Volume 13, No. 3, 2022, p. 146 - 152
https://publishoa.com
ISSN: 1309-3452

Recommendation system: National Institute rank prediction using Machine Learning

Gadi Himaja
Asst Professor, Gitam Deemed To be University, Visakhapatnam

Gadu Srinivasa Rao
Asst Professor, Gitam Deemed To be University, Visakhapatnam

Gali Akarsh Naidu

Student, Gitam Deemed University, Visakhapatnam

Tirumalesh Nagothi
Student, Gitam Deemed University, Visakhapatnam

Susmitha Dalli
Student, Gitam Deemed University, Visakhapatnam

ABSTRACT

Predicts future university or college rankings and to design a Recommendation system that notifies institutions of which factors could be addressed to improve their future rankings. The approach for constructing a system that predicts the ranking of university by evaluating the indicators of performance for national university is provided in this research. Here the datasets used are from the NIRF. Then based on the score of previous years, we predict the rank by giving the performance indicators to the model. Later, a recommendation system was built which tells a university or college which parameters must be improved so as to get better rankings in the future[1][2].

INTRODUCTION

The creation of national rank- ing systems has resulted from the recent interest in various educational institute rankings. The Ministry of Education's National Institutional Ranking Framework (NIRF), which provides access to statistical databases and reputation surveys to generate national league tables, is primarily responsible for the Indian national university ranking system. This national ranking system contains a more comprehensive collection of factors due to their access and in-depth understanding of local institutions. The Dataset used here includes the top 100 national universities or colleges for each year in this research. The collected Dataset is from the NIRF website[1] for the years 2016 to 2021. This Dataset has 600 samples and 12 parameters or columns such as Institute Id, Institute Name, City, State, Rank, Score, TLR, RPC, GO, OI, Perception and Year. To predict the ranks of the universities or colleges, we created Machine Learning models using Machine Learning algorithms like Ridge Regression, Decision Tree Regression, KNN Regression, Linear Regression, Lasso Regression and Random Forest Re- gression. Then a table is generated, which consists of the R Square scores, Mean Absolute Errors (MAE), Mean.

Square Errors and Root Mean Square Errors and select the model which has the highest R Square scores along with low Mean Square Errors, Mean Absolute Errors, and Root Mean Square Errors with both training and testing datasets. A Recommendation model was developed by finding out the Z Scores of parameters like Score, GO, TLR, RPC, OI, and Perception and comparing the Z scores with fixed Threshold values and then displaying the parameters that require improvement. And lastly, a web interface was developed that displays the results using the flask module so that the model can be used by users who have no programming background.

LITERATURE REVIEW

Given the stakes involved in selecting an accept- able educational college in terms of time commitment, future job opportunities, and financial resources, rank- ing and evaluating educational institutes has become increasingly significant. Rankings are important, espe- cially for students who want to acquire the greatest education available for their higher education. Furthermore, educational institutes are increasingly using rankings to establish strategies for supporting their institutions' growth and development. The existing research in this field has developed this prediction systems for ranking the university by analyzing the university performance that has a global influence indicates [4][5][30][32]. Comparing the ranking systems of national and international institutions in terms of metrics, coverage, and ranking outcomes has also been done [6][7][8][29].There are

Volume 13, No. 3, 2022, p. 146 - 152
https://publishoa.com
ISSN: 1309-3452

researches that examined the methodology and key features of all existing university ranking systems throughout the world [9]. Research has also been done on the impact of rankings on the institutions and how it affects them [10][11].Though many university ranking prediction models and also the comparison of various ranking systems have been made, there is no proper system to suggest to the universities which influential parameters can improve so as to increase their future rankings.

METHODOLOGY

- *Decision Tree Algorithm*
- *K-Nearest neighbor Algorithm*
- *Random Forest Algorithm*
- *Linear-Regression Algorithm*
- *Lasso-Regression Algorithm*
- *Ridge-Regression Algorithm*

DECISION TREE ALGORITHM

To forecast data and provide meaningful con- tinuous output, decision tree regression analyses an object's attributes and the model is trained as a tree shape. In the perspective as that might not be represented only by a different, known set of values or numbers, the output/result is not discrete. A model that states the prediction is likely profit that might get earned from the sale of a product is a continuous output example.This model is used for predicting the values which are continuous in this case [12]..

RANDOM FOREST ALGORITHM

It's an ensemble method that combines numerous decision trees with a technique called Bootstrap and Aggregation, sometimes known as bagging, to solve regression and classification problems. The main idea is to acquire the final conclusion by combining sev- eral decision trees rather than relying on individual decision trees. The Random-Forest-Algorithm expands the decision tree and is useful because it solves the decision tree's challenge of putting data points into a slightly incorrect category unnecessarily [14].

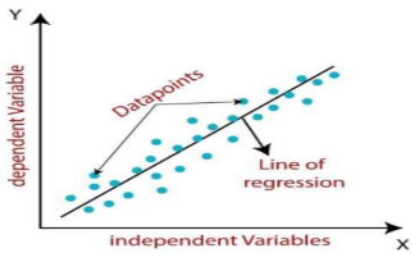
$$S(T,X) = \sum_{c \in X} P(c)S(c)$$

K-NEAREST NEIGHBOUR ALGORITHM:

The K-nearest neighbours algorithm is a straight- forward method of storing all relevant data and fore- casting numerical targets using a similarity score. KNN been utilised in statistical estimation and pattern identification as an approach which is non-parametric. The KNN regression method is a non-parametric approach for assessing the association among independent factors and variable results. Calculating the average of the numerical target of the K closest neighbour is a simple and direct application of KNN regression. Another option is to use the K nearest neighbours' inverse distance weighted average[13].

LINEAR-REGRESSION ALGORITHM

Linear-regression is a machine learning approach for supervised learning. It's good at regressing. Based on independent variables, regression models the desired predicted value. It's generally used for predicting and assessing relationships between variables. A linear relationship between either one or more independent variable and one dependent variables is established using the concerned procedure. Linear regression depicts how the value of thevariables which depend varies due to the change of the value of the variable which changes. The bonding between the variable is represented in the linear-regression model by a straight line in an inclined way. [15].



Mathematically, we can represent a linear regression as:

$$y = a_0 + a_1 x + \varepsilon$$

Volume 13, No. 3, 2022, p. 146 - 152
https://publishoa.com
ISSN: 1309-3452

Here,
Y is the Variable been Dependent (Target Variable)
X is the Variable been Independent (predictor Variable) a0 is the Line_intercept (Gives an additional degree of freedom)
a1 is the Linear-regression_coefficient (scale factor to each input value). = random stuff error

LASSO REGRESSION ALGORITHM

It's a form of normalisation. It is chosen over regression techniques for a more exact forecast. In shrinkage, values are compressed towards the mean, a centre point. The lasso approach promotes models to be basic and sparse. This regression type is best for models with a lot of multicollinearity or when you want to automate parts of the model selection process like variable selection and parameter removal. The L1 regularisation technique is used in Lasso Regression (which will be discussed later in this article). It is utilized when there are a large number of features since it performs feature selection automatically [16].

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

RIDGE REGRESSION ALGORITHM Ridge re-

gression is a method of determining the coefficients of multiple regression models in circumstances when linearly independent variables are highly linked. Ridge regression is a model tuning technique that may be applied to multicollinear data analysis. This approach is used to produce L2 regularisation. Least-squares are unbiased and variances are high when there is an issue with multicollinearity, resulting in predicted values that are distant from the actual values [17].

$$\widehat{\beta} = (X^T X)^{-1} X^T y$$

(where X^T is the transpose of X).

Algorithm Selection:
The Regression algorithms used in this research are Ridge Regression, Decision Tree, Random Forests, KNN, Linear Regression and Lasso Regression.

TRAIN MODELS:
The models are trained using regression algorithms like Linear Regression, Decision Tree, Random Forests, KNN, Ridge Regression and Lasso Regression. The data set collected from the NIRF website is taken to train the models.

MODEL EVALUATION:
Model assessment is important because it helps us better understand our model's performance and makes it simpler to demonstrate it to others. There are many evaluation metrics to choose from, but only a handful are suitable for regression.
So basically, there are four types of model evaluation metrics:

R SQUARE:

R Square is a model evaluation metric in which the sum of the squared predicted error is divided by the entire sum of the square. The value might vary from 0 to 1, with a higher number, the forecast and actual value are more closely aligned. Because determining the accuracy of regression models is difficult, R Square is commonly used as the model's accuracy [18][21].

MEAN ABSOLUTE ERROR (MAE):
MAE is calculated by adding the absolute values of the errors. The MAE represents the sum of error terms in a more direct manner [19][21].

MEAN SQUARE ERROR (MSE):
The sum of squares of prediction error, which is the difference between real and predicted output, is multiplied by the number of data points to get the MSE. It gives an absolute number that shows how far the predicted values differ from the actual value [20][21].

Volume 13, No. 3, 2022, p. 146 - 152
https://publishoa.com
ISSN: 1309-3452

ROOT MEAN SQUARE ERROR (RMSE):
It's the square root of MSE (RMSE). Because the MSE figure might often be too huge to compare easily, it is more widely employed than MSE. Since the MSE is based on the square of error, omitting the square root decreases the prediction error to the same level as the original and simplifies the explanation [21].

MODEL SELECTION:
Based on the above model evaluation metrics, the best model is chosen for further development. We select a model which has the highest R Square scores along with low Mean Absolute Errors, Mean Square Errors and Root Mean Square Errors with both training and testing data sets.

Regression Algorithms	R Square		MAE		MSE		RMSE	
	Train	Test	Train	Test	Train	Test	Train	Test
Linear	0.69	0.64	13.08	14.56	255.34	301.07	15.98	17.35
Decision Tree	1.0	0.86	0.0	7.05	0.0	130.23	0.0	11.41
Random Forest	0.98	0.87	2.64	6.92	14.52	99.29	3.81	9.96
KNN	0.74	0.68	10.47	11.81	214.67	256.94	14.65	16.03
Ridge	0.70	0.56	13.51	13.06	266.43	265.64	16.32	16.30
Lasso	0.67	0.71	13.66	13.27	268.20	263.70	16.38	16.24

RECOMMENDATION SYSTEM:

Z SCORE CALCULATION:
The Z-score is calculated for the Score, TLR, RPC, GO, OI and Perception parameters. The standard score, commonly known as the Z-score, is a measurement of how much a data point deviates from the mean. It expresses how far an element deviates from the mean in standard deviations.
As a result, the Z-Score is expressed as a standard deviation from the mean [22].
The formula for Z Score:

$$z = \frac{\text{data point} - \text{mean}}{\text{standard deviation}} \quad \text{(OR)} \quad z = \frac{x - \mu}{\sigma}$$

THRESHOLD VALUE COMPARISON:
Once the Z-Scores are calculated for the Score, TLR, RPC, GO, OI and Perception parameters, we find the maximum value of each column. We then multiply that maximum value with some percentage as a threshold (let's consider a 30% threshold).
Now we compare this threshold with the rest of the column Z-Scores values. If the Z-score is greater than the considered threshold value, then we get the output as 1; else, we get the output as 0.

IMPROVEMENT RECOMMENDATION:
Once we stored the outputs of the comparison in other columns, we just have to retrieve the rows which have the output as 0 and then display their respective column indices.

	Year	Score Parameter	TLR Parameter	RPC Parameter	GO Parameter	OI Parameter	Perception Parameter
45	2016	0	0	0	0	0	1
144	2017	0	0	0	0	1	0
244	2018	0	0	0	0	1	0
344	2019	0	0	0	0	0	0
444	2020	0	0	0	0	0	0
544	2021	0	0	0	1	0	0

Volume 13, No. 3, 2022, p. 146 - 152
https://publishoa.com
ISSN: 1309-3452

```
The college/university needs improvement in:
Score Parameter
TLR Parameter
RPC Parameter
GO Parameter
Perception Parameter
```

```
dff=pd.concat([df16,df17,df18,df19,df20,df21])
dff
```

	Year	Score_ZScore	TLR_ZScore	RPC_ZScore	GO_ZScore	OI_ZScore	Perception_ZScore	Score Parameter	TLR Parameter	RPC Parameter	GO Parameter	OI Parameter	Perc Para
0	2016	2.868440	1.835069	2.204328	0.578280	0.236926	1.251479	1	1	1	1	0	
1	2016	2.464522	1.208862	2.227255	0.394724	-0.165026	1.175402	1	1	1	1	0	
2	2016	2.356619	1.519441	1.511221	0.578280	0.702230	1.200761	1	1	1	1	1	
3	2016	2.261130	0.575081	1.796693	0.171703	1.813426	1.200761	1	0	1	0	1	
4	2016	2.152273	1.163412	1.647020	0.165164	0.934518	0.896450	1	1	1	1	1	
...	...	...	...	...	...	...	...	...	...	...	...	...	
595	2021	-1.098475	-0.873619	-0.737556	-0.862044	-0.272081	-0.822420	0	0	0	0	0	
596	2021	-1.103686	-0.806066	-0.723855	-0.888087	-0.263336	-0.962897	0	0	0	0	0	
597	2021	-1.104968	-0.982454	-1.101309	-0.877069	0.511218	0.027431	0	0	0	0	0	
598	2021	-1.107584	-2.070802	-0.897169	0.438119	0.175162	-0.437331	0	0	0	0	0	
599	2021	-1.108896	-1.173853	-0.735501	-0.940174	0.479986	-0.690329	0	0	0	0	0	

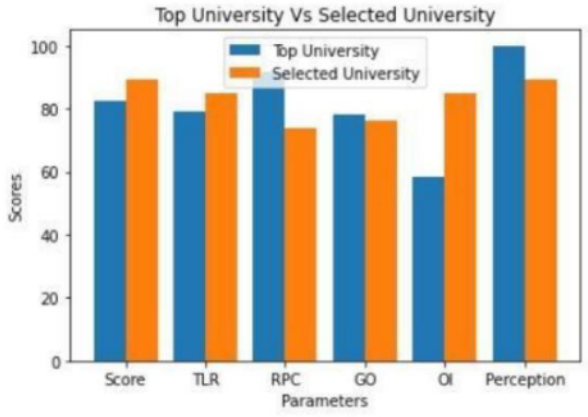
600 rows × 13 columns

DISCUSSIONS

Observe that Random Forest Regression gets the second-highest. R Square value and has consistently low Mean Absolute Error, Mean Square Error and Root Mean Square Error. So for these factors, we say that Random Forests Regression Algorithm gives the best results when tested with real-time data. Though the R Square value of the Decision Tree Regression Algorithm is higher than that of the Random Forests Regression Algorithm, we cannot recommend it for real-time use since its errors, such as Mean Absolute Error, Mean Square Error, and Root Mean Square Error, are not minimal.

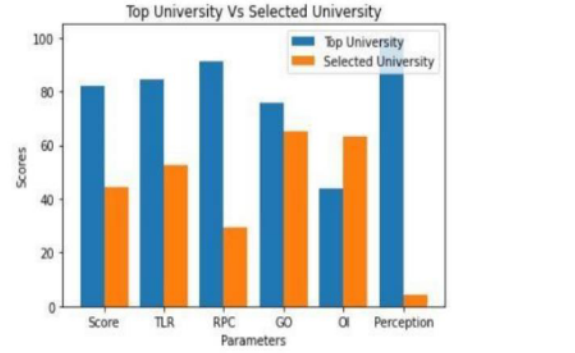
RESULTS:

```
Top University: [82.67, 79.13, 91.48, 78.23, 58.39, 100.0]
Selected University: [89, 85, 74, 76, 85, 89]
```



Volume 13, No. 3, 2022, p. 146 - 152
https://publishoa.com
ISSN: 1309-3452

```
Top University: [82.16, 84.54, 91.08, 75.48, 43.7, 100.0]
Selected University: [44.34, 52.38, 29.43, 65.31, 63.14, 4.3]
```



CONCLUSION:

Meant to develop different types of Machine Learning Models and compare them with each other and find the model which is best suited for Real-time usage. The model (Random Forest Regression Model) predicts the future ranks of the universities or colleges. As a part of this, developed a Recommendation system that tells the universities or colleges which parameters should be improved so as to improve their future ranks.

FUTURE SCOPE:

Creating a mobile application that connects to the application server giving users mobile access to the application. In-depth data analysis can be done by linking the

REFERENCES

1. MoE, National Institute Ranking Framework (NIRF), https://www.nirfindia.org.
2. NIRF Ranking Dataset https://www.kaggle.com/datasets/apoorvgupta25/nirf- rankings-from-2016-to-2021
3. MoE, National Institute Ranking Framework (NIRF),https://www.nirfindia.org/parameter.
4. Tabassum, Anika, et al. "University ranking prediction system by analyzing influential global performance indicators." 2017 9th International Conference on Knowledge and Smart Technology (KST). IEEE, 2017 , https://sci-hub.se/10.1109/KST.2017.7886119.
5. Singh, Vaibhav, et al. "University Ranking Prediction System." IJSRD - International Journal for Scientific Research Development—, vol. 9, no. 6, 2021, pp.49-53, http://www.ijsrd.com/articles/IJSRDV9I60029.pdf.
6. Çakır, Murat Perit, et al. "A comparative analysis of global and national university ranking systems." Scientometrics 103.3 (2015): 813-848,https://www.researchgate.net/publication/277351587 A comparative analysis of global and national university ranking systems
7. Mohammed, Fatima, and RehamanBee Abdul Subhan.University Classification and Prediction Using Data Mining , https://core.ac.uk/download/pdf/55305284.pdf
8. Comparing university rankings: (PDF) Comparing university rankings (researchgate.net)
9. Global ranking of higher education institutions : (PDF) Global Ranking of Higher Education Institutions (researchgate.net)
10. Global university rankings and their impact: global university rankings and their impact - report ii.pdf (eua.eu)
11. University rankings: Theoretical basis, methodology and impacts on global higher education: University Rankings: Theoretical Basis, Methodology and Impacts on Global Higher Education — SpringerLink.
12. "Decision Tree Regressor — scikit-learn 1.0.2 documentation", https://scikit-learn.org/stable/modules/generated/sklearn.tree..html.
13. "K Neighbors Regressor — scikit-learn 1.0.2 documentation" , https://scikit-learn.org/stable/modules/generated/sklearn.neighbors. KNeighborsRegressor.html
14. "Random Forest Regressor — scikit-learn 1.0.2 documentation", https://scikit-learn.org/stable/modules/generated/sklearn.ensemble. RandomForestRegressor.html

Volume 13, No. 3, 2022, p. 146 - 152
https://publishoa.com
ISSN: 1309-3452

15. "Linear Regression — scikit-learn 1.0.2 documentation", https://scikit-learn.org/stable/modules/generated/sklearn.linear model.LinearRegression.html
16. "Lasso Regression — scikit-learn 1.0.2 documentation", https://scikit-learn.org/stable/modules/generated/sklearn.linear model.Lasso.html
17. "Ridge Regression — scikit-learn 1.0.2 documentation", https://scikit-learn.org/stable/modules/generated/sklearn.linear model.Ridge.html.
18. "sklearn.metrics.r2 score — scikit-learn 1.0.2 documentation", https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2 score.htmlsklearn.metrics.r2 score
19. "mean absolute error — scikit-learn 1.0.2 documentation", https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean absolute error.htmlsklearn.metrics.mean absolute error
20. "mean squared error — scikit-learn 1.0.2 documentation", https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean squared error.htmlsklearn.metrics.mean squared error
21. Brownlee, Jason. "Regression Metrics for Machine Learning." Machine Learning Mastery, https://machinelearningmastery.com/regression- metrics-for-machine-learning/
22. zscore — SciPy v1.8.0 Manual https://docs.scipy.org/doc/ scipy/reference/generated/scipy.stats.zscore.html.
23. "NumPy" NumPy , https://numpy.org/doc/. 24."Pandas— pandas 1.4.2 documentation." Pandas ,https://pandas.pydata.org/docs/.
25. Welcome to Flask — Flask Documentation (2.1.x) ,https://flask.palletsprojects.com/en/2.1.x/.
26. scikit-learn: machine learning in Python — scikit-learn1.0.2 documentation , https://scikit-learn.org/stable/
27. "Matplotlib — Matplotlib 3.5.0 documentation."Matplotlib , https://matplotlib.org/stable/.
28. TheJupyterNotebook — Jupyter Notebook 6.4.10 documentation, https://jupyter-notebook.readthedocs.io/en/stable/ 29.Hushyar Sherwani, Karwan. "Comparative Analysis of National University Ranking System in Kurdistan- Region and Other National University Rankings:
29. An Emphasis on Criteria and Methodologies. " International Journal of Social Sciences Educational Studies 5.1 (2018): 7-15.
30. "World University Rankings Research Papers." Academia.edu, https://www.academia.edu/Documents/in/World University Rankings. https://www.elsevier.com/research-intelligence/university-rankings-guide.
31. Tu, P-L., and J-Y. Chung. "A new decision-tree classification algorithm for machine learning." . IEEE Computer Society,1992.
32. "Performance Ranking of Scientific Papers for National Universities." Wikipedia, https://en.wikipedia.org/wiki/Performance Ranking of Scientific Papers for National Universitie.
33. Garc ía, Nicolás Robinson, et al. "An insight into the importance of national university rankings in an international context: The case of the I-UGR Rankings of Spanish universities." Scientometrics.
34. Pusser, Brian, and Simon Marginson. "University rankings in critical perspective." The journal of higher education 84.4 (2013): 544-568.
35. Comparative study of international rankings: Comparative study of international academic rankings of universities — SpringerLink