Prediction of MS Graduate Admissions using Decision Tree Algorithm

Janani P[1], Hema Priya V[2], Monisha Priya S[3]

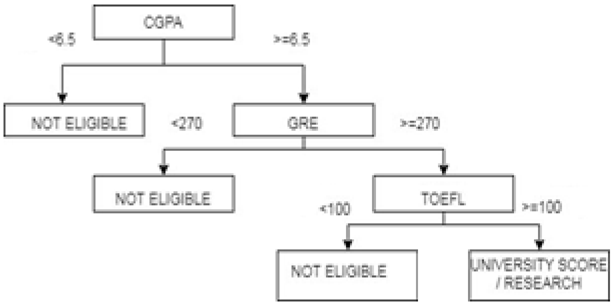
Department of Computer Science and Engineering, R.M.K. College of Engineering and Technology

Abstract: ***Every year, the number of students wanting to pursue higher studies in abroad especially in US is more and they find difficult to search the best university. And thus, this paper helps on predicting the eligibility of Indian students getting admission in best university based on their Test attributes like GRE,TOEFL,CGPA, Research papers published etc. According to their scores the possibilities of chance of admit is calculated. This project uses machine learning technique specifically decision tree algorithm(version 0.22.1) for predicting the output which helps them to admit in the best university and also helps the student to find the possibility of admitting in other universities based on their scores.***

Keywords: Machine learning, decision tree, best university, admission

1. Introduction

The future of the traditional higher education model [1] is an important topic of discussion. These issues are of course highly dependent on the type of academic institution being considered, such as public colleges and universities. Most Indian students face difficulties in enrolling in abroad universities especially US based university and thus Graduate admission is useful in bridging the gap between Indian students and US based university where student can find better university they needed, without any inconvenience.The number of Indian students (both undergraduate and graduate) enrolling in the US crossed 1M for the first time in 2015-16 as per an IIE report backed by the State department [2]. Thus, there must be web application to find their corresponding universities that might help the students to reduce the cost on Admission counselling and apply to those universities which they are eligible and which they are interested. The admission prediction uses binary classification problem using supervised machine learning by classifying the student into one of two categories: *Eligible* and *Not eligible*. Supervised machine learning problems are a class of problems which can generalize and learn from labelled training data, i.e., a set of data where the correct classification is known [3].



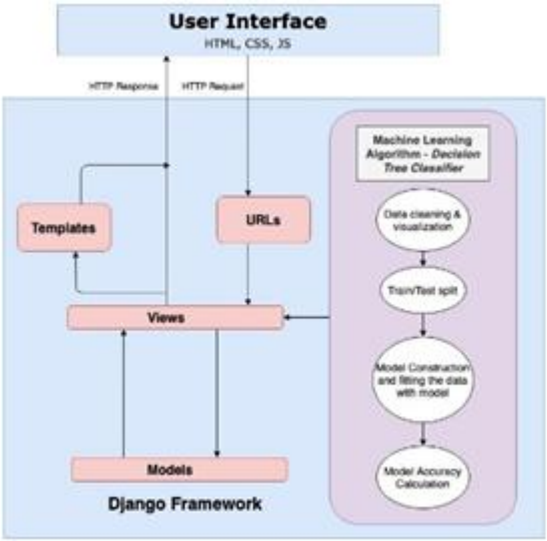
The above decision tree classifier illustrates the students getting admitted based on the various scores and predicts whether they are eligible or not in corresponding university which they are interested. There are various aspects to calculate the score (i.e.) CGPA should be above 65% and GRE should be above 270 and TOEFL score should be greater than 100 , if the student fulfils the above criteria then he/she is eligible to admit in the university which they are interested.

2. Literature Review

There are many applications of machine learning techniques to analyse data and other information in the context of educational settings. This area of study is generally known as "educational data mining" (EDM) and it is a recently emerging field with its own journals [7], conferences [8] and research community [9] .A subset of EDM research that focuses on analysing data in order to allow institutions of higher education better clarity and predictability on the size of their student bodies is often known as enrolment management. Enrolment management is "an organizational concept and systematic set of activities whose purpose is to exert influence over student enrolment"[6].A recent study [11] published this year reveals some key factors in the decision process and, consequently, allows to propose a recommendation algorithm that provides applicants the ability to make an informed decision regarding where to apply. There are many websites which predict college admission from the perspective of an aspiring student, each website are unique and our website displays top ten universities with their official websites and their score range which they expect from their student.

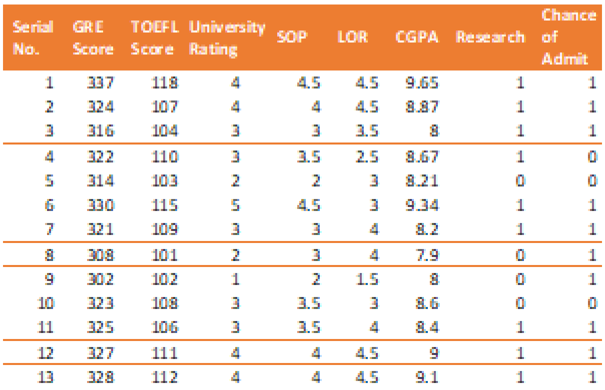
3. System Architecture

Architecture consists of User Interface and Django framework. HTML, CSS, JS is used to implement user interface. And decision tree classifier algorithm is used for analysing the data and Django framework which consist of collection of models which makes development easier and it is also used for both frontend and backend, it makes a convenient way to generate dynamic HTML pages by using template system. It consists of view where business logic is performed and returns response to the user. And template is used to build dynamic web pages and system uses HTTP for requesting information from backend.


User Interface
HTML, CSS, JS
HTTP Response
HTTP Request
Templates
URLs
Views
Models
Django Framework
Machine Learning Algorithm - Decision Tree Classifier
Data cleaning & visualization
Train/Test split
Model Construction and fitting the data with model
Model Accuracy Calculation

4. Data

This project uses dataset which is obtained from Kaggle platform .The dataset contains 500 rows and 9 features.

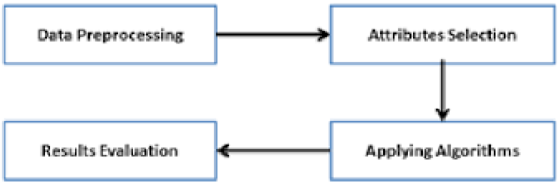


Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	1
2	324	107	4	4	4.5	8.87	1	1
3	316	104	3	3	3.5	8	1	1
4	322	110	3	3.5	2.5	8.67	1	0
5	314	103	2	2	3	8.21	0	0
6	330	115	5	4.5	3	9.34	1	1
7	321	109	3	3	4	8.2	1	1
8	308	101	2	3	4	7.9	0	1
9	302	102	1	2	1.5	8	0	1
10	323	108	3	3.5	3	8.6	0	0
11	325	106	3	3.5	4	8.4	1	1
12	327	111	4	4	4.5	9	1	1
13	328	112	4	4	4.5	9.1	1	1

GRE is a complex feature consisting of GRE Quant, GRE Verbal and GRE AWA score. Similarly, TOEFL is also a complex feature consisting of TOEFL score and essay score. Chance of admit is calculated based on the inputs .After obtaining the dataset it is mandatory to pre-process the data which includes checking missing variables, checking null values etc by analysing data.

5. Exploratory Data Analysis

It is important to analyse the data such as data gathering & data cleaning. The major issues which we deal are
1) Missing Data
2) Inconsistent Data.



Data cleaning and visualization involves analysing the dataset which consist of dropping irrelevant columns. In this dataset "Serial No" is not required.

```
ds.drop([‘SerialNo’],axis=1,inplace=True)
```

And by using above python code the serial no is removed and next process involves checking whether the data contains any null values in the dataset.

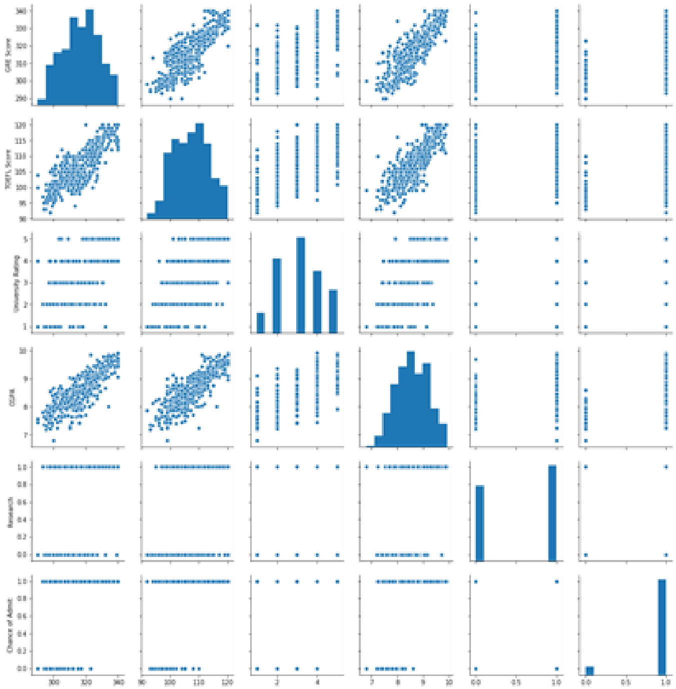
```
ds.isnull().sum()
```

The above line calculates the null value, if the sum is 0 it doesn't contain any value.

The data after cleaning must be trained and tested to fit the data into the decision tree model .There are many libraries available in python to visualize the data. One such library includes seaborn and matplotlib etc. One way is by using pair plot which visualize the relationship between two variables and it is usually a grid of plots for each variable in the dataset.

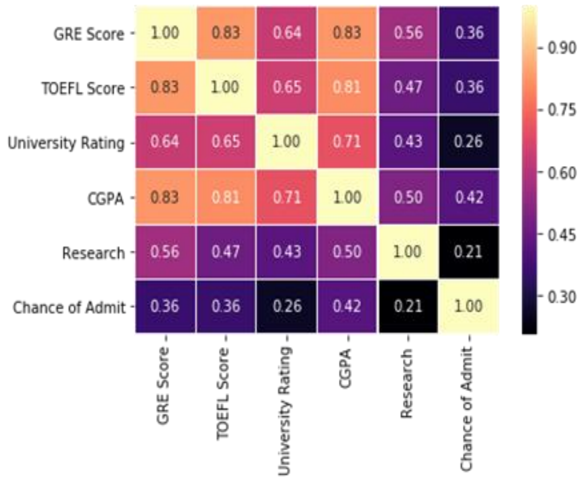
```
sns.pairplot(ds)
plt.show()
```

The pair plot compares all the attributes in dataset and produces different types of plots.



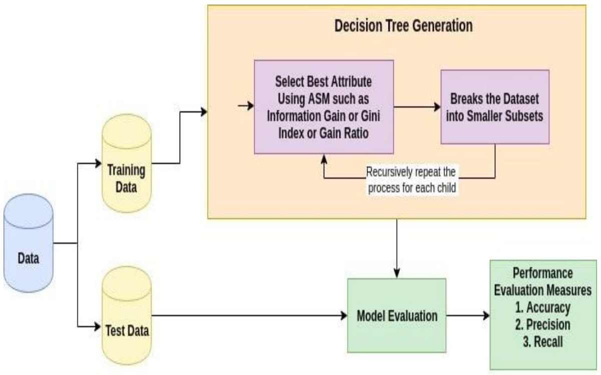
And using heatmap we can visualize the data. It is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors.

```
sns.heatmap(ds.corr(),annot=True,linewidths    =0.10,fmt =
'.2f',cmap='magma')
plt.show()
```



6. Algorithm

Decision tree classifier is used to predict the output, due to its simple logic, effectiveness and interpretability, it is the most widely used classification algorithm. The model works by creating a tree-like structure by dividing the data-set into several smaller subsets based on different conditional logic. The main components of the decision tree are the decision nodes, leaf nodes and the branches. Nodes with multiple branches are the decision nodes, nodes with no branches are called the leaf nodes, and the top node is called the root node of the decision tree. The nodes are connected to each other via branches based which are different conditions. The root and decision nodes are created by computing the entropy and information gain for the data-set.


Decision Tree Generation
Select Best Attribute Using ASM such as Information Gain or Gini Index or Gain Ratio
Breaks the Dataset into Smaller Subsets
Recursively repeat the process for each child
Training Data
Data
Test Data
Model Evaluation
Performance Evaluation Measures
1. Accuracy
2. Precision
3. Recall

7. Modeling and Results

After fitting the data with appropriate model the result obtained will be either 1 (i.e.) they are eligible or 0 (i.e.) they are not eligible. Based on the visualization from pairplot and heatmap it is observed that CGPA is most important feature in predicting the admission of student. There are many libraries available in python for computation such as NumPy, SciPy and matplotlib.

```
from          sklearn.metrics          import
accuracy_score
y_pred = mdl.predict(X_test)
print(accuracy_score(y_pred,y_test))
```

Accuracy obtained by decision tree classifier in predicting the output is 93%.

8. Conclusion

In this paper we focused on factors that influence the enrolment of applicants. We use machine learning methods to measure the level of correlation between enrolment and such factors. The results show that our proposed models can predict enrolment with reliable accuracy using only a small set of features related to student. This project helps the student to find the best university in US without affording.

9. Limitations

1) Authenticity:
The value which we took from Kaggle platform is accurate and accuracy is more. If input data produced by the student is not correct or gives the wrong score and it produces wrong output and the accuracy will be lowered. If we authentic the student, we can come up with more accurate prediction

2) Advanced Modelling (NLP):
We should advance NLP techniques to capture SOP, LOR, etc. We have considered only score such us GRE, CGPA, TOEFL for prediction, but SOP & LOR play a major role in Graduate student admissions. Accuracy is reduced if we consider SOA and LOR. We should try to rank the students based on Internship or full-time work done by them. Furthermore, we should also categorize conferences and journals where applicant has submitted his work.

References

[1] Lapovsky, L. The Changing Business Model For Colleges And Universities. *Forbes* 2018. Available online: https://www.forbes.com/sites/lucielapovsky/2018/02/06/the-changing-business-model-for-colleges-and-universities/#bbc03d45ed59 (accessed on 15 December 2018).

[2] IIE Report http://www.iie.org/Services/Project-Atlas/United-States/International-Students-In-US

[3] Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* 2007, *4*, 249–268. [Google Scholar]

[4] Alpaydin, E. *Introduction to Machine Learning*, 3rd ed.; MIT Press: Cambridge, MA, USA, 2010. [Google Scholar]

[5] Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*, 2nd ed.; McGraw Hill; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2014. [Google Scholar]

[6] Hossler, D.; Bean, J.P. *The Strategic Management of College Enrollments*, 1st ed.; Jossey Bass: San Francisco, CA, USA, 1990. [Google Scholar]

[7] Journal of Educational Data Mining. Available online: http://jedm.educationaldatamining.org/index.php/JEDM (accessed on 15 December 2018).

[8] Educational Data Mining Conference 2018. Available online: http://educationaldatamining.org/EDM2018/ (accessed on 15 December 2018).

[9] Romero, C.; Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert Syst. Appl.* 2007, *33*, 135–146. [Google Scholar] [CrossRef]

[10] Peña-Ayala, A. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Syst. Appl.* 2014, *41*, 1432–1462. [Google Scholar]

[11] American graduate admissions: both sides of the table http://hdl.handle.net/2142/92866