International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Recommendation System for Higher Studies using Machine Learning

Alcina Judy[1], Kesha D'cruz[1], Janhavi Kathe[1], Kirti Motwani[2]

[1]*U. G. Student, Department of Computer Engineering, Xavier Institute of Engineering, Mumbai- 400016*
[2]*Professor, Dept. of Computer Engineering, Xavier Institute of Engineering, Mumbai- 400016*

Abstract - "Recommendation System for college" is a system framework that will suggest universities to the users of the system who are searching for universities to seek after education for their higher studies and recommends the top universities to the students or users of the system. The users need to fill the registration form to get the login details and this system will suggest universities. The main parameters used to recommend universities are CGPA percentage, University ratings etc. This System utilizes machine learning algorithms and compare accuracy among algorithms.

Keywords—*Higher Study Abroad, Recommender System, Data Mining, Naïve Bayes, K-means*

1. INTRODUCTION

The system gathers the data with respect to items. They assemble inclinations and profiles and break down the equivalent to encourage the user to settle on right choices with respect to items, individuals, approaches, and administrations. As for a long time, the availability of electronic and page is developing quick, analysts are depending more on substance to separate the imperative data for better suggestions. Thus, recommendation systems got well known in helping various dynamic settings. Since there are numerous alternatives for universities the graduates need to invest a ton of energy for investigating the subtleties and they may not do it in an appropriate manner. The graduates need a system that acknowledges their preferences and prescribes the correct university. University determination is one among the issues that the graduates will in general get confused. This system will help the graduate students to choose in which universities they should consider.

1.1 Problem Statement

The college recommendation system assumes a significant role as university choice requires a great deal of study of the quantity of variables according to the future perspective. Looking for a decent university is a tough job for a graduate who needs to seek after his/her higher studies. Graduates look for different angles like university campus, teaching staffs, extracurricular exercises in universities, foundation of universities, and so on, even the surveys of university is looked to get additional affirmation about the subtleties. College Recommendation System will prescribe great universities to a graduate depending on his/her decision of field, top evaluations, area and past rate. It is significant as it diminishes the manual work and robotizes this with the assistance of software.

1.2 SCOPE OF THE PROJECT

This system will be recommend colleges and that will predict chances of admit to the students to pursue their higher studies so that they will get clear idea where they can do progress or which factor like GRE or TOEFL or grade point is imp. Even other factors are also imp and that should also be considered as research paper or internships. And this system will be useful for final year students.

System will be trained by the dataset that we are using and we will try to get accuracy as much as possible by comparison of the number of the algorithms K-Means, Naïve Bayes, KNN, regression, SVM etc. and comparing the accuracy of the algorithms and to check which work better for dataset.

The systems available do not consider a lot of parameters so this system can easily help students for recommended colleges and check where they can stand. By the use of different machine learning algorithms to come up with a good recommendation system.

2. EXISTING WORK

The way toward getting chance of higher studies with full financing is methodical just as serious. Heaps of graduates apply various colleges of various nations for their graduate studies with their scholastic profile and systematize test scores, for example, GRE, TOEFL, and

IELTS. Universities offer admission to appropriate applicants depending on their scholastic profile, test scores, job and research. University selection is the most critical for applying to higher studies. The information obtained from the database of graduates will be adequate to discover answers to such inquiries as: Which elements decide the financing opportunity for the graduates to a specific university? What categories of graduates usually get full fund in M.Sc. or PhD in a university? Which key variables are important to accomplish subsidizing in higher studies in the wake of choosing fitting university? Information mining procedures are particularly helpful to find such sort of concealed information from the compound data types. USA is one of the most educated country worldwide .So in the wake of finishing graduate studies heaps of gradates attempt to seek higher studies abroad. Some of them succeed and get entrance into their ideal courses in ideal universities. A non-benefit management gathers those graduate's information and structures a general database so that alternative graduates get advantage from that. The principle target of this investigation study is to construct and build up a recommender system for graduate admission seekers which can help them to determine university coordinating their whole profile utilizing scholarly information of graduates who have just got the chance to seek higher studies abroad.

Recent papers used algorithms Naive Bayes and Decision Trees which can be useful to solve the given problem. The algorithm which is expected to have higher accuracy in recommending the best college is used. Thus technique would be helpful for students minimizing their time in searching colleges. Our system consists of in all four modules which describe the various aspects for recommending colleges, details of the same, branches and comparison with other colleges. A. Our Application comprises of four modules

• **College Search:** In this module user will give name of college and location by which he has to search as input. Accordingly a list of colleges will be displayed to the user.
• **Comparison:** In this module user will be given a choice to select colleges he wants to compare. By this he will get a clearer idea for the distinguished college.
• **Advanced Search:** We have collected interest rating from a survey using google form. Interest fields includes attributes like infrastructure, cultural and technical activities, sports, NSS, Edc and other attributes include faculty, hostel, placement and fees. Based on the feedback given by the students of various colleges we have averaged their ratings to get a mean value. In this module user will be asked to give his candidature details and his interests in cocurricular and extra curricular activities. According to his merit and his interests, colleges will be shortlisted based on their merit, interest, fees and locality.

• **Branch Search:** In this module user will be asked to give his candidature details and the college name. According to his merit a list of branches will be recommended with the student is most likely to get according to merit in that college.

Above work WEKA tool is used for comparing algorithms based on inputs. There are 160 entries as input which include cut off of 5 colleges namely COEP, VIT, PICT, VIIT and SCOE. We considered 6 attributes for classification including College, Branch, Gender, University, Caste, Merit Number. After giving training data we passed testing data for classification. Here we took college's names as class labels for classification. And we found that Naive Bayes gave an accuracy of 50 percent that means it classified 5 out of 10 colleges correctly. While for J48 we found the accuracy to be 80 percent. The text around the Confusion Matrix is arranged so as row labels are on the left instead on the right but we read it just the same. The row indicates the true

2.1 PROFILE BASED

The profile based system will display the list of universities will be displayed on the basis of the student's GRE verbal score, GRE quants score TOEFL score and CGPA. It will calculate the average and that average will be converted into the range. So first we will calculate average for each score that is GRE verbal, GRE quants and TOEFL score. So the formula for GRE verbal and GRE quants is (1),

$$GREV/GREQ= ((score-130)*100)/40 + 0$$

So 130 is the minimum score of GRE verbal or GRE quants, 100 is the maximum range , 0 is the minimum range, score is the value of GRE verbal or GRE quants and 40 is the difference between the maximum and the minimum score that is 170 and 130 respectively.

Same goes for TOEFL score is (2),

$$TOEFL=((score -0)*100)/120 + 20 \quad (2)$$

So here 0 is the minimum score , 100 is the maximum range, 20 is the minimum range, score is the TOEFL

score of the student and 120 is the difference between maximum score and minimum score that is 120 and 0 respectively. Now after calculating average of each score we can now obtain the range by formula given below is (3),

Range=(GREQ*2+GREV*1.5+TOEFL*0.4+percent*1)/3.9

So here 2, 1.5, 0.4 and 1 are the weights which are assumed. More weightage is given to GRE quants because normally universities gives first priority to GRE quants score. 3.9 is the addition of all the weights assumed . Now this range will be compared to the ranges of the universities. On the basis of that the list of universities will be listed with the user and system ratings.

3. PROPOSED DESIGN

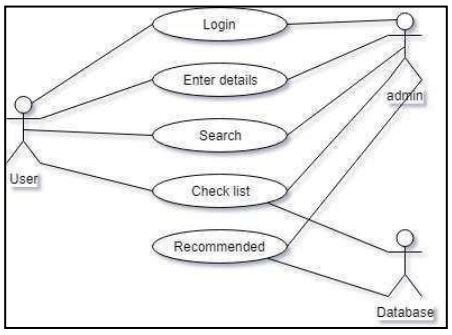

Fig.1 Use Case Diagram

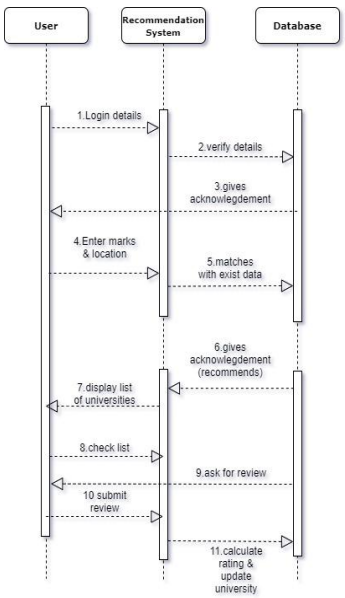

Fig.2 Sequence Diagram

User of the framework is executing login procedure and afterward in the wake of entering exceedingly significant subtleties user can check that he/she is qualified or not for getting affirmation for M.S.

The user at that point examines for universities and afterward that user list of the universities on the basis of the details he/she filled in the form and ratings and reviews given by other users.

4. ALGORITHMS USED

1. Collaborative Filtering

More weightage is given to GRE quants because normally universities gives first priority to GRE quants score. On the basis of that the list of universities will be listed with the user and system ratings.

Steps:

1) Users similarity calculation :It uses Pearson correlation, cosine similarity and Euclidean distance
2) Top N nearest neighbors' selection and
3) Prediction.

So, in collaborative based the list of universities will be displayed on the basis of student's GRE total score and TOEFL score. It will compare the cut-off of the universities and on the basis of that the list of universities will be listed with the user and system ratings.

2. Content Based Filtering

Recommendation to user is given only by the users individual behavior and data. First it analyses the description of items preferred by user to decide the preferences that can be utilized to describe these items. based on users' choices user profile is created, next each item attribute is compared with user profile so that only related items are recommended to the user. So in collaborative based the list of universities will be displayed on the basis of student's GRE verbal score , GRE quants score and TOEFL. It will compare the cut off of the universities and on the basis of that the list of universities will be listed with the user and system ratings.

3. Similarity And Distance

[1] Pearson correlation

Pearson relationship quantifies the straight relationship between persistent factors. (alluded to as ρ).The unique formula for correlation, created by Pearson himself, utilizes unanalyzed

$$\rho_{X,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad (4)$$

Formula for two factors, X and Y is in equation (4)

Raw inspections are focused by subtracting their means and re-scaled by an estimate of standard deviations. An alternate

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (5)$$

Method to show a similar amount is as far as presume values, means μX, μY, and standard deviations σX, σY in equation (5)

Notice that the numerator of this division is much the same as above meaning of covariance, since mean and desire are frequently utilized reciprocally. Separating the covariance between two factors by the result of standard deviations guarantees that correlation will consistently fall between - 1 and 1. This makes deciphering the correlation coefficient a lot simpler. Let the set of items evaluated by the two users and be meant by I, at that point similarity coefficient between them is determined as in equation (6)

Here i signifies the rating of user u for item i , and ru is the mean of all items given by user u. Additionally, rv ,i signifies the rating of user v for item i , and ru is the mean of

$$sim(u,v) = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{v,i} - \bar{r}_v)^2}} \quad (6)$$

items given by user v .Let the set of items judged by the two users and be meant by I, at that point closeness coefficient between them is decided as in equation (7)

$$sim(u,v) = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{v,i} - \bar{r}_v)^2}} \quad (7)$$

[2] Cosine Similarity

The cosine distance between two points is the edge that the vectors to those points make. This point will be inside the range 0 to 180 degrees, regardless of what rate measurements the space has.

$$s(u,v) = \frac{r_u \cdot r_v}{\| r_u \|^2 \| r_v \|^2} = \frac{\sum_i r_{u,i} r_{v,i}}{\sum_i r_{u,i}^2 \sum_i r_{v,i}^2} \quad (8)$$

The similarity sim(u,v) between user u and v is calculated as shown in equation (8)

4. K-MEANS :-

Choose a number of clusters "K"
Randomly assign each point to Cluster
Until cluster stop changing, repeat the following
1. For each cluster, compute the centroid of the cluster by taking the mean vector of the points in the cluster.
2.Assign each data point to the cluster for which the centroid is closest.
The K-Means algorithm can cluster observed data. But how many clusters (k) are there is sloved by elbow method finds that finds optimal value for k (#clusters).

The technique to determine K, the number of clusters, is called the "Elbow method".

K-Means can be used for deciding the number of clusters by using elbow method.

5. Naive Bayes classifier

Bayes' Theorem:

- Bayes' theorem is also known as Bayes' Rule or Bayes' law, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Working of Naïve Bayes' Classifier:

Python Implementation of the Naïve Bayes algorithm:

Now we will implement a Naive Bayes Algorithm using Python. So for this, we will use the "user_data" dataset, which we have used in our other classification model.

Therefore we can easily compare the Naive Bayes model with the other models.

Steps to implement:

- Data Pre-processing step
- Fitting Naive Bayes to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.

5. SPLITTING OF DATA SET

To check how algorithm works the CSV file that is dataset is split into train and test data in python machine learning.

We can use supervised machine learning algorithm to split data around 30% to 70% between testing and training stages. We used pandas for importing the dataset that is CSV file and sklearn for splitting the dataset. Here prediction is on the basis of GRE score, TOEFL score and CGPA.

The trained data is fitted in the model to make predictions by using any machine algorithm. We can use hybrid concept that is more than one algorithm to get better accuracy for the system. So we used algorithm as k-means, naive bayes machine learning algorithms and can briefly study accuracy with another algorithms as linear regression (which is used to solve regression problems), SVM(which is used to solve classification or regression problems) and KNN(which is used to solve both classification problems and regression problems). The most accurate algorithm amongst the three is KNN for this recommendation system. This system will the recommend top 10 universities to the users.

6. DATA ANALYSIS

Personal and scholarly data of a particular graduate is reserved in the general table. They are gathered for the information preprocessing and information observation

Most important elements of the database are CGPA, GRE Score, TOEFL Score, University Rating, University Name
The personal and scholastic information are considered for information analysis with respect to scholarly profile of graduates and considered ratings given by other users.

7. IMPLEMENTATION & RESULTS

By using python (jupyter notebook) algorithms are implemented and output is shown in the following figures.

7.1 K-Means

The elbow method shows the number of the clusters are 3 as shown in the figure 2.

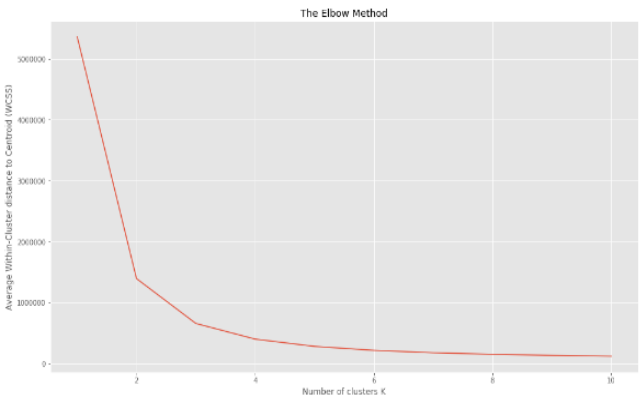

Fig.3 Elbow Method

Clusters are formed by using k-Means and min number of clusters are 3 that are shown in the graph in figure 3.

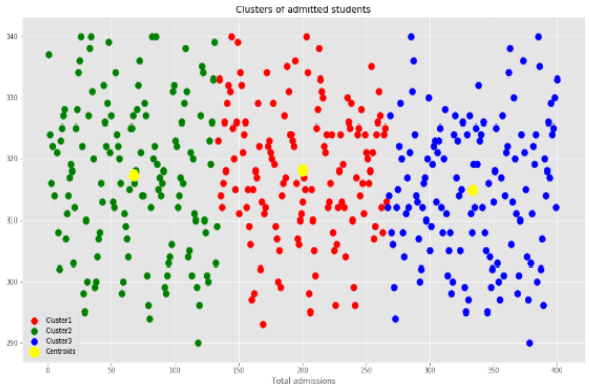

Fig.4 Clusters of admitted students

7.2 Naïve Bayes classifier

```
# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test,pred)
print(cm)
```

```
[[  2  65]
 [  6 127]]
```

Fig.5 Confusion Matrix

Confusion matrix shows 124 are correctly predicted and 71 are not correctly predicted.

```
In [25]: print(classification_report(y_test, pred))
```

	precision	recall	f1-score	support
no	0.00	0.00	0.00	58
yes	0.64	1.00	0.78	102
accuracy			0.64	160
macro avg	0.32	0.50	0.39	160
weighted avg	0.41	0.64	0.50	160

Fig.6 Classification Report

7.3 Content based filtering

By using tkinter and listbox output is shown for content based filtering. Top 20 universities displayed by using List box.

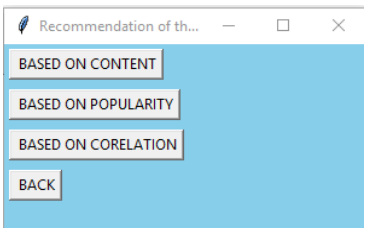

Fig.7 Tkinter buttons

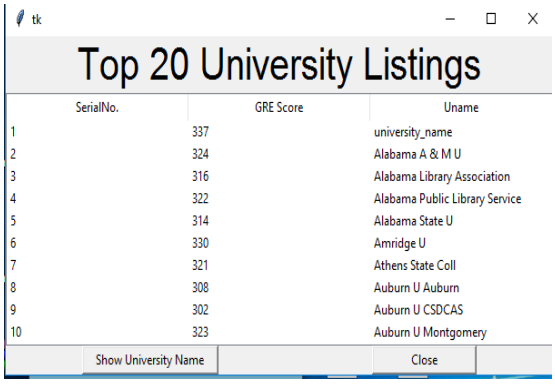

Fig.8 University List

8. CONCLUSION

We have applied KNN, SVM and linear regression on the attributes such as CGPA, GRE, and TOEFL. KNN gives preference to GRE score, SVM gives preference to CGPA and Linear Regression gives preference to TOEFL score. Hence the conclusion is that in this recommendation system KNN gives the highest accuracy that is 91% as compare to SVM and linear regression which is 81% and 83% respectively by comparing the graphs of each model.

We have studied new features of python language as list box, tables, graphs, TKINTER and animation.

9. FUTURE SCOPE

1. In future we will use rule base classification on the basis of GRE verbal score, GRE quants score, TOEFL score and CGPA.
2. We will consider more parameters for this system like limit of the fees of the user, choice of the location of the, job experience.
3. We will use sematic analysis on the reviews of the other users for the recommendation system.
4. This project can be hosted as a weblink in any college.

REFERENCES

[1] A Collaborative Filtering Based Library Book Recommendation System, by Chaloemphon Sirikayon, Panita Thusaranon, Piyalak Pongtawevirat,978-1-5386-5254-1/18/$31.00©2018 IEEE

[2] Graduate School Recommender System: Assisting Admission Seekers to Apply for Graduate Studies in Appropriate Graduate

Schools, by Mahamudul Hasan, Shibbir Ahmed, Deen Md.Abdullah, and Md.Shamimur Rahman, 978-1-5090-1269-5/16/$31.00 ©2016 IEEE

[3] An Improved Approach For Movie Recommendation System, by Shreya Agrawal (ME Student) and Pooja Jain (Assistant Professor),978-1-5090-3243-3/17/$31.00©2017 IEEE

[4] For training dataset- https://data-flair.training/blogs/train-test-set-in-python-ml/