The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 3, No. 1. Online ISSN: 2831-4794
Available at https://www.isods.org/publication/

Vietnam University Ranking in Data Science and Artificial Intelligence through Principal Components Analysis

Linh Dinh[a], Thanh Ha[a], Cuong Do[b], Nam Vu[a], Tan Nguyen[a], and Robert Pham[a]

[a]George Washington Institute of Data Science and Artificial Intelligence (ISODS), Lexington, MA, USA
[b]Department of Mathematics, University of Connecticut, 341 Mansfield Road U1009, Storrs, Connecticut 06269-1009

ABSTRACT

As Data Science (DS) and Artificial Intelligence (AI) disciplines are rising within the global technological sphere, students are actively seeking options that potentially enhance their knowledge of these domains. Nevertheless, the absence of a robust ranking system in Vietnam poses a considerable challenge for prospective college students aspiring to pursue careers in Data Science or Artificial Intelligence. The team is driven by the primary goal of establishing a comprehensive ranking framework to aid students in selecting universities conducive to their academic and career pursuits in DS and AI. The methodology comprises a systematic process of gathering and transforming data, conducting analytical tests such as correlation tests and comparative analysis of data points, and constructing linear regression models and principal components analysis utilizing Excel and R Studio employed for execution. Principal Components Analysis (PCA) was chosen as the main method to approach the dataset in ranking various features and concluding features that are significant in determining quality universities. The study successfully yields the first ranking system within the Vietnamese Data Science and Artificial Intelligence college sectors that captures the unique advantages and values of 34 Vietnamese universities. Notably, students, lecturers, and lawmakers who seek counsel in the dynamic field of data science and artificial intelligence education can gain useful perception and development from the ranking system, exponentially making more comprehensive evaluations for future studies of data and AI.

Keywords: Data Science, Artificial Intelligence, ranking, Principal Components Analysis

1. INTRODUCTION

In modern education critical fields such as Data Science and Artificial Intelligence (DS/AI) have established their popularity, with the expertise of future professionals, shaping the pivotal role of universities. Noticing the significance, these disciplines were based on to create a comprehensive ranking system for universities. Due to the immense impact, these technologies had on multiple industries lead to focusing on DS/AI, imperative to appraise and accept the educational prospect.

The chosen methodology for this study involves the application of Principal Components Analysis, a powerful statistical technique that enables the identification of essential features influencing the quality of education. By leveraging PCA, the research seeks to unveil the intrinsic factors contributing to the excellence of DS/AI programs in Vietnamese universities. This analytical approach allows for a nuanced understanding of the educational landscape and ensures a data-driven assessment.

This study develops a ranking system for universities in Vietnam that reflects a few unique features in their strength and quality in DS/AI. In terms of the factors participating, valuable recommendations are proposed for future students, lecturers, and AI policymakers. This study is very meaningful to guide colleges in making precise strategies in the rapidly changing field of Data Science and Artificial Intelligence.

2. METHODOLOGY

2.1 Data Collection and Data Processing

In this exploratory study, researchers systematically collected data from 34 educational institutions in Vietnam offering Data Science and Artificial Intelligence (DS/AI) programs, including courses, majors, minors, or concentrations. The raw data has been processed through sorting and filtering procedures for the purpose of dealing with non-normal N.N.A. entries and missing values. Thus, the data have been homogeneous by correcting discrepancies and inconsistencies, which have been detected by monitoring. The team chose to select eleven variables in building the ranking based on five categories:

Academic credentials: number of PhDs, and professors.
Research impact: number of citations, H-index, i10 index.
Publication metrics: number of national publications, international publications.
Educational quality: English courses, threshold entry scores.
Collaborative efforts: international partnerships, exchange programs.

The five categories and their seven variables were chosen based on the team's extensive research of suitable various school ranking methodologies. For instance, according to A Ranking of Vietnam's Universities The 2016-2017 Report, the majority of rankings available in Vietnam, such as Webometrics, Scientometrics, or U.S News, contain important factors regarding publication metrics, and research impact are highlighted as significant, aligning with the academic credentials of the faculty. [1][2]

Consequently, academic credentials, research impact, and publication metrics are the first categories the researchers developed. Within those categories, the first variables examined are the number of PhDs and professors within each school, as these indicators reflect the academic quality of the faculty and enhance the overall educational experience for students. Then researchers delved into the research impact by assessing faculty credentials through citation counts, H-index, and i10 index. Last but not least, publication metrics include the number of national and international publications of the schools as a whole.

After developing the three categories with their corresponding variables, the team took a look at educational quality measurements in Vietnam. A crucial performance indicator in the Vietnamese market is the threshold entry scores, where universities in Vietnam use scores from entrance exams - High School Graduation Examination - to assess student's abilities to be a part of the community. This metric is included in evaluating a school's educational quality, as admissions are becoming increasingly competitive, and the required entrance scores are rising. [3] With the increasing number of schools aiming for international standards, the availability of English courses qualified for international programs has become significant. Therefore, we evaluated the school's offerings of English courses in the DS/AI program. Researchers also examined their collaborative efforts with international partners and exchange programs to enhance student education and prepare them for global success.

In essence, these categories — Academic Credentials, Research Impact, Publication Metrics, Educational Quality, and Collaborative Efforts — reflect the factors typically considered by students, parents, and reputable ranking systems when evaluating universities. The goal was to select variables that are crucial, accessible, and suitable for analysis. This ensures a holistic evaluation of the programs rather than focusing on one or two aspects alone and also ensures the data is consistent across the board as these features are widely accepted in academia for assessing impact, allowing for a standardized comparison across different institutions.

Data is then collected from various sources such as school websites, third-party information on the metrics, or academic sources like Google Scholar, Research Gate, etc. The features are all collected up-to-date within the time in the year in which the research was conducted - 2023, with the exception of the entry threshold that the team was able to collect data from 2021 - 2023. Group 1 comprises continuous numerical variables, which will undergo standardization to a range between 0 and 1 and will subsequently be integrated into the principal

component analysis. On the other hand, Group 2 encompasses other discrete numerical values, specifically 0 and 1 representing yes and no, respectively.

One special case regarding variables is the data processing of citations, H-index, and i10. These three variables have been collected based on each citation, H-index, and i10 of faculties within the department up-to-date at the time of data collection in 2023. After all faculties publication metrics have been collected, they then would be transformed into averages across all researchers within the department - becoming a representation across the entire university through the measure of central tendency. This ensures that the general trend regarding the research quality of a school is captured and able to compare with others within the same field. The team also utilizes the same concept for the entry threshold by averaging the scores of each school from 2021 to 2023.

The data intended for the unsupervised model underwent a standardized process to ensure equality contribution to the analysis using Eq.1:

$$\frac{\text{Value} - \min(\text{All Data Points})}{\max(\text{All Data Points}) - \min(\text{All Data Points})} \tag{1}$$

PCA then would be used to identify important factors affecting the quality of DS/AI education, prioritizing the parameters with the most significant differences, and using standardized data to estimate the principal factors. The PCA results were scrutinized to examine the contribution of each factor to the total variance [4]. A side note is that sometimes data points that come from different sources or universities are not consistent with each other, some academies have their own entry points system. Lead to conversion is needed since consistency is needed for a successful and accurate correlation and PCA testing. An example such as RMIT University which doesn't employ the standard 100-point system but instead uses its 10-point structure. In order to convert the authors to compare the maximum and minimum scores of RMIT and another academy from there it can result in some theoretical entry scores, which is close enough for consistency. However, being unorthodox is the best way to overcome this milestone since no standard conversion methods exist.

2.2 Principal Component Analysis

Initially, the team acquired the necessary packages: dplyr, psych, ggrepel, and tidyverse for data manipulation, statistical tests, and visualization. The universities' dataset will then be exported as a data frame. Prior to this, the researchers categorized the data into two groups: continuous and discrete. Continuous constitutes the majority of the dataset and is deemed efficient for Principal Component Analysis as the continuous variables are numeric non-binary, opposite with discrete variables that are binary. The continuous variables will be extracted as the following standardized variables from the main data frame for analysis: Number of PhD, Number of Professors, Number of International papers, averaged number of Citation, H Index, i10, and Average Entry Threshold across the years 2021, 2022, and 2023. This subset of variables is termed "Test_features" and is the main testing set for PCA.

The Principal Component Analysis test starts with checking the variance of the dataset to ensure that the variables are significant for analysis, allowing the team to rank the importance of different features [5]. The researchers applied the Bartlett test to assess the equality of variances among the data samples. The obtained P-value was $3.30561e^{-26}$, indicating that not all groups exhibit the same variance and possess a significant relationship with each other, rendering it suitable for further analysis. Subsequently, the team conducted the first Principal Component Analysis - PCA test with no rotation to determine the significance of principal components and identify the primary distribution of the data in terms of variance [6].

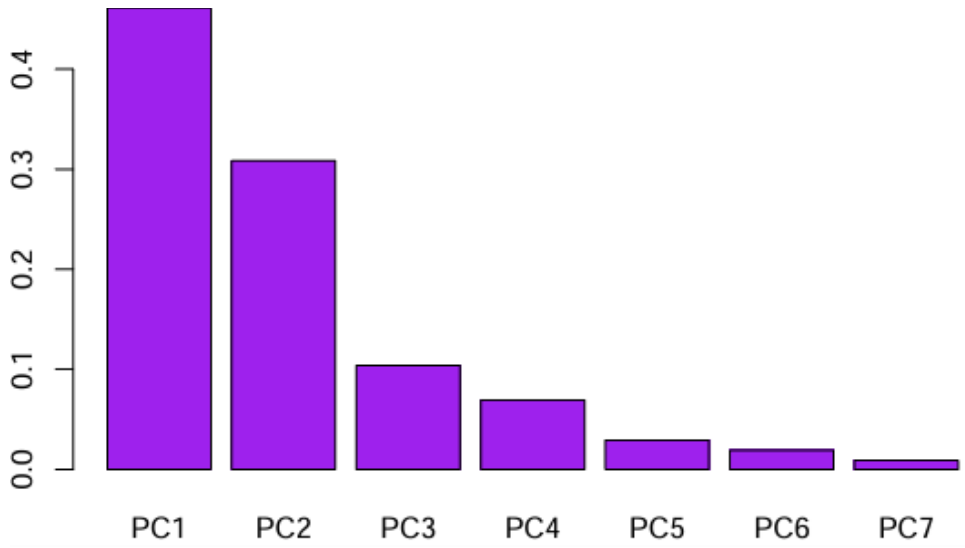


Figure 1: Variance of data explained based on principal components

From the insights obtained from the graph, PC1 and PC2 hold the highest values for variance. These features will then be assessed through Kaiser's rule to observe further how variance manifests across the data. The authors examine eigenvalues from the variance test and identify those exceeding 1 as significant. Upon analysis, both PC1 and PC2 meet the criteria set by the Kaiser test with corresponding eigenvalues: 3.22755455 and 2.15892624.

Consequently, the variance test is re-conducted to focus solely on PC1 and PC2 and remove other insignificant principal components for ranking [7]. Together, they account for the majority of the data variance and exhibit close association: PC1 explains approximately 46%, PC2 31% of the proportion of the data, and with a total of approximately 77% data is explained through these two components.

```
pca_no_rotate$values

## [1] 3.22755455 2.15892624 0.72798366 0.48569848 0.20419524 0.13419847 0.06144337
pca_no_rotate$Vaccounted #both pass! PC1 & PC2

##                             PC1       PC2       PC3       PC4        PC5
## SS loadings           3.2275546 2.1589262 0.7279837 0.4856985 0.20419524
## Proportion Var        0.4610792 0.3084180 0.1039977 0.0693855 0.02917075
## Cumulative Var        0.4610792 0.7694973 0.8734949 0.9428804 0.97205117
## Proportion Explained  0.4610792 0.3084180 0.1039977 0.0693855 0.02917075
## Cumulative Proportion 0.4610792 0.7694973 0.8734949 0.9428804 0.97205117
##                              PC6         PC7
## SS loadings           0.13419847 0.061443372
## Proportion Var        0.01917121 0.008777625
## Cumulative Var        0.99122238 1.000000000
## Proportion Explained  0.01917121 0.008777625
## Cumulative Proportion 0.99122238 1.000000000
```

Figure 2: First PCA test to determine the principal components that pass the Kaiser test - No Rotation, 1st Dimension

```
##                            PC1       PC2
## SS loadings           3.2275546 2.1589262
## Proportion Var        0.4610792 0.3084180
## Cumulative Var        0.4610792 0.7694973
## Proportion Explained  0.5991954 0.4008046
## Cumulative Proportion 0.5991954 1.0000000
```

Figure 3: The second variance test after only PC1 and PC2 were chosen. PC1 explains approximately 46% and PC2 31% of the proportion of the data

An examination of the correlation existing between the variables under consideration based on the two Principal Components is the next step in the analysis. Notably, the correlation coefficients exhibit a range between approximately 0.4 to 0.7, which is more than 0.3 - the threshold to be able to run PCA – and indicative of varying degrees in variables' relations, ranging from moderate to robust associations.

Variables	PC1	PC2
PhD	0.7316271	0.3913399
Professors	0.6587407	0.6426347
Citation	0.6762535	-0.6554656
H-index	0.7621623	-0.5052323
i10	0.7512550	-0.6209006
International Publications	0.6494094	0.5889049
Threshold Entry Score	0.48374458	0.4190210

Table 1: Correlation between variables (No Rotation, 1st Dimension)

Looking at a clear lens through graphical representations, discernible patterns emerge – where H-index, Citations, and i10 manifest a contrasting orientation compared to the number of PhDs, Professors, International Publications, and Scores. This dichotomy suggests a positive correlation among variables such as Entry Threshold, professors, PhDs, and international publications. Similarly, an affirmative relationship is evident between H-index, Citations, and i10. After observing relationships between variables on the first dimension where the projection of the points within the data is the most spread out, the team looked at the perpendicular line to the first dimension - rotate to the second dimension that accounts for the second highest variance [8].

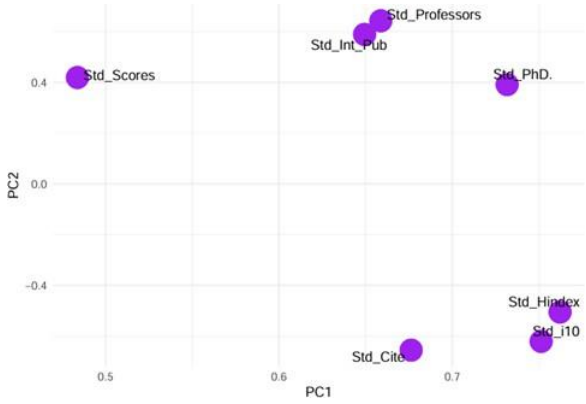


Figure 4: Visualizing Relationships Among Features Using PC1 & PC2 (No Rotation, 1st Dimension)

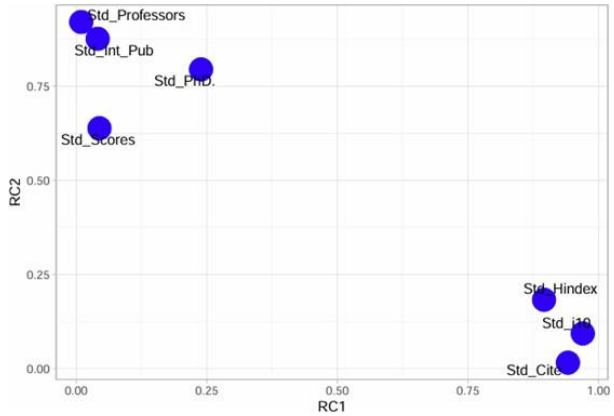


Figure 5: Visualizing Relationships Among Features Using PC1 & PC2 (Rotation, 2nd Dimension)

The graphical representation of the second dimension exhibits a widened spread variance, with variables maintaining their relative positions albeit undergoing slight rotation - a favorable and consistent observation compared to the first-dimension rotation. In order to confirm this assertion, a final examination is conducted through a comparison of commonalities. Comparison of commonalities aims to examine the extent of variance loss between the non-rotated, 1st Dimension vertical and rotated 2nd Dimension horizontal configurations. The findings indicate minimal differences irrespective of rotation, suggesting uniform variances across the dataset.

Variables	Commonalities no rotate	Commonalities w/ rotate
PhD	0.6884252	0.6884252
Professors	0.8469187	0.8469187
Citation	0.8869540	0.8869540
H-index	0.8361510	0.8361510
i10	0.9499017	0.9499017
International Publications	0.7685416	0.7685416
Threshold Entry Scores	0.4095886	0.4095886

Table 2: Assessing Variance Loss Between Non-Rotated (1st Dimension Vertical) and Rotated (2nd Dimension Horizontal) Configurations.

Last but not least, the principal component analysis then shifts to its ranking phase. However, the researchers in this particular case focus on obtaining the scores - distance from point O where two PCs meet to PC1 & PC2 [6] - for the final calculation of ranking where they will be combined with other features. The ranking from the PCA will only serve as a comparative metric for the team's final result, meaning it will not be included in the actual ranking. Its calculation can be explained through the proportion of variances attributed to the Principal Components, multiplied by the scores of the six factors, and applied to both rotate and non-rotate configurations in Eq.2 and Eq.3 [7]:

$$Ranking = (Score\ of\ PC1 \times Variance\ of\ PC1) + (Score\ of\ PC2 \times Variance\ of\ PC2) \quad (2)$$

$$Ranking = (Score\ of\ RC1 \times Variance\ of\ RC1) + (Score\ of\ RC2 \times Variance\ of\ RC2) \quad (3)$$

- PC represents the principal component in the first rotation (1st Dimension - highest variance).
- RC represents the principal component in the second rotation (2nd Dimension - second highest variance and perpendicular with 1st Dimension).

With the completion of the ranking process, numerical labels are assigned to each subset within the Test_feature subset for organizational clarity and future comparison.

Two distinct rankings have been predicated upon PCs and RCs. The initial ranking corresponds to the scenario where principal components remain non-rotated, while the subsequent ranking pertains to PC rotation. It is expected that discrepancies will emerge between these rankings given the expanded variance range associated with rotated components. [10] However, upon examination, the rankings exhibit notable similarities with little differences between the rankings of the schools. This observation underscores the robustness of the ranking methodology, even with rotational adjustments. It further validates PCA as an effective method for ranking universities, demonstrating that the scores are reliable and can be confidently integrated with other ranking factors. From this juncture, the derived scores and rankings can be effectively integrated with additional features not considered in the Principal Component Analysis.

2.3 Additional Features

Analysts determine the weight assigned to each element through a careful process. For standardized elements in Principal Component Analysis, examine the results of the variance test for both Principal Component 1 (PC1) and Principal Component 2 (PC2). The evaluation is based on their respective ability to recognize data sets in both rotated and non-rotated scenarios. The mean is then calculated, and the integrated index is 0.77, which can be referred back to figure 2.

For the remaining portion of the dataset: 6 groups are remaining, subsequently, a correlation test is conducted to elucidate their interrelationships.

```
> corr.test(All_features)
Call:corr.test(x = All_features)
Correlation matrix
                        Internship English International_Partnership Exchange_Program Num_Na_Pub
Internship                    1.00    0.24                     -0.06             0.25      -0.09
English                       0.24    1.00                      0.24             0.16       0.18
International_Partnership    -0.06    0.24                      1.00             0.55      -0.05
Exchange_Program              0.25    0.16                      0.55             1.00      -0.01
Num_Na_Pub                   -0.09    0.18                     -0.05            -0.01       1.00
Sample Size
[1] 36
Probability values (Entries above the diagonal are adjusted for multiple tests.)
                        Internship English International_Partnership Exchange_Program Num_Na_Pub
Internship                    0.00    1.00                      1.00             1.00          1
English                       0.16    0.00                      1.00             1.00          1
International_Partnership     0.71    0.16                      0.00             0.01          1
Exchange_Program              0.14    0.36                      0.00             0.00          1
Num_Na_Pub                    0.62    0.29                      0.76             0.95          0

 To see confidence intervals of the correlations, print with the short=FALSE option
```

Figure 6: Correlation test between remaining factors in the dataset.

In this next section under analysis, the fraction that remains will also be separated similarly using all other records still left; it forms 1 – 0.77 = 0.23. Next involves keeping the same weight distribution as had been used before based on how important relationships were inside such a dataset. In particular, International_Partnership or Exchange_Program have emerged as essential components having weights of 0.2 each following their discoverable salience vis-à-vis relative prominence in relationships. A suitable method is followed for other variables; namely, a fractional weight of 0.023/3 = 0.069 is designated for each of them. However, due to their relative importance which makes its perceived insignificance since half of this weight (0.0115) is assigned to Num_Na_Pub.

As for Internship and English, their weights are 0.023 each, to say that they have a significant impact on the present state of affairs. Among all variables in the model, Num_Na_Pub is the variable represented with the least data points. The decision was made to allocate half of this weight (0.023/2) towards partnership and exchange that matter most in order to reduce its effects substantially. Distribution of weights, as such, ensures an accurate

reflection of respective levels of importance across dataset variables thus adding considerable robustness to future analyses.

2.4 Combining PCA and Outside Features

Finally, the teams utilize the equation that represents the produced total number of consistent scores and weights to calculate the scores of the universities in Eq.4:

$$\sum_{i=1}^{n}(xi \cdot wi) \tag{4}$$

- n is the number of elements (scores and weights).
- xi represents the score for the i-th element.
- wi represents the weight for the i-th element.

The scores of these and the weights in the equation are important to identify a complex interrelationship, and thus adequately assess the combinatorial effects. For example, this composite framework is vital for moving forward to the next stage of their work plan where they utilize PCA and external sources to give valuable insights. With that, it is not just deepening the analysis but also leads to sound data interpretation which enhances strictness and reliability of analysis results.

3. RESULTS AND DISCUSSION

Following the completion of ranking through both PCA and additional features, we have successfully compiled a ranking for 36 schools specializing in Data Science and Artificial Intelligence. Among them, the University of Science, Vietnam National University Hanoi secured the top position with a score of 0.89 out of 1, while the Thai Nguyen University Of Information And Communication Technology (ICTU) attained the lowest score of -0.32.

School Name	Scores	Ranking
University of Science, Vietnam National University Hanoi	2.46	1
International University - Vietnam National University HCM City	1.05	2
University of Science - VNUHCM	0.86	3
Ton Duc Thang University	0.81	4
FPT University	0.72	5
Hanoi University of Science and Technology	0.70	6
University of Information Technology - VNUHCM	0.62	7
Thuy Loi University	0.52	8
UEH - University of Economics Ho Chi Minh City	0.44	9
Phenikaa University	0.44	10
Duy Tan University	0.42	11
RMIT University	0.39	12
National Economics University (NEU)	0.36	13
HCMC University of Technology and Education	0.33	14
HUTECH University	0.31	15
University of Science and Technology of Hanoi (Vietnam-France University)	0.30	16
Dong A University	0.25	17
HCMC University of Industry and Trade	0.18	18
Vietnam National University of Agriculture	0.14	19
Van Lang University	0.09	20
Hanoi University of Mining and Geology	0.02	21
HCMC Open University	0.00	22
University of Transport and Communications - HN	-0.01	23
Hong Bang International University (HIU)	-0.01	24
University of Transport and Communications - HCM	-0.01	25
Industrial University of Ho Chi Minh City	-0.04	26
CTUT - Can Tho University of Technology	-0.06	27
University of Economics - Technology for Industries (UNETI)	-0.18	28
Thu Dau Mot University	-0.19	29
Quy Nhon University	-0.21	30
Dalat University	-0.22	31
UEF - University of Economics and Finance Ho Chi Minh City	-0.22	32
Hue University	-0.27	33
Thai Nguyen University Of Information And Communication Technology (ICTU)	-0.32	34

Table 3: Final ranking of DS/AI schools in Vietnam.

The ranking is built upon scores from principal component analysis and combined with weight from other features outside of PCA. There are interesting insights the researchers are able to discover from the analysis, particularly in specific instances such as FPT and Ton Duc Thang University outperforming Hanoi University of Science and Technology while in terms of reputation, Hanoi University of Science and Technology has long been known as one of the most prestigious universities for students in STEM fields [9]. In order to explain this phenomenon, the authors aim to understand the underlying rationale for these discrepancies by determining the most important factor in influencing the ranking.

Linear Regression analysis and Predictive Power Score analysis are used in order to gain insight into the importance of features to the scoring from the Principal Component Analysis, which determines the final ranking that was accumulated from the scoring of features within the test and outside binomial features. There is also the fact that features within the PCA are linear and suitable for this type of regression for us to see the significance

of various features to the final ranking - rotate and non-rotate:

```
Call:
lm(formula = score_no_rotate ~ . - rank_yes_rotate - rank_no_rotate -
    score_yes_rotate, data = Test_features)

Residuals:
       Min         1Q     Median         3Q        Max
-1.523e-16 -5.113e-17  1.897e-17  5.238e-17  1.417e-16

Coefficients:
                Estimate Std. Error    t value Pr(>|t|)
(Intercept)    -6.092e-01  3.209e-17 -1.898e+16   <2e-16 ***
Std_PhD.        6.189e-01  8.695e-17  7.118e+15   <2e-16 ***
Std_Professors  9.783e-01  1.584e-16  6.178e+15   <2e-16 ***
Std_Cite        1.276e-02  1.452e-16  8.791e+13   <2e-16 ***
Std_Hindex      1.291e-01  9.905e-17  1.304e+15   <2e-16 ***
Std_i10         5.957e-02  1.355e-16  4.398e+14   <2e-16 ***
Std_Int_Pub     9.145e-01  1.324e-16  6.906e+15   <2e-16 ***
Std_Scores      3.986e-01  5.205e-17  7.659e+15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.051e-17 on 28 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.374e+32 on 7 and 28 DF,  p-value: < 2.2e-16
```

Figure 7: Linear Regression Analysis for non-rotate features from PCA.

```
Call:
lm(formula = score_yes_rotate ~ . - rank_no_rotate - score_no_rotate -
    rank_yes_rotate, data = Test_features)

Residuals:
       Min         1Q     Median         3Q        Max
-1.397e-16 -8.165e-17 -1.621e-17  5.540e-17  3.880e-16

Coefficients:
                Estimate Std. Error    t value Pr(>|t|)
(Intercept)    -7.560e-01  5.011e-17 -1.509e+16   <2e-16 ***
Std_PhD.        4.748e-01  1.358e-16  3.498e+15   <2e-16 ***
Std_Professors  5.822e-01  2.473e-16  2.354e+15   <2e-16 ***
Std_Cite        4.917e-01  2.266e-16  2.169e+15   <2e-16 ***
Std_Hindex      4.532e-01  1.547e-16  2.931e+15   <2e-16 ***
Std_i10         4.064e-01  2.115e-16  1.922e+15   <2e-16 ***
Std_Int_Pub     5.640e-01  2.068e-16  2.728e+15   <2e-16 ***
Std_Scores      2.512e-01  8.127e-17  3.091e+15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.257e-16 on 28 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 9.367e+31 on 7 and 28 DF,  p-value: < 2.2e-16
```

Figure 8: Linear Regression Analysis for rotate features from PCA.

Observing the linear regression analysis result where the rankings are the dependent variables in each case and other features are the independent variables, with the P - value less than 0.01, the model is suitable for assessment. With adjusted R squared being 1, the two models are a perfect fit for predicting the scoring from Principal Component Analysis, which is fully expected as all variables that are within the process of utilizing PCA are accounted to the scoring in both rotations, then became a part of the final ranking. Therefore, the team decided to move on to the Predictive Power Score of the features to understand further their predictive power towards the scoring, as the goal is to identify variables with strong predictive power:



Figure 9: Predictive Power Score (PPS) test to identify variables with strong predictive power for ranking.

Accessing the predictive power of variables with two fields: score_yes_rotate and score_no_rotate as the target, which are scores in rotating and non-rotating in PCA, two features that hold the most predictive power is the number of PhDs and H-index with an average of 0.29 and 0.26 from two rotation. Following those two measures is the number of Professors with an average of 0.24. Both citations and international publications are at 0.21, i10 at 0.20, and the threshold entry score comes in last at 0.17.

Furthermore, the team also looks at other features that affect the final scoring that determines the ranking through linear regression analysis and predictive power score analysis:

```
Call:
lm(formula = Scores ~ ., data = Cor_features)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52208 -0.27387  0.00667  0.14462  0.68791

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.3360     0.1902  -1.767   0.0875 .
Std_Na_Pub                 1.4888     0.2828   5.263 1.11e-05 ***
Internship                -0.1276     0.1630  -0.783   0.4399
English                    0.1989     0.1269   1.568   0.1275
International_Partnership  0.2720     0.1868   1.456   0.1558
Exchange_Program           0.2785     0.1557   1.788   0.0838 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3457 on 30 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.6295,    Adjusted R-squared:  0.5678
F-statistic: 10.19 on 5 and 30 DF,  p-value: 8.838e-06
```

Figure 10: Linear Regression Analysis for remaining features.



Figure 11: Predictive Power Score (PPS) test for remaining features.

While the weight has been pre-determined by the researchers, there is unexpected significance from features to the final scoring. Looking at both the linear regression model and the predictive power score test, national publications have the most significant P-value of less than 0.01 with a PPS of 0.02, indicating that schools that have national publications in Data Science or Artificial Intelligence have an influence on the ranking, although the predictive power is little to none, which is expected due to its weight. This will place the National Publication feature behind the Entry Threshold. Aside from national publications, in the regression model, the Exchange Program also has a P-value of 0.08, which is significant with a confidence level of 90%. PPS; however, tells a different story with the Exchange Program - where its predictive power with the final scoring is 0 - meaning there is no predictive value. This means that the exchange program factor does have a statical significant relationship with the scoring, but it will not be able to predict the scoring by itself, and the team will place the feature behind the National Publications.

Another interesting key point in the differences between these two tests has to do with the variable English, where its P-value is 0.12 and too high for there to be any significance towards the scoring, while its PPS is 0.07 - little predictive power for scores. These differences can be accounted for by the fact that the sample size of the data is limited. Nonetheless, due to the fact that there is no significant statistical relationship between the variable and scoring, combined with its low predictive power, it is safe to assume that having English in the curriculum of Data Science or Artificial Intelligence has little to no effect on determining the ranking of the universities. For the rest of the features: Internship and International Partnership, from both regression analysis and predictive power score analysis, they have no significance on the scoring, which also ties with having no influence on the final ranking.

To conclude, the study explains the ranking by differences in notability of factors in determining the scores, the ranking for Data Science or Artificial Intelligence universities in Vietnam. The number of PhDs has the most impact on the ranking, followed by the H - Index of faculties, and the number of Professors, which makes the top three and other features. Therefore, although there are schools that have a huge reputation for Data Science or Artificial Intelligence programs, differences in the number of PhDs, H - index of the faculties, and Professors can change the placement of the school on the schools' order. This means that - academic credentials from faculty and research impact from the H-index can be the top three factors that determine the quality of a school in Data Science and Artificial Intelligence.

4. CONCLUSION

Based on the ranking, the University of Science, Vietnam National University Hanoi secured the top position with a score of 0.89 out of 1 in the ranking for Data Science or Artificial Intelligence universities in Vietnam, while the Thai Nguyen University Of Information And Communication Technology (ICTU) attained the lowest score of -0.32. The reasoning for these results can be traced back to the factors that have the most significance on the ranking, which are: the number of PhDs and H-Index of faculties, followed by the number of Professors. This means that the number of PhDs, Professors, and their research quality measure such as the H-Index emerge as the top three factors that students should prioritize when selecting their ideal school in terms of quality. At the same time, colleges and universities are encouraged to further enhance their faculty quality to attract students and advance their standing within the Data Science and Artificial Intelligence field.

The ranking was able to achieve a concise ranking of universities within the Data Science and Artificial Intelligence sector in Vietnam by cooperating with various features and measuring their influence on the quality of one's school, making it a reliable first step in further enhancing DS/AI field in the Vietnamese educational world. The study will hold significance for students looking to choose a school for studying Data Science and Artificial Intelligence, offering insights into the trends and elements defining a good school. As humans continue to refine methodologies and expand their understanding of the Data and AI world, the researchers envision a future where Data Science and Artificial Intelligence education in Vietnam further blooms, producing skilled professionals and driving technological innovation forward.

Acknowledgments: We would like to thank William Nguyen, Toan V Luu, the VietSearch team, and other ISODS institute members for facilitating their support in this study.

REFERENCES

[1] Giap, Duong Ngo, Duc-The Thuy, Tran Quang Hung, Luu Huyen, Nguyen Anh, Ngoc. (2017). A Ranking of Vietnamese Universities 2017. *ResearchGate* 10.13140/RG.2.2.24358.19527.

[2] News, U. (2024). Best global universities - US news. *US News* https://www.usnews.com/education/best-global-universities

[3] VietNamNet News. (2020). Vietnamese universities to raise admission criteria. *VietNamNet* https://vietnamnet.vn/en/vietnamese-universities-to-raise-admission-criteria-621180.html

[4] Jaadi, Z. (2023). A Step-by-Step Explanation of Principal Component Analysis (PCA). *Build-In*. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[5] OriginLab Corporation. Help Online - Tutorials - Principal Component Analysis. https://www.originlab.com/doc/Tutorials/Principal-Component-Analysis

[6] Analytics Vidhya (2024). What Is Principal Component Analysis (PCA) & How It Works. https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

[7] Santos, G. (2023). Creating scores and rankings with PCA. *Medium*. https://towardsdatascience.com/creating-scores-and-rankings-with-pca-c2c3081fdb26

[8] Loukas, S. (2020). PCA Clearly Explained - When, Why, How To Use It and Feature Importance: A Guide in Python. *Towards Data Science*. https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e

[9] Ha, K. (2023). HUST is Ranked 360th in the World and 1st in Vietnam in the Field of Engineering and Technology. https://hust.edu.vn/en/news/campus-life/hust-is-ranked-360th-in-the-world-and-1st-in-vietnam-in-the-field-of-engineering-and-technology-623357.html

[10] Ananda, B.W.M. & Stephen, M.S. (2013) An Introductory Application of Principal Components to Cricket Data. *Journal of Statistics Education*, 21:3, DOI: 10.1080/10691898.2013.11889689