.1007/s11192-020-03790-1>
 19. Shearer, C.: The crisp-dm model: the new blueprint for data mining. *J. Data Warehous.* **5**(4), 13–22 (2000)
 20. Mori, M.: How do the scores of world university rankings distribute?. 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), pp. 482–485, (2016). <https://doi.org/10.1109/IIAI-AAI.2016.36>
 21. Gopal, K., Shitan, M.: Cluster analysis of top 200 universities in mathematics. In: 2015 International Symposium on Mathematical Sciences and Computing Research (ISMSC), pp 408–413, (2015). <https://doi.org/10.1109/ISMSC.2015.7594089>
 22. Hussain, M.M., Rahman, S.A., Beg, M.S., Ali, R.: Cognitive fuzzy rank aggregation for non-transitive rankings: an institute recommendation system case study. In: 218 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), pp 358–365. <https://doi.org/10.1109/ICCI-CC.2018.8482073>
 23. Shi, C., Quan, J., Li, M.: Information extraction for computer science academic rankings system. In: 2013 International Conference on Cloud and Service Computing, pp. 69–76, (2013). <https://doi.org/10.1109/CSC.2013.19>
 24. Szentirmai, L., Radacs, L.: World university rankings qualify teaching and primarily research. In: 2013 IEEE 11th International Conference on Emerging eLearning Technologies and Applications (ICETA), pp. 369–374, (2013). <https://doi.org/10.1109/ICETA.2013.6674461>
 25. Tabassum, A., Hasan, M., Ahmed, S., Tasmin, R., Abdullah, D. M., Musharrat, T.: University ranking prediction system by analyzing influential global performance indicators. In: 9th International Conference on Knowledge and Smart Technology (KST), pp 126–131, (2017). <https://doi.org/10.1109/kst.2017.7886119>
 26. Ramadhan, A., Masayu, L.K.: Ranking prediction for time-series data using learning to rank (case study: top mobile games prediction). In: 2014 International Conference of Advanced Informatics: Concept, Theory, and Application (ICAICTA), pp 214–219, (2014). <https://doi.org/10.1109/icaicta.2014.7005943>
 27. Uslu, B.: A path for ranking success: what does the expanded indicator-set of international university rankings suggest? *High. Educ.* **80**(1), 949–972 (2020). <https://doi.org/10.1007/s10734-020-00527-0>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).

Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.

We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.

While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.

These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.

To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.

If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com