DATA INNOVATORS

ANALYSING ENROLLMENT PATTERNS IN EDUCATION INSTITUTIONS

Team Members

Meghana Kota

Laasya Reddy Gaddam

Harshad Gupta Pasumarthy

Chandra Challa

ABSTRACT

Our research investigates the change in the enrollment patterns across United States educational institutions in the year 2022, mainly focusing on the factors influencing the enrollment rate. The research investigates the impact of institutional types, geographic location and various other economic factors which affect the enrollment rate by addressing the research question: Examining the changes in enrollment patterns and degree offerings across various US higher education institutions and the factors influencing enrollment.

Using the Integrated Postsecondary Education Data System (IPEDS) dataset, data was cleaned and aggregated for our analysis. Descriptive statistics, exploratory data analysis, and predictive models, including linear regression and decision trees, were used to explore enrollment trends and key influencing factors. Our analysis revealed a strong correlation between admissions and enrollment. Compared to men, women showed higher graduation rate.

According to the regression and random forest models, hospital facilities and institutional type (public vs. private) have very little impact on enrollment, but admissions, applications, graduation rate, and geographical characteristics all have a substantial impact. Further research expanding temporal and demographic studies may be possible, however the study offers valuable insights for strategic planning, policy-making, and well-informed decision-making in higher education.

INTRODUCTION

Our research project's objectives are to examine the reasons influencing these enrollment trends and the ways in which degree offers, and enrollment patterns have evolved in the year 2022, across various kinds of US higher education institutions. We are working on Integrated Postsecondary Education Data System (IPEDS) dataset, which provides comprehensive information on institutional characteristics, admissions and test scores, financial aid, enrollment, graduation rates, and more. This dataset combines data from various sources to focus on enrollment trends and factors influencing enrollment at institutions.

Through our research, we have investigated the patterns of enrollment in different kinds of higher education institutions (public, private for example), determine the economic, demographic, and policy-related elements that have impacted the enrollment. Our research is about changing pattern in educational institutions in the United States. So, our research can help in strategic planning, improving facilities, policies. The knowledge acquired can assist stakeholders in higher education, such as administrators, legislators, and potential students. As students, we have done lot of research while joining masters, by checking the universities, locations, tuition fees, weather, student-faculty ratio etc. We are trying to explore the same features with the educational institutions that we have in the dataset. The knowledge acquired can also assist stakeholders in higher education, such as administrators, legislators, and potential students, in more effectively navigating the industry's shifting dynamics.

LITERATURE REVIEW

In recent years, there has been a lot of interest in using learning analytics and educational data mining techniques to extract knowledge from huge educational datasets. Numerous studies have examined the state of these topics and how they are used in higher education.

Romero and Ventura (2024) offer a current overview of learning analytics and educational data mining, covering current approaches and trends. He provided a broader overview of learning analytics and educational data mining, addressing important techniques and resources.

Furthermore, a large-scale hierarchical dataset that may be used for tasks like knowledge tracing in interactive educational systems is provided by the EdNet dataset, which was first presented by *Lee et al. (2024).* Additionally, *Gašević, D., Dawson, S. & Siemens, G (2015)* delve into the potential of learning analytics to transform higher education. They discuss how learning analytics can be used to enhance student engagement, personalize learning experiences, and improve instructional practices. Their work emphasizes the importance of data-driven decision-making in higher education.

Examining the most recent developments and projected trends in the application of learning analytics in higher education, also examines the impact of data-driven decision-making on the education sector. Data science approaches are used to investigate the integration of adaptive learning systems. While the use of data science technology to enhance student learning outcomes, talk about the analytics and research methods involved in exploiting big data in education.

METHODOLOGY

The IPEDS dataset, offers extensive data on enrollment, financial aid, test and admission scores, graduation rates, and other topics, is what we are focusing on. To concentrate on enrollment trends and the variables affecting enrollment at universities, the dataset is integrated from multiple sources. It contains information on 6,256 universities across the United States for the year 2022, with 30 columns detailing key aspects such as institution ID, name, location, state, region, enrollment numbers, admissions, applications, instructional staff, and graduate counts.

To aggregate the data, we used Excel’s VLOOKUP function, referencing UNITID, the unique identifier for each institution. In the final dataset, we encountered numerous missing and negative values. Negative values indicated unavailable or irrelevant data, so we replaced these with 0. While there were no outliers, several rows had substantial missing data. We removed rows with more than 15 missing values, resulting in a dataset of 5,721 rows. For the remaining missing values, we filled them using the minimum value of each column within each region, as region is a significant factor influencing students' choices. The final dataset has a shape of (5721,30).

Target variable in our dataset is 'ENRLT', which is the total number of enrollments in each institute. Factors affecting our target variable are 'APPLCN', 'ADMSSN', 'ENRLT', 'SAINSTT', 'GRTOTLT', 'NPIST2'. We created a data dictionary for our dataset to provide a clear understanding of each column, including its description and data type.

Column Name	**Description**	**Datatype**
UNITID	Institute ID	INT
INSTNM	Institution Name	STRING
ADDR	Address of the institute	STRING
CITY	City	STRING
STABBR	State	STRING
APPLCN	Total number of applications	INT
APPLCNM	Total number of applications - Men	INT
APPLCNW	Total number of applications - Women	INT
ADMSSN	Total number of admissions	INT
ADMSSNM	Total number of admissions - Men	INT
ADMSSNW	Total number of admissions - Women	INT
ENRLT	Total Enrollments	INT
ENRLM	Total Enrollments - Men	INT

ENRLW	Total Enrollments - Women	INT
SAINSTT	Total Instructional Staff	INT
SAINSTM	Total Instructional Staff - Men	INT
SAINSTW	Total Instructional Staff - Women	INT
GRTOTLT	Number of students Graduating	INT
GRTOTLM	Number of students Graduating - Men	INT
GRTOTLW	Number of students Graduating - Women	INT
NPIST2	Average net price-students awarded grant or scholarship aid, 2021-22	INT
NPIS412	Average net price (income 0-30,000)-students awarded Title IV federal financial aid, 2021-22	INT
NPIS422	Average net price (income 30,001-48,000)-students awarded Title IV federal financial aid, 2021-22	INT
NPIS432	Average net price (income 48,001-75,000)-students awarded Title IV federal financial aid, 2021-22	INT
NPIS442	Average net price (income 75,001-110,000)-students awarded Title IV federal financial aid, 2021-22	INT
NPIS452	Average net price (income over 110,000)-students awarded Title IV federal financial aid, 2021-22	INT
OBEREG	Bureau of Economic Analysis (BEA) Regions 0 - US Service schools 1 - New England CT ME MA NH RI VT 2 - Mid East DE DC MD NJ NY PA 3 - Great Lakes IL IN MI OH WI	INT

	4 - Plains IA KS MN MO NE ND SD 5 - Southeast AL AR FL GA KY LA MS NC SC TN VA WV 6 - Southwest AZ NM OK TX 7 - Rocky Mountains CO ID MT UT WY 8 - Far West AK CA HI NV OR WA 9 - Outlying areas AS FM GU MH MP PR PW VI 3 - Not available	
CONTROL	A categorization of whether an organization is run by officials who are elected or appointed by the public or by officials who are appointed or elected privately and who obtain most of their funding from private sources.	INT
HOSPITAL	A code to indicate whether the institution has hospital.	INT
OPENPUBL	A code to indicate whether the institution is open for admission to the public.	INT

Table 1: Data Dictionary

After cleaning the data, we utilized descriptive statistics and various visualizations, including bar graphs, pie charts, and correlation matrices, to gain insights into the data and identify underlying patterns.

Descriptive statistics of predictive variables and target variables (ENRLT):

	APPLCN	ADMSSN	ENRLT	SAINSTT	GRTOTLT	NPIST2	OBEREG	CONTROL	HOSPITAL
count	5721	5721	5721	5721	5721	5721	5721	5721	5721
mean	6936.13	4097.30	895.51	344.66	3461.68	3409.72	4.63	2.05	0.82
std	8013.20	3903.81	895.68	515.38	5160.19	5754.01	2.18	0.85	0.98

min	0	0	0	2	2	0	0	1	0
25%	3022.58	2203.02	564.66	98	819	0	3	1	0
50%	6573.61	3935.09	866.88	324	3008	0	5	2	0
75%	7645.65	5316.58	1005.53	368.29	3608.75	6503	6	3	2
max	149801	61739	15151	8414	72219	45657	9	3	2

Table – 2: Descriptive statistics of predictive and target variables

As we are working on different educational institutions’ data, it is important to have diverse data which involves public and private institutes. So, we have created a pie-chart to understand the proportions of public, private for profit and private not-for-profit institutes.

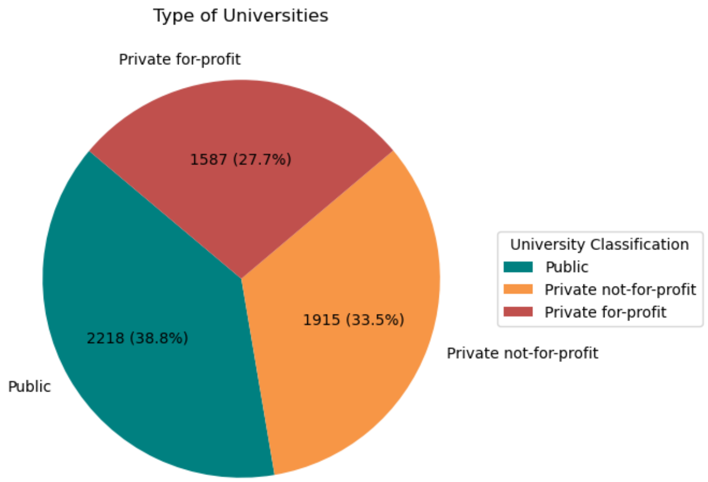


Figure – 1: Classification based on type of university.

Explanation:

By looking at the above pie-chart, we can say that the dataset is diverse. It has 1587 private non-profit institutions, 1915 private not-for-profit institutions and 2218 public institutions.

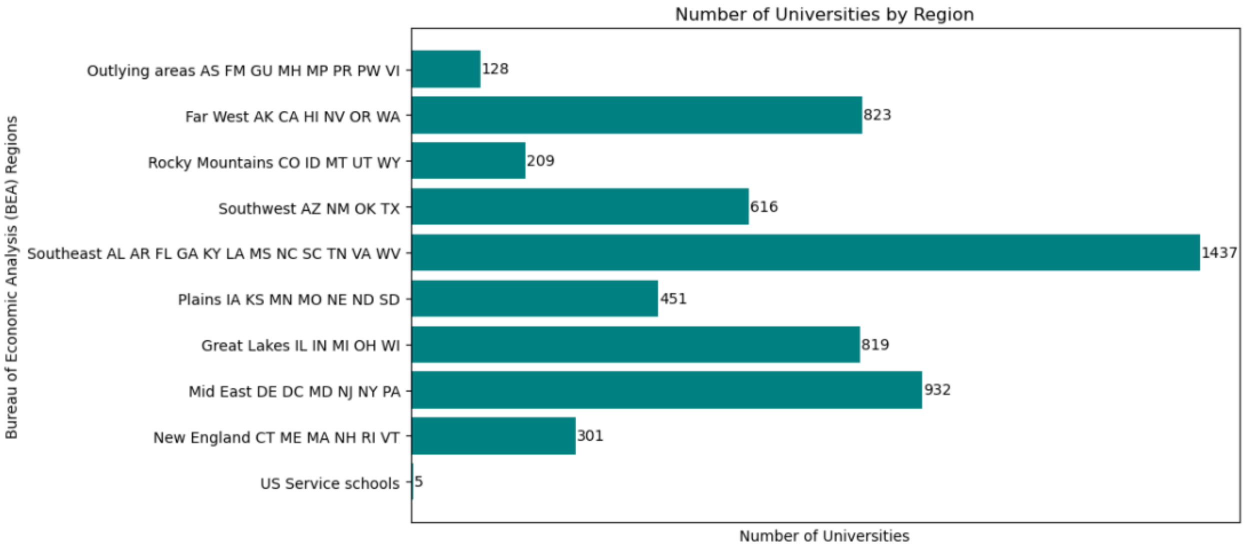


Figure – 2: Number of institutions in each region

Explanation:

The above bar graph shows the total number of educational institutions in each region. The southeast states – AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV has highest number of educational institutions. There are very few US service schools.

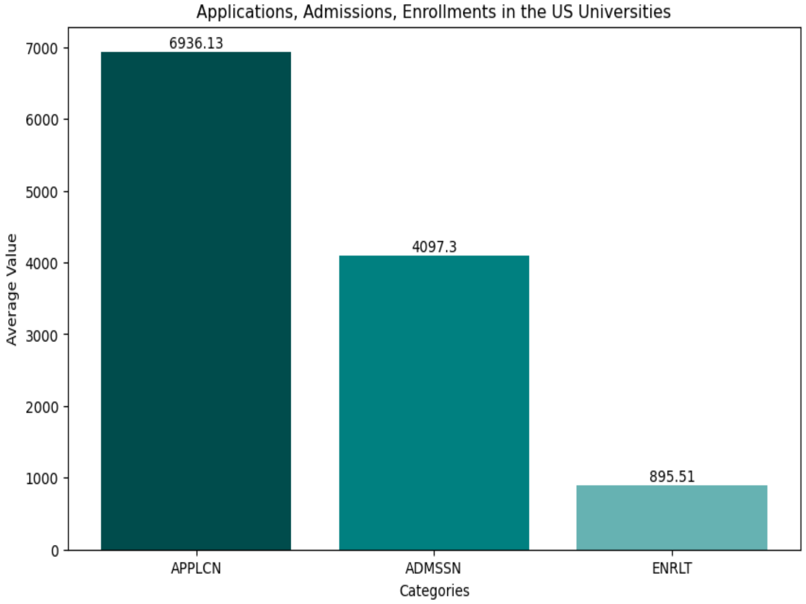


Figure – 3: Average number of applications, admissions and enrollment in the United States

Explanation:

Number of applications, Admissions and Enrollment are interdependent on each other. The above bar graph shows the average number of applications, admissions and enrollment across different institutes in the United States.

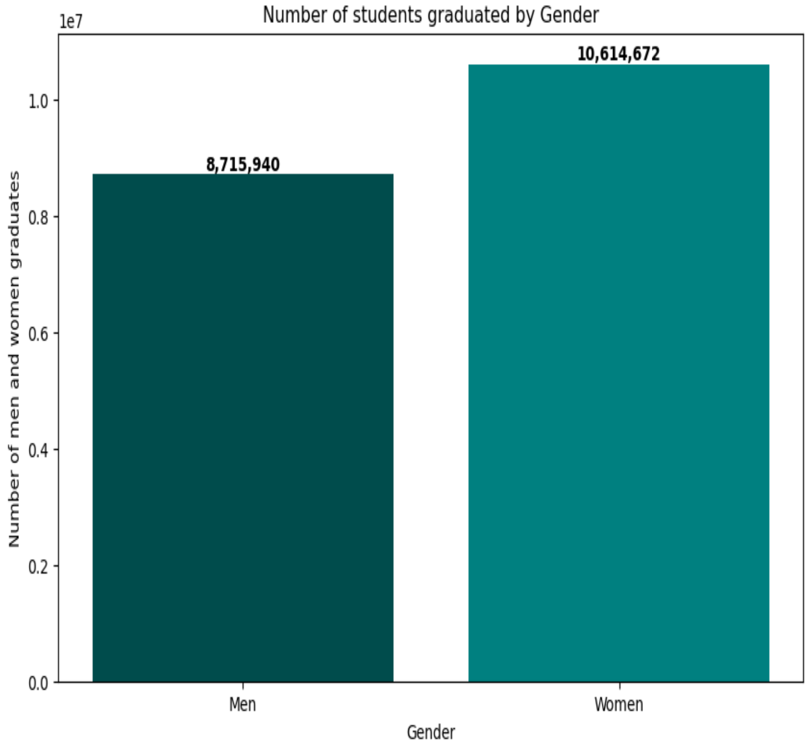


Figure – 4: Number of students graduated by gender

Explanation:

The above bar graph shows the total number of men and women students graduated in the year 2022. From Figure-4, we can say that the graduation count is high in women when compared to men. Total women graduates in 2022 are 10, 614,672, whereas men are 8,715,940.

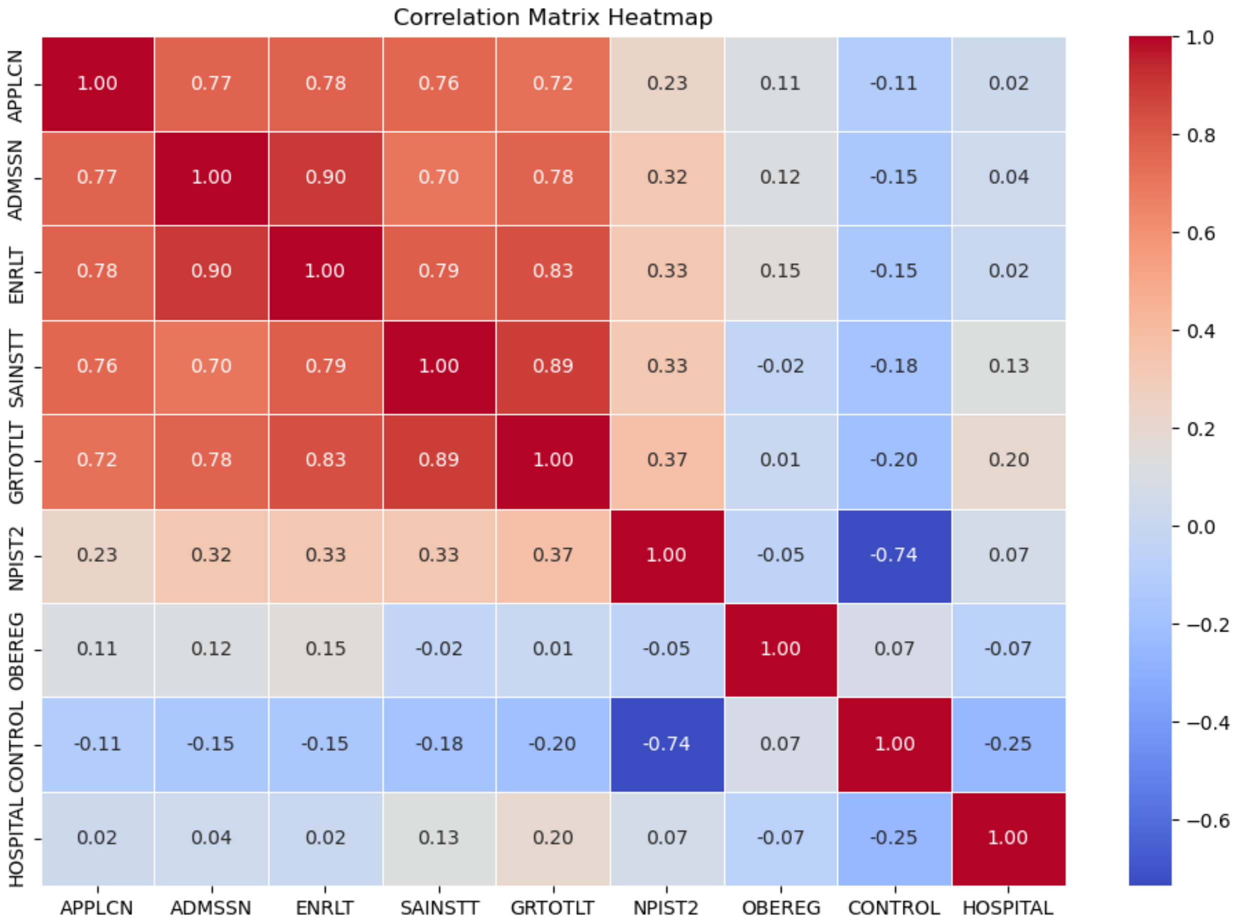


Figure – 5: Correlation Matrix

Explanation:

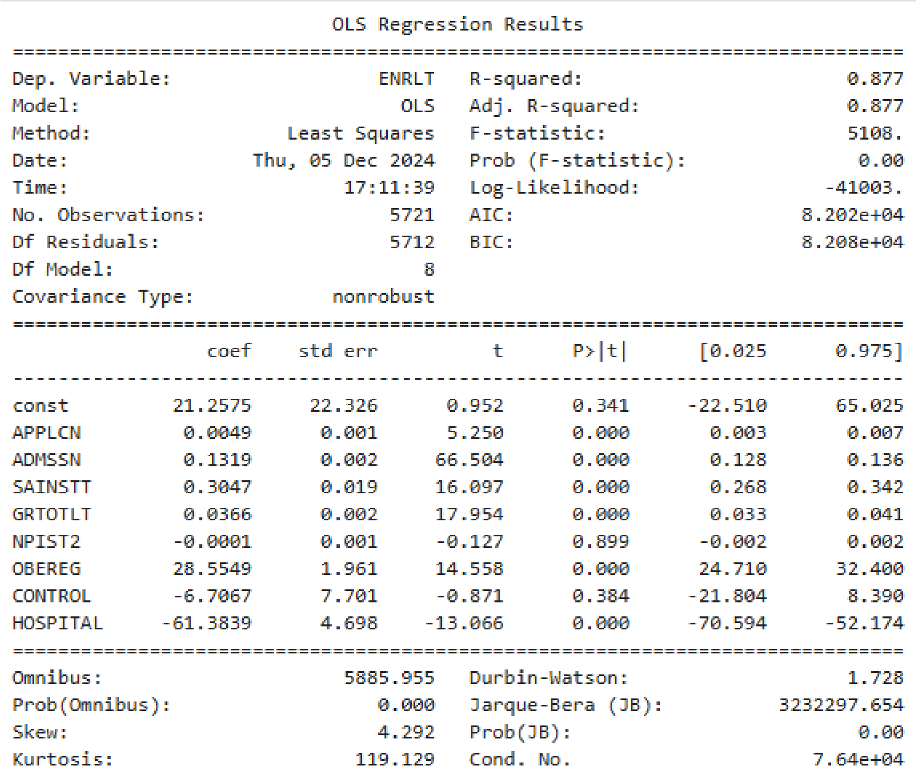
To understand the correlation between the variables in the data, we have created a correlation matrix. The highest correlation is observed in between enrollment and admissions with the value of 0.90, followed by enrollment and graduation rate with the value of 0.78.

For our analysis, we used linear regression, a regression-based statistical model that can predict the factors influencing an institution's enrollment. In our dataset, geographic information about institutions is represented as categorical variables, while the remaining data is numeric. As most of our data is numeric and to find factors affecting enrollment, it is best to use linear regression.

We have built the model by dividing the data into training, testing and validation data. Additionally, we tried other models such as random forest and checked their performance. We have compared the results of different models.

RESULTS AND DISCUSSION

The analysis of IPEDS dataset revealed the factors affecting the enrollment in the educational institutions. The results of OLS regression model are shown below.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  ENRLT   R-squared:                       0.877
Model:                            OLS   Adj. R-squared:                  0.877
Method:                 Least Squares   F-statistic:                     5108.
Date:                Thu, 05 Dec 2024   Prob (F-statistic):               0.00
Time:                        17:11:39   Log-Likelihood:                -41003.
No. Observations:                5721   AIC:                         8.202e+04
Df Residuals:                    5712   BIC:                         8.208e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         21.2575     22.326      0.952      0.341     -22.510      65.025
APPLCN         0.0049      0.001      5.250      0.000       0.003       0.007
ADMSSN         0.1319      0.002     66.504      0.000       0.128       0.136
SAINSTT        0.3047      0.019     16.097      0.000       0.268       0.342
GRTOTLT        0.0366      0.002     17.954      0.000       0.033       0.041
NPIST2        -0.0001      0.001     -0.127      0.899      -0.002       0.002
OBEREG        28.5549      1.961     14.558      0.000      24.710      32.400
CONTROL       -6.7067      7.701     -0.871      0.384     -21.804       8.390
HOSPITAL     -61.3839      4.698    -13.066      0.000     -70.594     -52.174
==============================================================================
Omnibus:                     5885.955   Durbin-Watson:                   1.728
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          3232297.654
Skew:                           4.292   Prob(JB):                         0.00
Kurtosis:                     119.129   Cond. No.                     7.64e+04
```

Figure – 6: Regression Summary

Regression equation of the model is:

ENRLT = 21.2575 + 0.0049 * APPLCN + 0.1319 * ADMSSN + 0.3047 * SAINSTT + 0.0366 * GRTOTLT – 0.0001 * NPIST2 + 28.5549 * OBEREG – 6.7067 * CONTROL – 61.3839 * HOSPITAL

The R^2 value and adjusted R^2 of the model are 0.877, which indicates that the model is good fit as variance in the dependent variable (ENRLT) is explained by the independent variables in the model. The p-value of F-statistics is 0.000, which indicates the model is statistically significant.

Variable Name	**Coefficient**	**p-value**
APPCLN	0.005	0.000
ADMSSN	0.132	0.000
SAINSTT	0.305	0.000
GRTOTLT	0.037	0.000
NPIST2	0.000	0.899
OBEREG	28.555	0.000
CONTROL	-6.707	0.384
HOSPITAL	-61.384	0.000

Table – 3: Variables and Coefficients

The results says that APPCLN, ADMSSN, SAINSTT, GRTOTLT, OBEREG has high significance on enrollment. Hospital, OBEREG has no significance on enrollment. This regression model reveals that significant predictors like APPCLN, ADMSSN, and GRTOTLT are the crucial factors affecting enrollment.

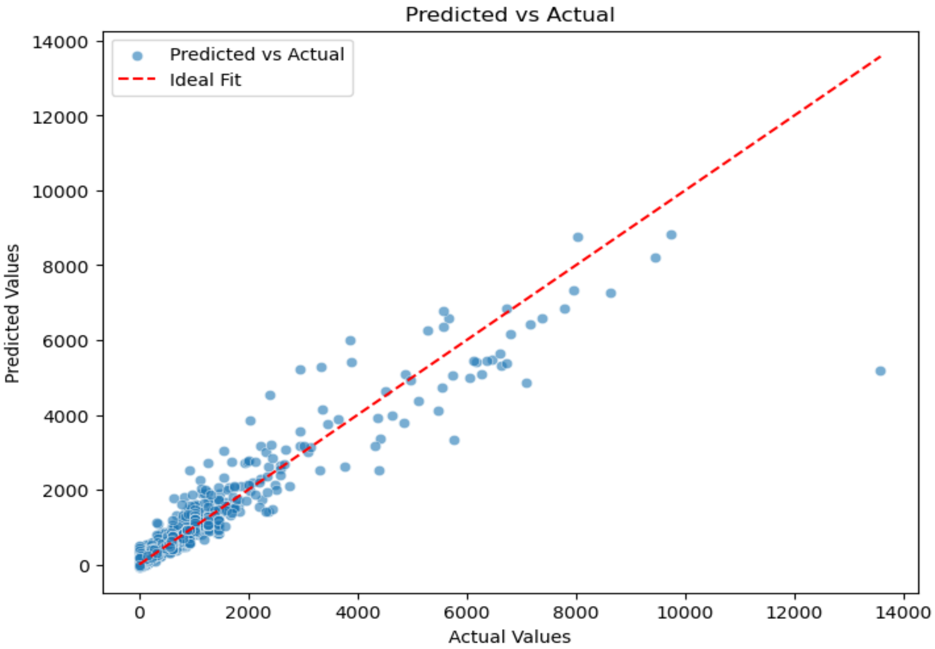


Figure – 7: Scatter plot for predicted Vs actual values

The blue dots follow an upward trend, indicating that the model's predictions tend to increase as the actual values increase. The model suggested that some of the underlying relationships between the independent and dependent variables are identified.

Random Forest Regression

R^2 value of the random forest regression is 0.88, which means that approximately 88.53% of the variability in the target variable can be explained by the model. The mean square error is low.

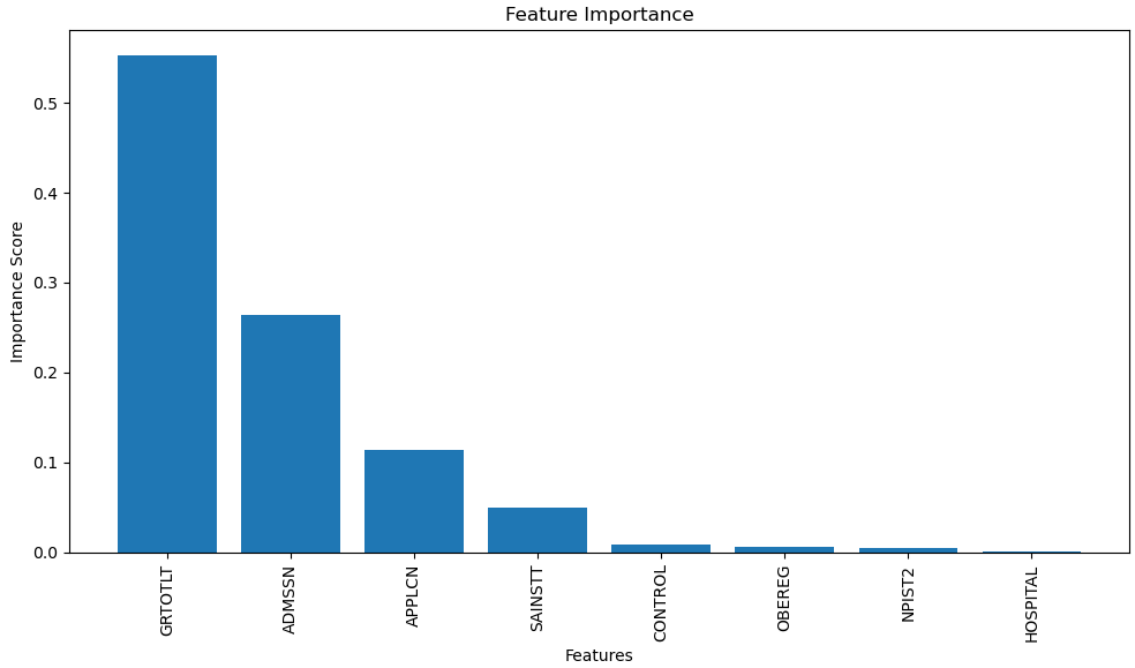


Figure – 8: Feature Importance

The factors affecting the enrollment are admissions, applications, graduation rate as per the random forest regression model. In Figure – 9, some dots are closer to the line, while others are farther away. This indicates that the model's predictions are not perfectly accurate, and there is some variability in the predictions.

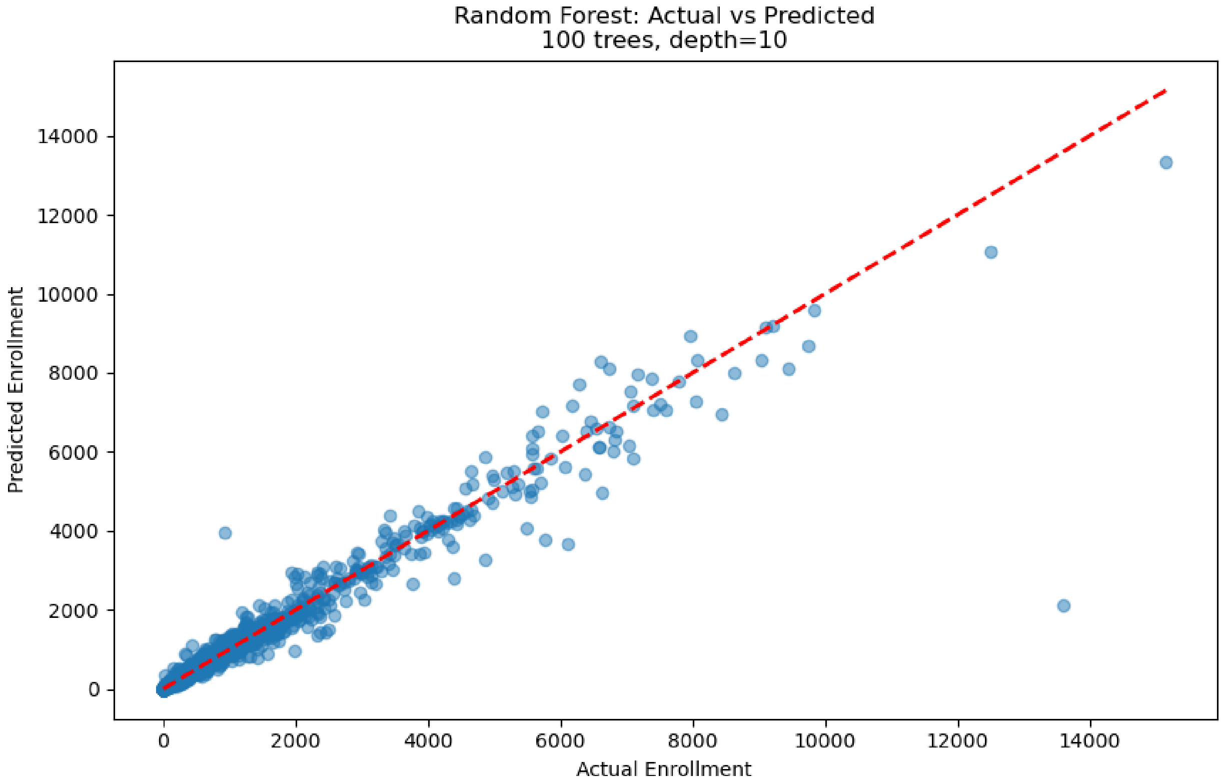


Figure – 9: Scatter plot for predicted Vs actual values

Considering the results of both the models, we can say that admissions, applications, graduation rate and regions highly affect the enrollment in any educational institution. Hospital facilities and the type of institution, like private and public institutions, do not affect the enrollment rate.

The study offers information for institutional strategic planning. Program creation and comprehension of changing dynamics in higher education. Policymakers can create more focused educational initiatives, and the administrators can use data-driven insights.

Prospective students are able to make better informed decisions about their education. Although the study only includes data from 2022. Future investigations could extend temporal analysis, include more detailed demographic information, examine new developments in online and hybrid education.

REFERENCES

Romero, C., & Ventura, S. (2024). Educational Data Mining and Learning Analytics: An Updated Survey. arXiv preprint arXiv:2402.07956.

Global Data Set on Education Quality (1965-2015)" - Altinok, N., Angrist, N., & Patrinos, H. A. (2018). World Bank

Lee, Y., et al. (2024). *EdNet: A Large-Scale Hierarchical Dataset in Education*. *arXiv preprint arXiv:1912.03072*.

Gašević, D., Dawson, S. & Siemens, G. Let's not forget: Learning analytics are about learning. *TECHTRENDS TECH TRENDS* **59**, 64–71 (2015). https://doi.org/10.1007/s11528-014-0822-x