CIS7031-Programming for Data Analysis

Module Leader : Dr. Angesh Anupam
Student Name : Vyshnavi Muthumula
Student Id : 20219772

Table of Contents

Task 1: What features must be considered to get good quality of teaching and which country is providing best quality of teaching?

Problem Statement: Education sector is very important sector in all sectors. When it comes to select the university student must need to take wise decision. Majority of the students choose universities based on Teaching of quality. The study of this research is to find the factors which affects teaching and predict r2 score.

2. Exploratory data analysis (EDA)

Analysing the data for business purposes is very crucial, for this if developer use programming for solving it will make the process easier.

There are different methods while analysing a dataset. They are

1. Identifying the data

2. Collection

3. Data Cleaning

4. Statistical / computational analysis

5. Machine learning

6. Visualizations

2.1 Data identification: Choose the data based on requirement.

2.2 Collection: The researcher collected this data from Kaggle in which users can upload their data collection, analyse it, and construct models in a web-based environment.

The shape of this dataset is (2063,14). The number of rows is 2063 and columns are 14. These are the columns world_rank, university_name, country, teaching, international, research, citations, income, total-score, num_students, student_staff_ratio, international_students, female_male_ratio, year. Out of these parameters only some parameters are affecting Teaching mainly research, citations, income, country, number of students and international. This dataset contains null values; the count of null values is around 153.

From researcher point of view these three columns (student_staff_ratio, international_students and female_male_ration) are not useful so they can be dropped.

2.3 Data Cleaning:

The process of recognising, updating or removing inaccurate, missing or unwanted data from a dataset is known as Data cleaning. Based on the dataset, user decides how to handle missing or inconsistent data.

If the dataset contains following issues then that data set needs to be cleaned otherwise it will affect the accuracy of the analysis.

1. Null / Missing values
2. Outliers
3. Repetitive and duplicate data
4. Inconsistent data

The world university ranking data set contains null (missing) values, outliers, and in some places "-" is present instead of values.

Handling missing values:

The researcher performed two different data cleaning techniques for visualisations and analysis.

Drop the column / entire row which contains nan:

For visualisations analysis author dropped all the null values from the data because replacing the null values with mean, mode or median cause issues for value distribution.

Imputing with mean or median or mode:

Income, total score and international fields contain special characters meant for some universities; those three values are filled with '-'. First thing, author can't delete those columns because those are important factors for describing teaching. So, either author can drop the rows or impute with mean or mode or median. After analysis, researchers found that dropping the rows causes more error. Instead, researchers did analysis by replacing nan values with median for those three columns income, total score and international (converted all unpredictable values ('-') with nan then those nan values were replaced with median).

Outliers:

Outliers are the unwanted data present in the data set which will cause noise for analysis and removing these will result in better performance. The target variable teaching contains many outliers so for better performance remove the outliers from the teaching column as shown in the notebook. By using quantiles, the outliers are removed. Box plots or histogram can detect the outliers. Handling outliers is the same as handling missing values either the user can drop or impute it.

Repetitive and duplicate data:

In this data set the country column contains duplicate values which lead to bad visualizations. This needs to be avoided by taking unique values from the data set. Trimmed unwanted spaces from this column.

Data inconsistency:

The university name column data is mixed with some special character. This column is no longer useful for analysis so hasn't performed any operations for solving that data inconsistency issue.

2.4 Statistical / computational analysis

By using describe function researchers can identify the statistical values for all the features.

The maximum value for teaching is 99.7 and min value is 9.9. Mean 37.8 and standard deviation is 17.6. The remaining statistical values for each column as shown in below figure.

```
uniDataframe2.describe()
```

	teaching	international	research	citations	income	total_score	num_students	student_staff_ratio	year
count	2603.000000	2603.000000	2603.000000	2603.000000	2603.000000	2603.000000	2603.000000	2544.000000	2603.000000
mean	37.801498	52.001537	35.910257	60.921629	53.252785	54.004303	23382.905878	18.445283	2014.075682
std	17.604218	22.065792	21.254805	23.073219	24.714967	10.239459	17769.283295	11.458698	1.685733
min	9.900000	7.100000	2.900000	1.200000	24.200000	41.400000	462.000000	0.600000	2011.000000
25%	24.700000	33.500000	19.600000	45.500000	33.600000	49.000000	12199.500000	11.975000	2013.000000
50%	33.900000	50.300000	30.500000	62.500000	42.600000	49.000000	20584.000000	16.100000	2014.000000
75%	46.400000	69.000000	47.250000	79.050000	69.500000	54.800000	29787.000000	21.500000	2016.000000
max	99.700000	100.000000	99.400000	100.000000	100.000000	96.100000	379231.000000	162.600000	2016.000000

Figure 1

The z-score of the teaching is shown in the notebook and the highest z-score is 2.4.

The median of the university data set is shown below

```
[63] uniDataframe.median()

/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning:

Dropping of nuisance columns in DataFrame reductions (with 'numeric_only=None')

teaching               46.20
international          61.60
research               47.65
citations              79.40
income                 46.60
total_score            55.45
student_staff_ratio    14.90
year                 2014.00
dtype: float64
```

Figure 2

2.5 Visualisations

Visualising the data will allow the user to understand, observe or analyse the patterns easily.

In this entire study the researcher observed how teaching is varying with other input variable.

Correlogram analysis: It is used to find the relationship between target variable with all the inputs. And able to decide which factor is more affecting the target. It varies from (-1, 1). If the value is equal to 0, means it is on the line, greater than 0 for positive correlation, less than 0 is negative correlations.

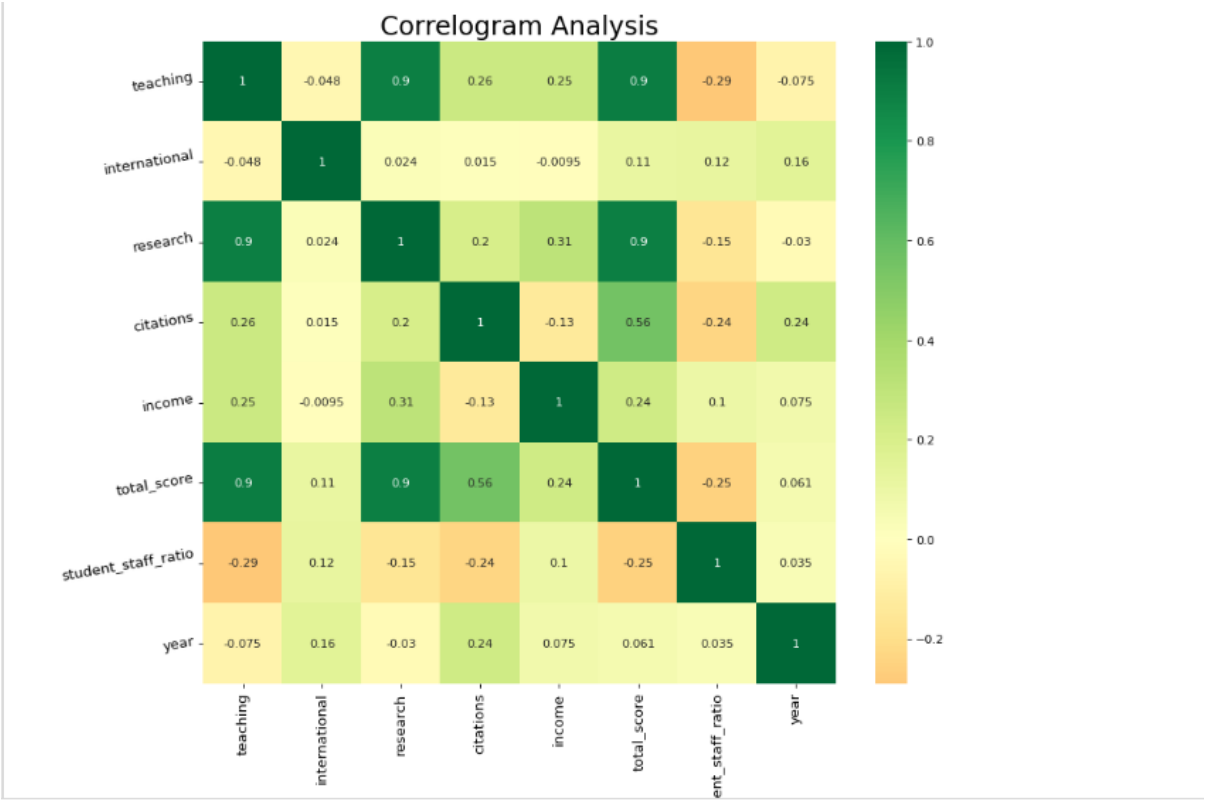

Figure 3

By using the above graph user can say teaching has only 3 negative values,

How researcher affecting teaching quality:

The purpose of this plot is to find the relationship between researcher and teaching quality. By using **"linear model"** (lmplot) user can predict the linearity between target and source. The reason choosing this graph is to find the correlation between research (input variable) and teaching. Based on the analysis Teaching variable is increasing with respect to input.

Here Teaching variable have positive correlation with research.

```
sns.set_style('whitegrid')
sns.lmplot(x ='research', y ='teaching', data = uniDataframe)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fc199b9b1d0>
```

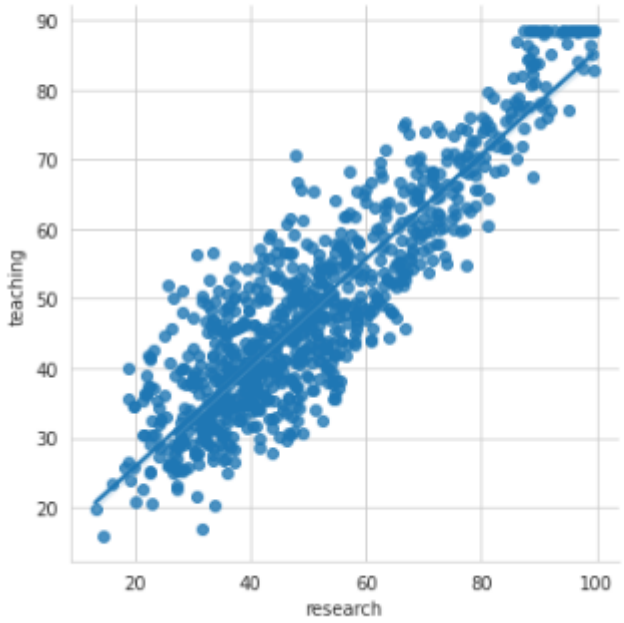

Figure 4

How total score affecting teaching quality:

Based on the analysis it found out that teaching is also affected by total score and it also has positive correlation. The reason for positive correlation is that the total score is calculated by using number of citations, research, num_ students, Female_male _ratio, student _staff _ratio and income along with teaching. So, if total score effecting then wise versa will also happens. If she / he observe the plot then will understood that the correlation is much higher than researcher. But both have positive correlation only means both input and output variables move in same direction. User can plot the graph either by using liner model or scatter plot here user considered to take Scatter plot with hue as Country. Hue can be also considered as input only. By using the below user can say there are still some more unwanted data which needs to have more cleaning.

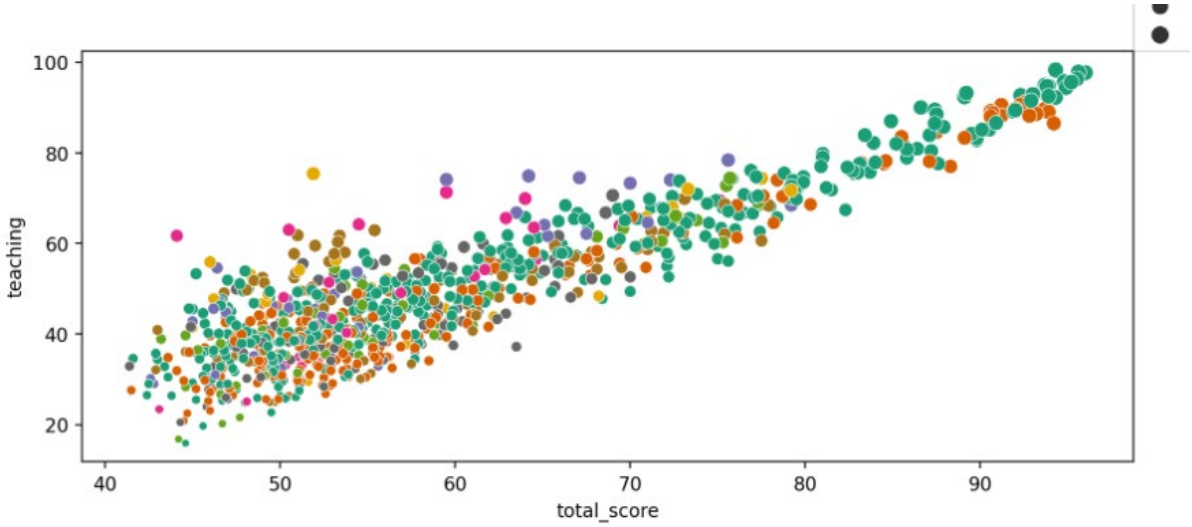

Figure 5

Affect of citations on Teaching: The relation between citations and teaching is defined by using swarm plot. This plot is used to plot the data which calculates non-overlapping positions for avoiding unclear values. The citations and teaching have positive correlation. The plots are represented in .ipynb file.

How Teaching quality depends on income: By using scatter plot the relation is described. The income field has more missing values and special characters even after imputing also the results are pretty good for this plot. Consecutive missing / unpredictable are present, so after replacing those with median the columns which have same values are more. That's why some values are straight line in the graph it shown (some values are constant at the last).

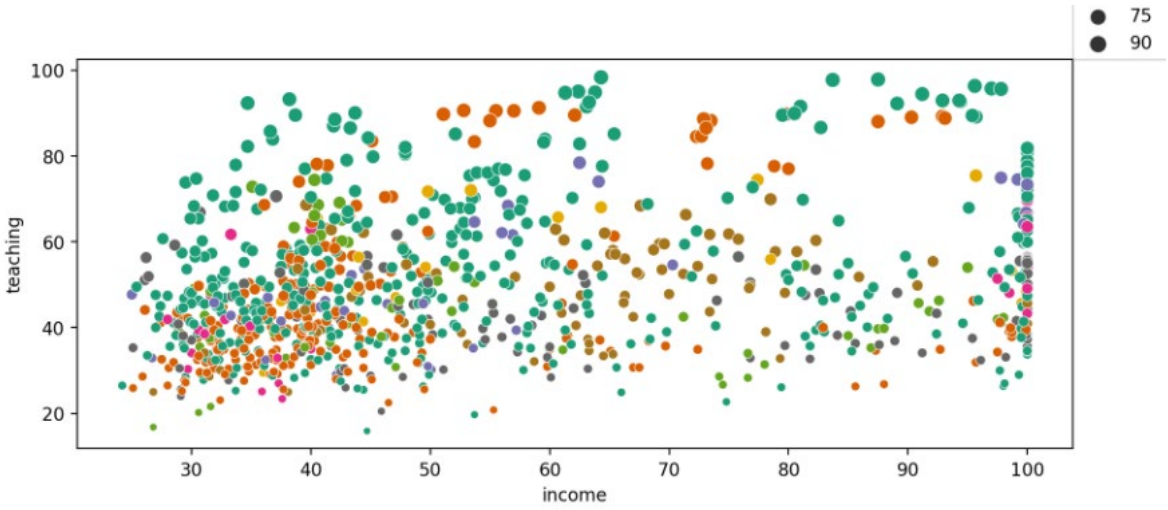

Figure 6

2.5 Machine learning algorithms:

These algorithms trained the data to predict the future output. For this analysis author performed 3 algorithms such as Linear Regression, K-Nearest Neighbours and SVR because the target variable is continuous that's why it is a regression supervised model.

The country data type conversion from String to float is done by using **LabelEncoder** and used for normalizing the input.

Linear Regression

This regression model is used to find relationships between input variables and target variables. The mean absolute error is 4.8 and mean squared error is 6.3. The r2 square error is 88% which means the correlation between source and target variables is high.

Figure 7

Support Vector Regression (SVR)

This regression supports both linear and non-linear regressions but in this analysis author perform only linear regression as the data is in linear format. This algorithm is mainly used to find the relation between linear models and it provides the flexibility to adjust the model to produce correct outcomes. The SVR result for this university ranking data is 6.1 absolute error, 8.5 mean squared error and 0.78 r2 score. This model also gives a good r2 score for teaching target variables.

Figure 8

K-Nearest Neighbours Regression (KNN)

This KNN uses a similarity concept to predict the target variable destination. The point which is closer to the target new point is considered. The university ranking algorithm results output for KNN is 0.87 r2 score.

```
from sklearn import metrics
print("Mean absolute error: ", metrics.mean_absolute_error(Y_test, testDataAccuracy))
```

```
Mean absolute error:  4.7832906530089625
```

```
Ks = 20
r2Score = np.zeros((Ks-1))
abs_error = np.zeros((Ks-1))
for n in range(1,Ks):

    #Train Model and Predict
    neigh = KNeighborsRegressor(n_neighbors = n).fit(X_train,Y_train)
    testDataPred=neigh.predict(X_test)
    r2Score[n-1] = metrics.r2_score(Y_test, testDataPred)
    abs_error[n-1] = metrics.mean_absolute_error(Y_test, testDataPred)


r2Score
abs_error
```

```
array([5.1099872 , 4.78329065, 4.85057618, 4.86338028, 4.93162612,
       5.01899274, 5.08562283, 5.14081306, 5.19418125, 5.2678233 ,
       5.33425678, 5.41407384, 5.46999902, 5.51119444, 5.55690141,
       5.59383803, 5.67655344, 5.74561815, 5.81921962])
```

```
plt.plot(range(1,Ks),r2Score)
plt.title('R2 score vs Neighbours')
plt.ylabel('R2 score ')
plt.xlabel('Number of Neighbours (K)')
```

```
Text(0.5, 0, 'Number of Neighbours (K)')
```

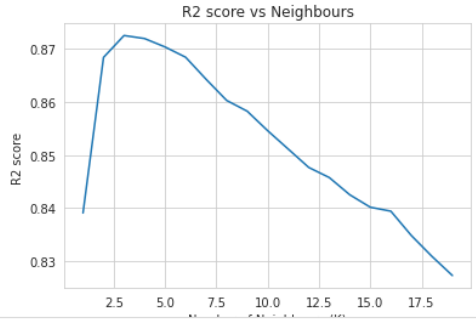

Figure 9

Based on the analysis, the researcher concluded the best fit model for this analysis is ***Linear Regression because of a good r2 score***.

Conclusion:

The result of this analysis is that **Research** column is most affecting variable for determining **Teaching**. United States is providing good quality of teaching and then Russia. The top universities are also listed in United States of America. So that researcher can conclude that by using word count plot and count plot United stated of America is providing best quality teaching. The KNN algorithm produced best r2 score compared to Linear regression and support vector regression.

Limitations:

The data set is taken from Kaggle and it contains more number of null values, irregular data, missing data and improper data because of this problems the analysis and visualizations are not proper. Researcher performed more cleaning but still means absolute error is high. There are more machine learning algorithms which provide better r2 score.

References:

Cox, V., 2017. Exploratory data analysis. In *Translating Statistics to Make Decisions* (pp. 47-74). Apress, Berkeley, CA.

Velleman, P.F. and Hoaglin, D.C., 1981. *Applications, basics, and computing of exploratory data analysis*. Duxbury Press.

Pandey, A. and Jain, A., 2017. Comparative analysis of KNN algorithm using various normalization techniques. *International Journal of Computer Network and Information Security*, *9*(11), p.36.

Yan, X. and Su, X., 2009. *Linear regression analysis: theory and computing*. World Scientific.

Yeganefar, A., Niknam, S.A. and Asadi, R., 2019. The use of support vector machine, neural network, and regression analysis to predict and optimize surface roughness and cutting forces in milling. *The International Journal of Advanced Manufacturing Technology*, *105*(1), pp.951-965.