Prediction for University Admission using Machine Learning

Chithra Apoorva D A, Malepati ChanduNath, Peta Rohith, Bindu Shree.S, Swaroop.S

Abstract: In today's education world there are many number of students who want to pursue higher education after engineering or any graduate degree course. Higher education in the sense, some people want to do M.tech through GATE or through any educational institute entrance examination and some people want to do MBA through CAT or through any respective educational institute entrance examination and some people want to do Masters in abroad universities. we are focusing on only the students who want to pursue their higher education in abroad universities. Generally Higher education in abroad universities means we have many options like canada, USA ,UK Germany, Italy, Australia etc. But we are focusing on only the students who want to do their Masters in America. Students who want to do masters in America have to write GRE (Graduate Records Examination) and TOEFL/IELTS (Test of English as a Foreign Language/International English Language Testing System). Once they have attended the exams they have to prepare their SOP(statement of purpose) and LOR(letter of reccomendation) which are one of the crucial factors they have to consider. These LOR and SOP plays a vital role if the student was looking for any scholarship. Then the students have to choose the universities they want to study or apply, we cannot apply to all the universities that will lead to lot of application fees. Here comes the problem that the student dontt know to which university he might get admission. There are some online blogs which help in these matter but they are not that much accurate and dont consider all the factors and there are some consultancy offices which will take lot of our money and time and sometimes they will give some false information.so our goal is to develop a model which will tell the students their chance of admission into a respective university. This model should consider all the crucial factors which plays a vital role in student admission process and should have high accuracy. The model name is UAP. To access this model we will develop a simple user interface.

Keywords: College admission predictor; Machine Learning

Revised Manuscript Received on February 01, 2020.
***** Correspondence Author

Chithra Apoorva D.A his/her department, Name of the affiliated College or University/Industry, City, Country. Email: xyz1@blueeyesintlligence.org

Malepati ChanduNath, department, Name of the affiliated College or University/Industry, City, Country. Email: xyz2@blueeyesintlligence.org

Peta Rohith, department, Name of the affiliated College or University/Industry, City, Country. Email: xyz3@blueeyesintlligence.org

Bindu Shree.S, department, Name of the affiliated College or University/Industry, City, Country. Email: xyz2@blueeyesintlligence.org

Swaroop.S, department, Name of the affiliated College or University/Industry, City, Country. Email: xyz2@blueeyesintlligence.org

I. INTRODUCTION

Specific preparation plays a crucial part in your life. Thus education preparation students often have multiple questions about universities which they can get admission and scholarship and accommodation. One of the main concerns is getting admitted to their dream university[1]. It's seen that students still choose to obtain their education from universities that are known internationally. And when it comes to international graduates, the United States of America is the first preference of the majority of them. With most world-renowned colleges, Wide variety of courses available in each discipline, highly accredited education and teaching programs, student scholarships, are available for international students. According to estimates, there are more than 10 million international students enrolled in over 4200 universities and colleges including both private and public across the United StatesMost number of students studying in America are from Asian countries like India, Pakistan, Srilanka, Japan and China. They are choosing not only America but also UK, Germany, Italy, Australia and Canada. The number of people pursuing higher studies in these countries are rapidly increasing. The background reason for the students going to abroad universities for Masters is the no. of job oppurtunities present are low and number of people for those jobs are very high in their respective countries[2]. This inspires many students in their profession to pursue postgraduate studies. It is seen that there is quite a large number of students from universities in the USA pursuing Masters in the field of computer science, the emphasis of this research will be on these students.

Many colleges in the U.S. follow similar requirements for student admission. Colleges take different factors into account, such as the ranking on aptitude assessment and academic record review[3]. The command over the English language is calculated on the basis of their performance in the English skills test, such as TOEFL and IELTS. The admission committee of universities takes the decision to approve or reject a specific candidate on the basis of the overall profile of the applicant application.

II. LITERATURE SURVEY

This section includes the literature review of previous research on the assessment of student enrolment opportunities in universities. Numerous programs and studies have been carried out on topics relating to university admission used many machine learning models which helps the students in the admission process to their desired universities. Previous research done in this area used Naive Bayes algorithm which will evaluate the success probability of student application into a respective university but the main drawback is they didn't consider all the factors which will contribute in the student admission process like TOEFL/IELTS, SOP, LOR and under graduate score.

Bayesian Networks Algorithm have been used to create a decision support network for evaluating the application submitted by foreign students of the university[5]. This model was developed to forecast the progress of prospective students by comparing the score of students currently studying at university. The model thus predicted whether the aspiring student should be admitted to university on the basis of various scores of students. Since the comparisions are made only with students who got admission into the universities but not with students who got their admission rejected so this method will not be that much accurate.

III. METHODOLOGY

Problem Understanding: Initially first we have to spend some time on what are the problems or concerns students having during their pre admission period and we should set the solutions to those problems as objectives of this research.

Data Understanding: Data should be collected from multiple sources like yocket and also consider all the factors including which will play a tiny role in student admission process.

Data Preparation: Data should be cleaned that is removing the noise in the data and filling the missing values or extreme values and finalising the attributes/factors which will have crucial importance in student admission process.

Building Models: several ML models have to be developed using various machine learning algorithms for admission to a particular university and the user interface has to be developed to access those models[6].

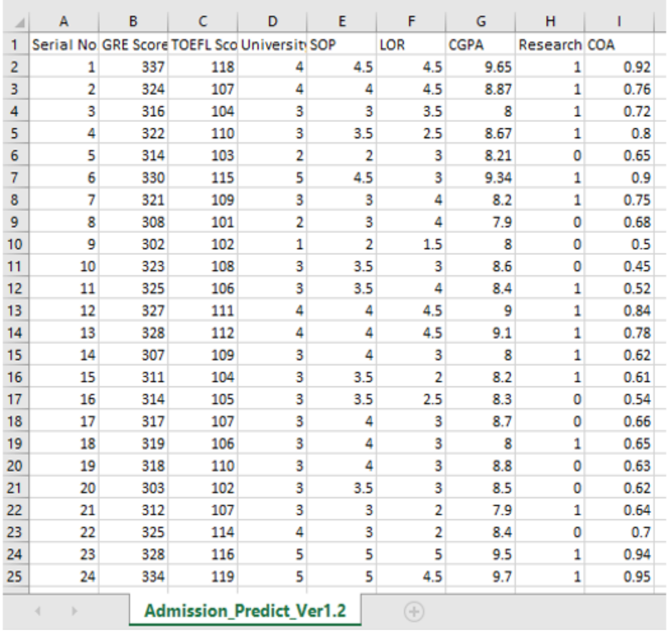
Evaluation: Developed models are evaluated according to their accuracy scores. Once the model is finalised that model will be merged with node red for final deployment.

Data Cleaning and Analysis:

- Inspecting feature values that help identify what needs to be done to clean or pre-process until you see the range or distribution of values typical of each attribute.
- You may find missing or noisy data, or anomalies such as the incorrect data form used for a column, incorrect measuring units for a particular column[7], or that there are not enough examples of a specific class.
- You can know that without machine learning, the problem is actually solvable.

The data cleaning process has several key benefits to it:

1. This eliminates major errors and inconsistencies which are unavoidable when dragging multiple data sources into one dataset.
2. Having data cleaning software will make everyone more effective as they will be able to get easily from the data what they need.
3. Fewer mistakes mean happy clients, and less unhappy workers.
4. The ability to chart the various functions, and what your data is supposed to do and where it comes from your data.



	A	B	C	D	E	F	G	H	I
1	Serial No	GRE Score	TOEFL Sco	Universit	SOP	LOR	CGPA	Research	COA
2	1	337	118	4	4.5	4.5	9.65	1	0.92
3	2	324	107	4	4	4.5	8.87	1	0.76
4	3	316	104	3	3	3.5	8	1	0.72
5	4	322	110	3	3.5	2.5	8.67	1	0.8
6	5	314	103	2	2	3	8.21	0	0.65
7	6	330	115	5	4.5	3	9.34	1	0.9
8	7	321	109	3	3	4	8.2	1	0.75
9	8	308	101	2	3	4	7.9	0	0.68
10	9	302	102	1	2	1.5	8	0	0.5
11	10	323	108	3	3.5	3	8.6	0	0.45
12	11	325	106	3	3.5	4	8.4	1	0.52
13	12	327	111	4	4	4.5	9	1	0.84
14	13	328	112	4	4	4.5	9.1	1	0.78
15	14	307	109	3	4	3	8	1	0.62
16	15	311	104	3	3.5	2	8.2	1	0.61
17	16	314	105	3	3.5	2.5	8.3	0	0.54
18	17	317	107	3	4	3	8.7	0	0.66
19	18	319	106	3	4	3	8	1	0.65
20	19	318	110	3	4	3	8.8	0	0.63
21	20	303	102	3	3.5	3	8.5	0	0.62
22	21	312	107	3	3	2	7.9	1	0.64
23	22	325	114	4	3	2	8.4	0	0.7
24	23	328	116	5	5	5	9.5	1	0.94
25	24	334	119	5	5	4.5	9.7	1	0.95

Figure 1. Admission Prediction csv data

- There are no missing values and outliers because we analysed the data, so for this data there is no need to fill the missing values and deal with outliers. If there are any missing values and outliers we can fill (or) drop using the fillna method and drop method and we can also standardize the data using the min-max scaler, if necessary.

Data Visualization:

- After analysing the data, we will be able to know what the features and labels are, so from the above data, the label we have to consider is Chance of Admission[8] and then we have to consider the parameters that influence or play a major role in Chance of Admission
- We can get to know certain features that are more affected by the visualization (or) analysis or the use of feature importance method in decision tree
- Below are some of our data visualizations[9]

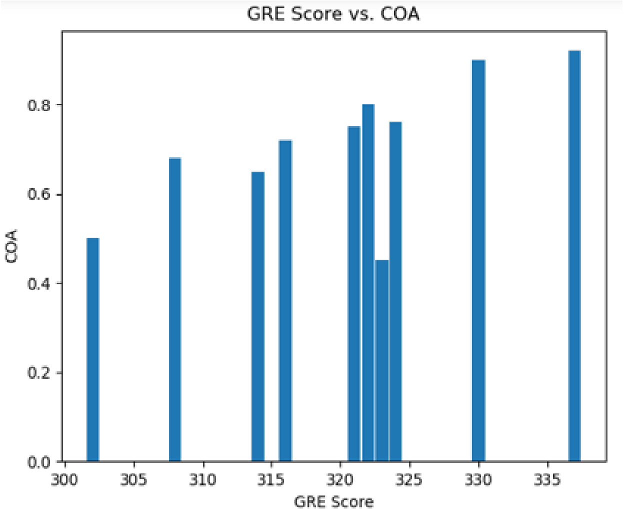


Figure 2. GRE Score Vs Chance of Admit

As we can see from above, students with strong GRE score have high chances of being admitted.

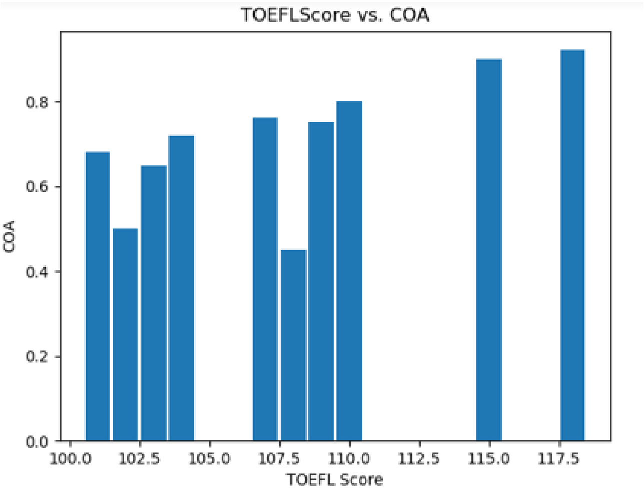


Figure 3. TOFEL Vs Chance of Admit

As we can see from above, students with strong TOFEL score have high chances of being admitted.

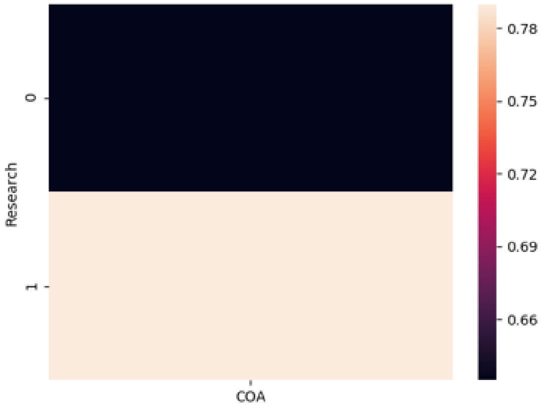


Figure 4. Research Vs Chance of Admit

As we can see from above, students who have had a high chance of accepting work.

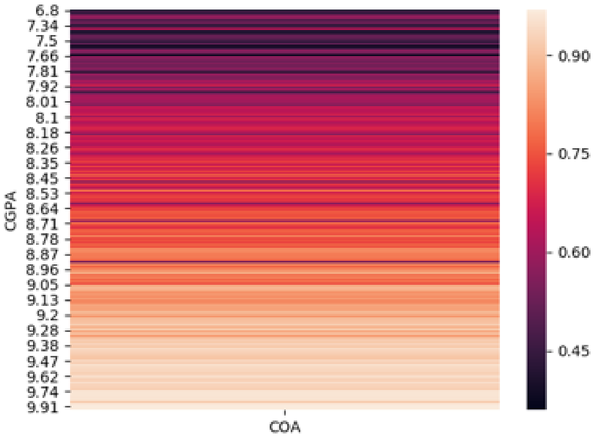


Figure 5. CGPA Vs Chance of Admit

As we can see from the figure above, students with strong CGPA are very likely to admit.

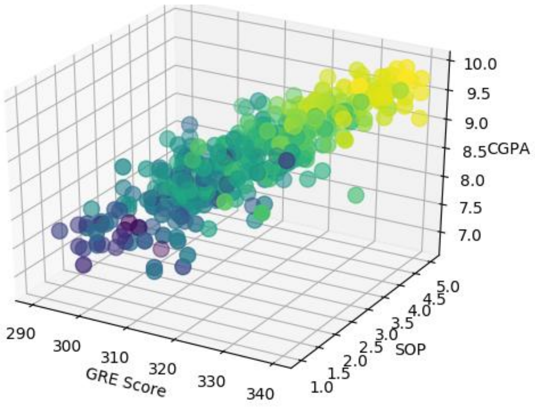
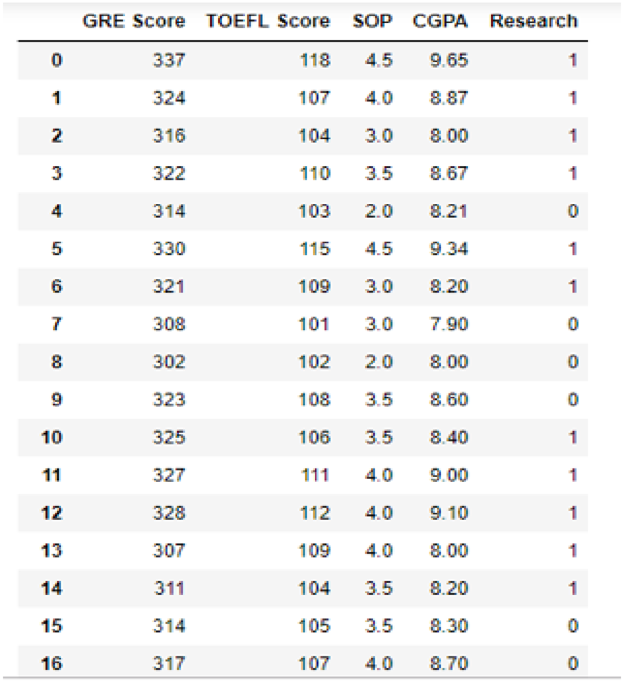


Figure 6. 3d Diagram

As a result of the above visualization and data analysis, the features given have a high impact on the probability of admission, so these features are considered only.



	GRE Score	TOEFL Score	SOP	CGPA	Research
0	337	118	4.5	9.65	1
1	324	107	4.0	8.87	1
2	316	104	3.0	8.00	1
3	322	110	3.5	8.67	1
4	314	103	2.0	8.21	0
5	330	115	4.5	9.34	1
6	321	109	3.0	8.20	1
7	308	101	3.0	7.90	0
8	302	102	2.0	8.00	0
9	323	108	3.5	8.60	0
10	325	106	3.5	8.40	1
11	327	111	4.0	9.00	1
12	328	112	4.0	9.10	1
13	307	109	4.0	8.00	1
14	311	104	3.5	8.20	1
15	314	105	3.5	8.30	0
16	317	107	4.0	8.70	0

FIGURE 7. Finalised Data

- Once the data visualization is done, we have to do predictive modelling for this purpose first we divide the data into train part and test part.
- we will develop model using machine learning algorithms on the train data and test model accuracy on the test data part.
- we will see which algorithms giving highest accuracy according to what parameters and take that for final consideration.

IV. ALGORITHMS

For this work, several machine learning algorithms have been used, K- Nearest Neighbour and Linear Regression, Ridge Regression, Random Forest[4] are used to predict students ' likelihood of university admission based on their profile.

K-Nearest Neighbours: KNN algorithm is the most commonly used algorithm for classification and regression purpose. KNN stands for k nearest neighbour, here k indicates a integer value which will tell that with how many neighbours comparisions should be made. It can be used for both classification and regression purpose. Suppose if it is classification and the k value is 5 it will compare with nearest 5 neighbours and gives the mode value, if it is regression and the k value is 6 it will take the nearest six values and return its mean value.

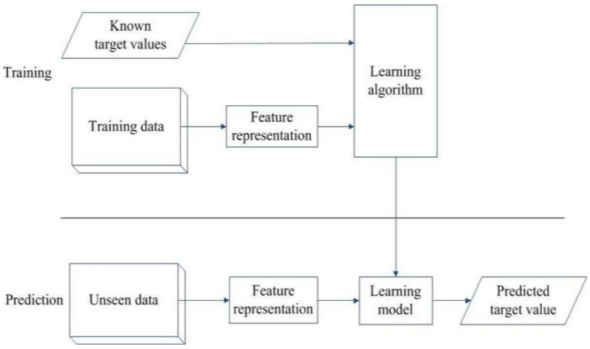
Linear Regression: It is an algorithm based on supervised learning of computers. It does the role of regression. Regression models[6] a predictive goal value based on the independent variables. Mostly it is used to figure out the relation between variables and forecasting. Different regression models vary on the basis–the form of relationship between dependent and independent variables, are considered, and the number of independent variables used.

Ridge Regression: Ridge regression is a regression method that is quite similar to unadorned minus squares linear regression: simply adding a ‖22 penalty on parameters β to the linear regression objective function gives the ridge regression objective function.

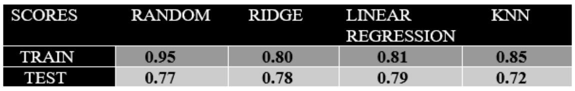
Ridge regression is an example of a shrinkage method: it shrinks the parameter estimates in the hopes of reducing uncertainty, increasing prediction accuracy, and aiding interpretation relative to the least squares.

Random Forest: Random forest is a machine learning algorithm which is a combined effect of classification and regression and other tasks which operate by erection of decision trees at training time and outputs the class that is the mode of the classes or mean value of individual trees.

V. FLOWCHART



VI. RESULT ANALYSIS



SCORES	RANDOM	RIDGE	LINEAR REGRESSION	KNN
TRAIN	0.95	0.80	0.81	0.85
TEST	0.77	0.78	0.79	0.72

Table: Comparision between Models

VII. CONCLUSION

The main goal of this work is to create a Machine Learning model which could be used by students who want to pursue their education in the US. Many machine learning algorithms were utilized for this research. Linear Regression model compared to other ones. Students can use the model to assess their chances of getting admission into a particular university with an average accuracy of 79 percent. A GUI was developed to make the program, from a non-technical perspective, usable and user-friendly. Using node-red the user interface was developed. The ultimate goal of research will be accomplished successfully, as the system allows students to save the lot of time and money that they would spend on educational mentors and application fees for colleges where they have less chances of getting admissions. The main limitation of this research is we developed models based solely on data from Indian Students studying Masters in Computer Science in the United States, we considered only few universities with different rankings. More information relating to new colleges and courses can be added to the curriculum in the future. The system may also be modified to a web-based application by making node-red modifications. To solve the problem, it is possible to test other classification algorithms if they have high accuracy score than the current algorithm, the framework can be easily modified to support the new algorithm by changing the server code in the Node Red. Finally students can have an open source machine Learning model which will help the students to know their chance of admission into a particular university with high accuracy.

REFERENCES

1. C. Haythorhwaithe, M. de Laat, and S. Dawson, Introduction to the special issue on the learning analytics. American Behavioral Science,57(10):1371-1379,2013.
2. Liu Jinpeng. Research on the application of Data Mining Technology in Analysis of Examinee Wish, Henan University,2009.
3. Alpaydin,E. Introduction to Machine Learning,3^{rd}ed;MIT press:Cambridge,MIT, USA,2010.
4. Kuncheva, LL combining pattern classifiers: Methods and Algorithms, 2^{nd} ed; McGraw hill;John wiley&sons,Inc:Hoboken,NJ,USA,2014.
5. D.M Blei, A.Y. Ng, and M. I. Jordan, Latent Dirichlet allocation,Journal of Machine Learning Research,3:993-1022, 2003.
6. L. Breiman, Accuracy Predictors, Machine Learning, 24(2):123-140,1996.
7. Data Cleaning and Analytics, Machine Learning https://archieve.ics.uci.edu/ml/index.php
8. Data Visualizaton, Machine Learning https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/
9. Jupyter Notebook, Implementing the Algorithms, Machine Learning, https://jupyter-notebook.readthedocs.io/en/stable/

AUTHORS PROFILE

Chithra Apoorva D.A Pursuing PhD in Artificial Intelligence, Assistant Professor in the Department of Computer Science and Engineering, GITAM School of Technology, Bengaluru. With 4 years of teaching experience , published 6 papers in international journals which has high impact factors. Awarded "Best Paper Award" from International Conference on Communication and Computing. Underwent 4 technical trainings and attended 3 workshops, 3 Faculty development program and 1 seminar.

Malepati ChanduNath I am currently pursuing B.tech final year in the stream of computer science and engineering in Gitam University. I studied Intermediate in Sri Chaitanya Junior College and I completed my schooling in Sri Chaitanya EM School. I participated spell bee competetions in School zone level and some Coding challenges like hackathon and Technical quizzes.

Peta Rohith I am currently pursuing B.tech IV year in the stream of computer science in Gitam University. I studied my SSC in RR high school and Intermediate in Sri chaitanya junior college. Participated in all India essay writings and some coding challenges and college level technical quizzes.

Bindushree.S currently pursuing B.tech IV year in the stream of computer science in Gitam University. I completed my SSLC in Amara jyothi School and Intermediate in Saint Francis de sales pu college. I participated in college level technical competitions and workshops like sakrobotix and I have done mini project based on iot (Automated Toll Tax Collection using RFID tag).

Swaroop.S I am Currently pursuing B tech final year in the stream of computer science and engineering in gitam University.I studied my SSLC in MVM central school and completed pu in Devari urs pu College . I participated in taluk level science project expo and completed work shop on automation anywhere.