**Global university ranking 2023**

Our Goal: To accurately predict the likelihood of 'Number of Studnet' based on key features such as Rank, University name, Location, Number of Students, Number of Students per Staff, International Student Proportion, and Female to Male Ratio.

To know if there is a significant association between features or not.

**Related Work**

Studies Related to the Current Work:

|  |  |  |
| --- | --- | --- |
| Reference | Year | Methods |
| Current Work | 2023 | LDA, NB, BBN , NN, KNN, |

As of 2023, the dataset used in this work is new, and there are no other studies available at the moment.

**Methodology**

**Results and discussion**

* **Dataset Description:**

The "Global University Rankings Dataset 2023" is a comprehensive collection of key metrics and characteristics for top universities worldwide. The dataset provides insights into the performance and demographics of renowned academic institutions on a global scale.

* + Included features: [Rank, University name, Location, Number of Students, Number of Students per Staff, International Student Proportion, Female to Male Ratio]
    - Rank: Represents the ranking position of the university in global rankings.
    - University Name: Unique identifier for each institution.
    - Location: Geographical location of the university, indicating the country or region.
    - Number of Students: Total number of students enrolled in the university.
    - Number of Students per Staff: Ratio of total students to academic staff members, indicating the student-to-faculty ratio.
    - International Student Proportion: Proportion of international students, reflecting the global appeal and diversity of the university.
    - Female to Male Ratio: Gender distribution among the university's student body, presenting the ratio of female students to male students.
* **Project different phases:**

**Preprocessing phase:**

* Data cleaning:

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screen shot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Data analysis phase:**

* Get min, max, var, std, skewness, kurtosis:
  + For n\_student feature:

A screen shot of a computer

Description automatically generated

* + For female\_prop feature:

A screen shot of a computer program

Description automatically generated

* + For male\_prop feature:

A screen shot of a computer program

Description automatically generated

* + For all features from describe function:

A screenshot of a computer

Description automatically generated

🡺 Here are some interesting insights from the data:

* + - * Number of Students:
* Average number of students per university: 19,416.
* Median number of students: 13,990. => Indicates a few universities with very high student numbers, impacting the average positively.
  + - * Student-per-Staff:
* Average student-to-faculty ratio: 16 students per academic staff member. => Lower ratio suggests potentially better student-to-faculty interaction and improved educational experiences.
  + - * International Student Proportion:
* Mean proportion of international students: 10%.
* Median proportion: 5%. => On average, 5% of students at each university are international, but some universities have a higher proportion.
  + - * Gender Imbalance:
* Slight gender imbalance with a slightly higher proportion of female students, on average. => Indicates that, on average, there are slightly more female students than male students at universities, but the difference is not significant.
* For Categorical Summary Statistics:

A screenshot of a computer

Description automatically generated

* + - * Location of Universities (2023):
* The majority of universities in the dataset are located in the **United States**, totaling around 174 universities. The United States stands out as a prominent study destination, offering a more extensive range of choices compared to other countries.

**Data visualization phase:**

* The distribution of all numeric variables using a box plot:

A screen shot of a computer program

Description automatically generated

A computer screen shot of a code

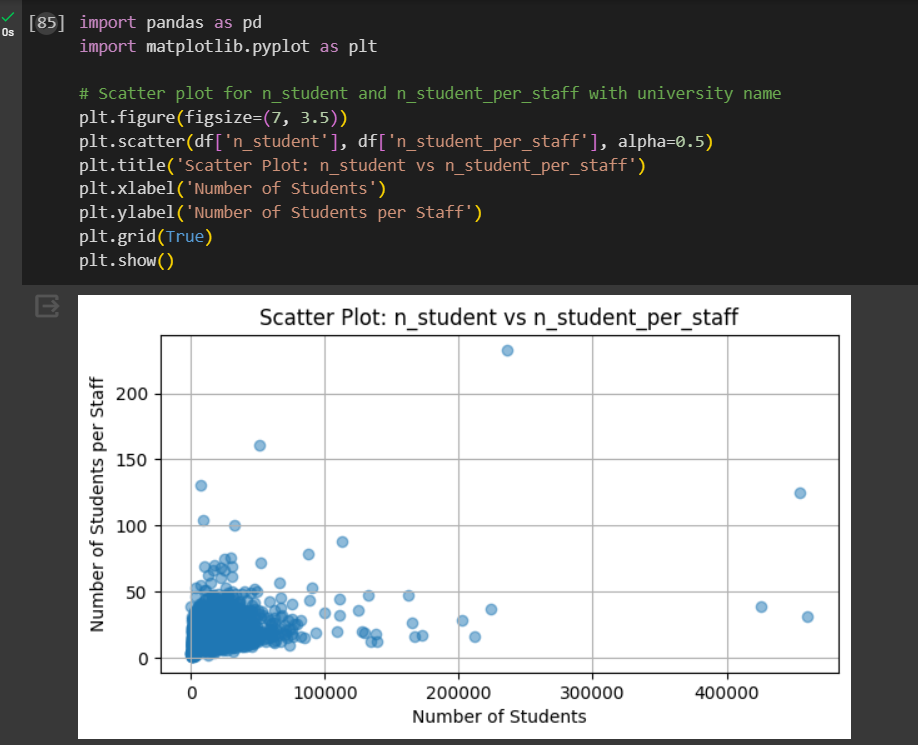
Description automatically generated

A group of graphs showing different types of students

Description automatically generated with medium confidence

🡺 Insights from data visualization:

* Overworked Staff: A few universities have a **high student-per-staff ratio**, with one university having more than 200 students per staff member.
* This suggests that these universities may have overworked staff and that the quality of education may be compromised.
* Gender Imbalance: A subset of universities has a remarkably high proportion of either male or female students, with some institutions having over 80% representation of one gender.
* This may indicate that these universities are not as inclusive as others.



A screen shot of a graph

Description automatically generated

A screenshot of a computer program

Description automatically generated

A group of graphs showing different sizes and numbers

Description automatically generated with medium confidence

A screen shot of a computer

Description automatically generated

* Some insights:
* n\_student (Number of Students)=> Skewness: 8.561805.
* n\_student\_per\_staff (Number of Students per Staff)=> Skewness: 5.403904.
  + - * This suggests that the distribution of the number of students and students per staff is skewed to the right. This means that there might be a few universities with an exceptionally large number of students ans also few universities have higher student-to-staff ratio.
* international\_student\_prop (Proportion of International Students)=> Skewness: 2.350499.
  + - * This implies that there might be a few universities with a notably higher proportion of international students.
* female\_prop (Proportion of Female Students) => Skewness: -0.287476 .
  + - * A small negative skewness suggests that most universities might have a slightly higher proportion of female students.
* male\_prop (Proportion of Male Students)=> Skewness: 0.287476.
  + - * This suggests a slight rightward skew in the distribution of the proportion of male students. Most universities might have a slightly higher proportion of male students.

**Data Analysis phase:**

* **Cov:**

A screenshot of a computer program

Description automatically generated

A screenshot of a graph

Description automatically generated

* **Corr:**

A screenshot of a computer program

Description automatically generated

**A graph of a graph

Description automatically generated with medium confidence**

* **Z-test:**
  + **I made z-test for n\_students and female\_pro and male\_prop:**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

* + - * **The conclusion for z-test:**

As: For Z-scores with absolute values less than the critical value we accept the null hypothesis. & For Z-scores with absolute values greater than the critical value, we reject the null hypothesis.

=> In previous outputs, most of the Z-scores have absolute values less than the critical value of 1.96 for a two-tailed test with a significance level of 0.05. Therefore, the null hypothesis is generally accepted for these data points.

=> However, there are a few cases where the Z-scores exceed the critical value (e.g., Z-scores with absolute values greater than 2), leading to the rejection of the null hypothesis for those specific data points.

* **ANOVA:**

**I performed Analysis of Variance (ANOVA) test to determine if there is a significant difference between the means of the specified columns in the dataset.**

A screen shot of a computer program

Description automatically generated

* In this case, the p-value is smaller than the chosen alpha (p-value << alpha). Therefore, **the null hypothesis is rejected.**
* There is a significant difference between the means of 'female\_prop' and 'male\_prop'. In other words, there is evidence to suggest that the proportions of females and males are **different in a statistically significant way.**

A screen shot of a computer program

Description automatically generated

* + - * This result indicates that the average values of 'n\_student' and 'international\_student\_prop' are not equal, and there is a **statistical significance in their differences.**
* **Chi-square:**

A screen shot of a computer

Description automatically generated

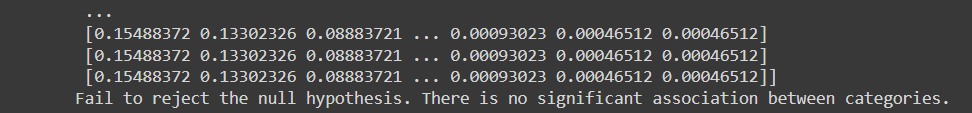
A black screen with white text

Description automatically generated

* A low p-value indicates that there is enough evidence to suggest a significant association between 'University name' and 'n\_student'.
* In this case, the null hypothesis is that there is no association.
* A higher chi-square statistic suggests a larger difference between the observed and expected values.

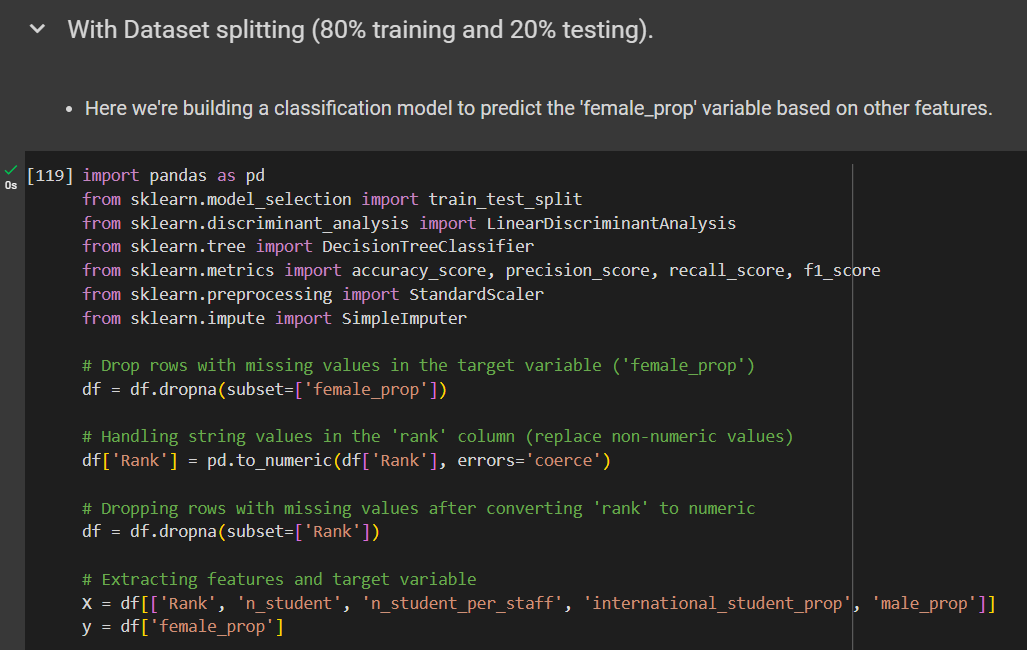
A computer screen with text and images

Description automatically generated



* The p-value is 0.35705597411449547, which is higher than the significance level (0.05). Therefore, you fail to reject the null hypothesis. The interpretation is that there is no significant association between 'University name' and 'international\_student\_prop' in the dataset.
* Also I made a chi-square for location with n\_student and for location with international student prop and the null hypothesis is rejected:
  + Reject the null hypothesis. There is a significant association between categories.
* Also for rank with international student prop and rank with n\_stundent and the null hypothesis is rejected.
  + - * **SO: There is a significant association between rank and international students and also between location and international students!**
* **Feature reduction phase:**

**Linear Discriminate Analysis (LDA)**



A screenshot of a computer program

Description automatically generated

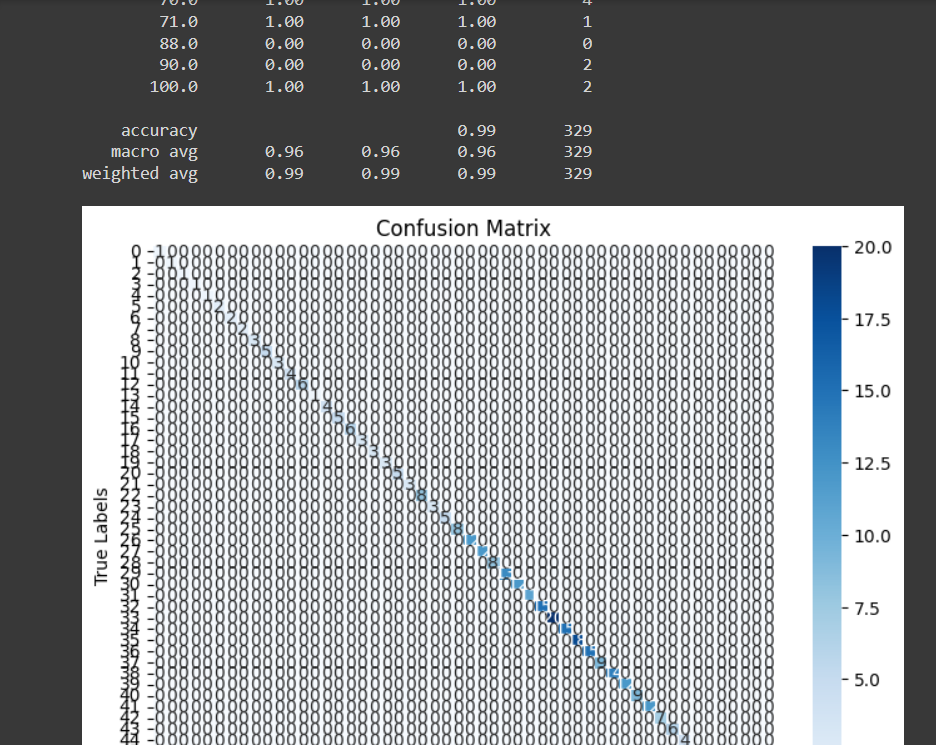
A screen shot of a computer program

Description automatically generated

* + - * **The model is correctly predicting the target variable (female proportion) for approximately 99% of the instances in the test set.**

A screen shot of a computer program

Description automatically generated



* + - * **Confusion matrix provide additional insights into the model's performance for each class.**
* **The classification report provides precision, recall, and F1-score for each class in our dataset.**

**Principle Component Analysis (PCA)**

**Applyiing PCA as SVD to reduce the dimensionality of the dataset while preserving as much of the original variability as possible**

A screen shot of a computer screen

Description automatically generated

A screen shot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

* **the model is now performing better in predicting the target variable.**

**Singular Value Decomposition (SVD)**

* Applying SVD to reduce the dimensionality of the dataset.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer program

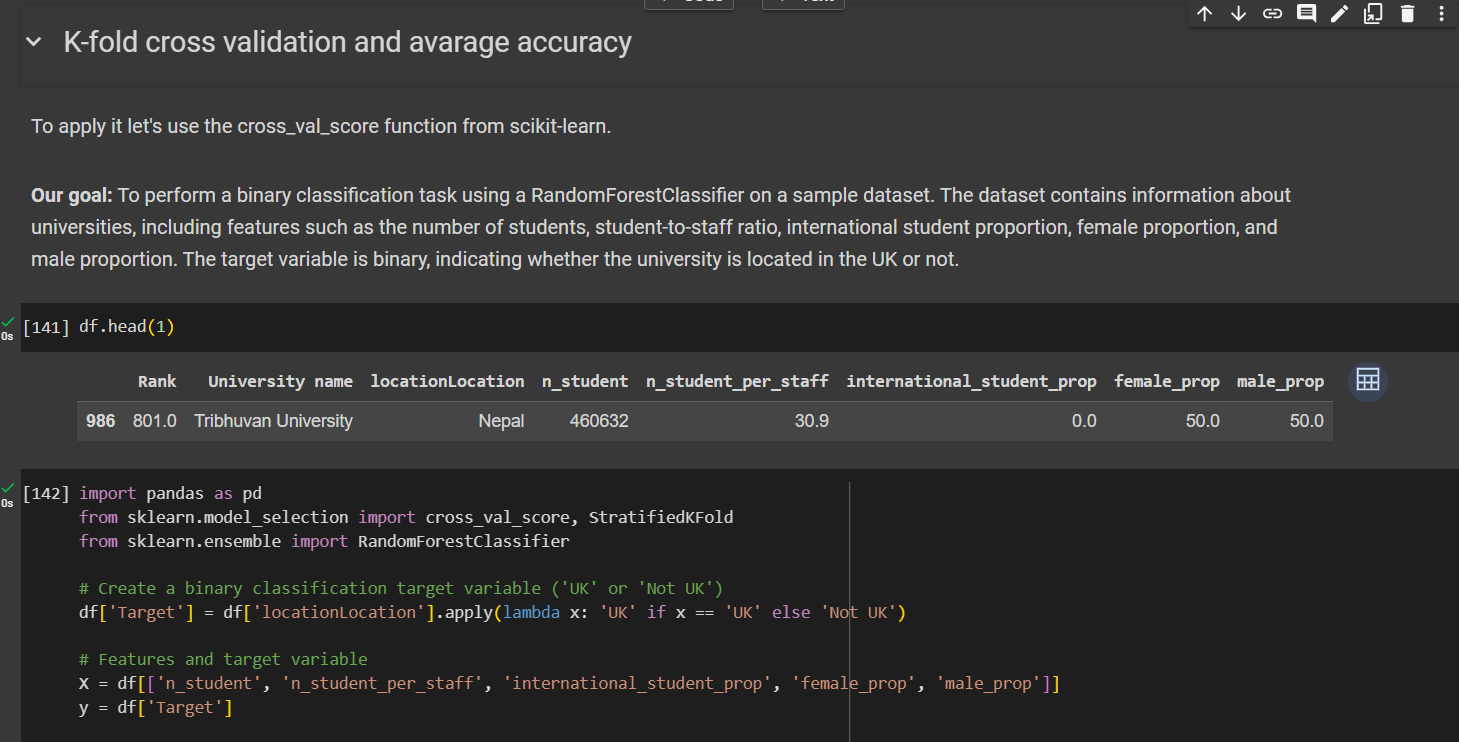
Description automatically generated

A screen shot of a computer program

Description automatically generated

* The model's performance, after applying Truncated Singular Value Decomposition (SVD), is quite poor.
* The accuracy, precision, recall, and F1 score are all extremely low (around 0.04), indicating that the model's predictions are not accurate, and it struggles to correctly classify instances.

**K-fold validation:**



A screenshot of a computer program

Description automatically generated

* + - * **The augmentation of the dataset has significantly enhanced the model's performance, resulting in perfect accuracy across all folds. However, it's essential to maintain a balance and avoid overfitting to the augmented data.**

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Results of confusion matrix:**

A screen shot of a graph

Description automatically generated

**Conclusion:**

There is a significant association between rank and international students and also between location and international students!

**References**

* Saritas, M. M., & Yasar, A. (2023, March 6). *Performance analysis of ann and naive Bayes classification algorithm for Data Classification*. International Journal of Intelligent Systems and Applications in Engineering. <https://ijisae.org/index.php/IJISAE/article/view/934>
* Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (1970, January 1). *KNN model-based approach in classification*. SpringerLink. <https://link.springer.com/chapter/10.1007/978-3-540-39964-3_62>
* Song, Y. (2015). Decision tree methods: applications for classification and prediction. *PubMed Central (PMC)*. <https://doi.org/10.11919/j.issn.1002-0829.215044>
* *Comparative analysis of PCA and LDA*. (2011, June 1). IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/abstract/document/5994243/>