**University Recommendation System**

Manish Katturu

<manish.katturu[@sjsu.edu](mailto:nagajanaki.dwadasi@sjsu.edu)>

Student Id: 013825201

Department of Computer Engineering

San Jose State University

San Jose, CA

**TABLE OF CONTENTS**

1. **Abstract**
2. **Introduction**
3. **Real – world Problem**
4. **Literature review**
5. **Work-flow Diagram**
6. **Data Collection and visualization**
7. **Data Cleansing and Transformation**
8. **Feature Extraction**
9. **Recommendation Models**
   1. **K-nearest neighbors**
   2. **Random Forest**

1. **Results**
2. **Conclusion**
3. **References**
4. **Abstract:**

For students applying to higher studies like Graduate programs or PHD programs, shortlisting the universities has become a difficult task. Many students get puzzled while thinking whether their profile is good enough to find an admit at a particular university. As a result, many students end up making wrong choices. So, in order to provide a better university recommendation, I have built a recommendation system with the help of various classification algorithms considering various criteria of the student and the university. Using the university profiles dataset and the student profiles data set, several models were trained which suggests a better university matching the student’s profile which simplifies the work of aspiring students.

1. **Introduction:**

As Over the time, the number of enrollments for the graduate programs is constantly increasing, the competition for getting admission to a better university increases year by year due to which shortlisting the universities by an aspiring student becomes more difficult. Given the growth of new programs and number of admissions, a student is often unaware of the existence of such programs. In this paper, a justifying attempt using K-Nearest neighbors, Random Forest was made to provide a solution to these issues by considering the target university’s perspective to evaluate whether a student’s profile is competitive enough to be admitted into their university. Hence the students could get a better picture of where they stand and can make an intelligent well-formed decision. As a first step, information regarding 45 universities were collected along with the details about students’ profile and their admission results. Section 2 explains about the problem that has been addressed and the approach to solve it. Section 3 includes a brief description about the existing literature on similar topics outlining their approach, techniques adopted, pros and cons analysis, results and conclusions. Section 4 describes the data set, populated by scraping and data cleansing and transformation. It also gives detailed explanation about the selection of relevant features and their impact on the model. Section 5 gives information about various models and a comparison between each other in terms of accuracy. Finally, the performance of the chosen models are analyzed using the results obtained and a summary are included under Section 6 and Section 7.

1. **Real – world Problem:**

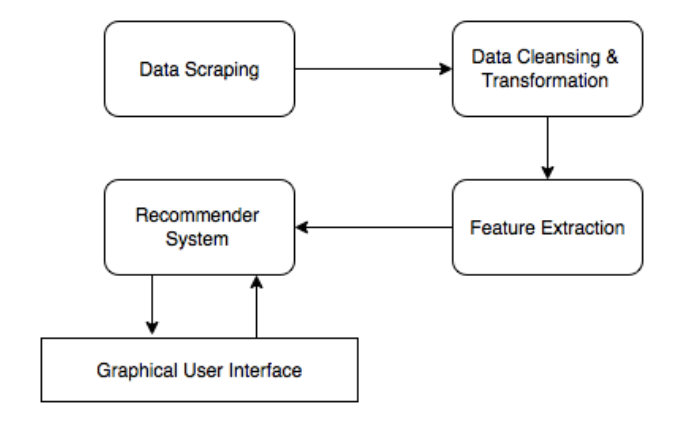
In today’s fast-paced world, every technological innovation influences the importance of higher education, especially the ones which serve as hubs to the latest researches and trends. Given that, the United States will be one of the top destinations for any student across the world. For international students who wish to pursue graduate studies in the United States of America, choosing a suitable college and earning an admit is a challenge. Although, many internet resources and forums are available, they do not offer satisfactory suggestions, as most of them are based on assumptions from college rankings and not the actual statistical relations. From a student’s point of view, the cost of the application and the amount of dedication to the process is also high. Thus, to guide the students in an efficient manner, the university recommender system has been developed, based on the input of the students’ academic data. Since the problem is extensive, for the sake of simplicity, a select list of 45 universities were considered.

1. **Literature review:**

In the past, a lot of work on employing data mining techniques in the field of education were undertaken. Few recommender systems to suggest course and university based on

a student’s academic record were developed. Those systems employed decision tree classifier and fuzzy c-means clustering techniques using WEKA tool kit and it was aimed to help the students choose a stream which will suit their skill setsAnother different recommender system was built to help the students with their academic itineraries. They help in making decisions about what course to select based on a student’s schedule, stream and professors. Here, the model was trained based on past 7 years data for a particular university and classifiers for every subject was modeled based on cumulative GPA. On the other hand, some recommender systems were modeled to help the university to know about their students by keeping track of their time, extracurricular activities and achievements, in addition to their academic potential. This helps them to identify and categorize the students depending on the need using two-step algorithm and K-means. However, there was no access to any of the data-set used in the above mentioned works. Although similarities exist with the topic considered in this paper, it is not appropriate to compare results directly with any previous work because the data set used in this paper is completely different

1. **Work-flow Diagram:**

****

1. **Data Collection and visualization:**

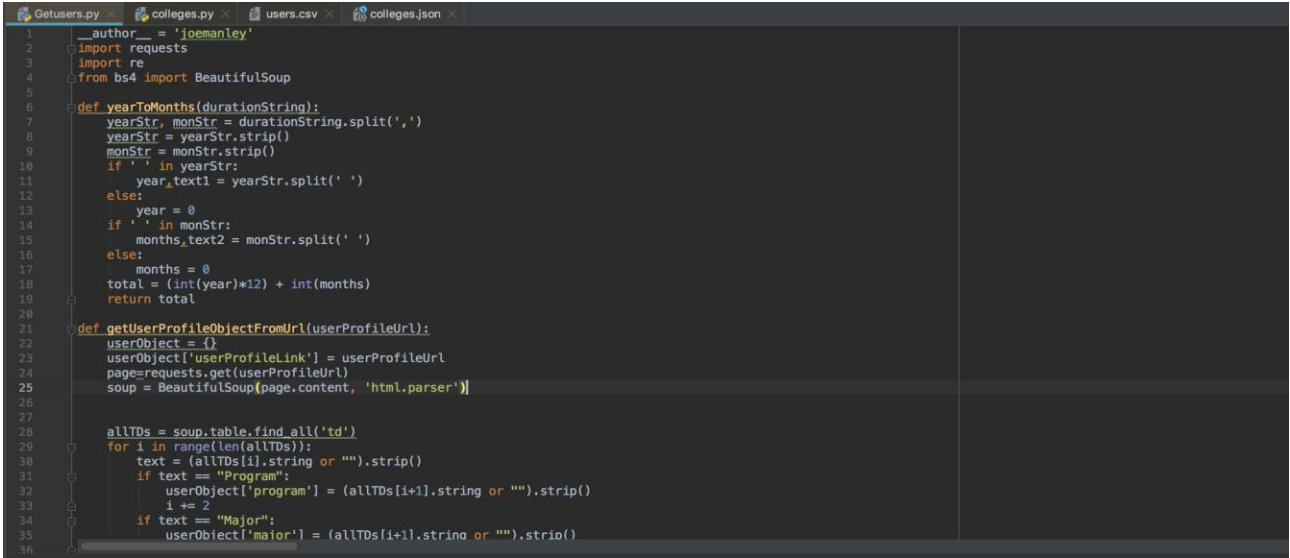
Initially the list of 45 universities was narrowed down, which had enough data to be scraped. Universities with skewed data were dropped down. Then a spider was built to get the list of students and the links to their profiles on Edulix. Once the unique set of students was identified, the data was scraped from each profile by using ’BeautifulSoup’. The tabular structure of Edulix’s web page, helped to identify the required data labels and points.

The following are the python code for scrapping the data and the corresponding output:

* The below two screenshots shows the python code for getting the users data containing

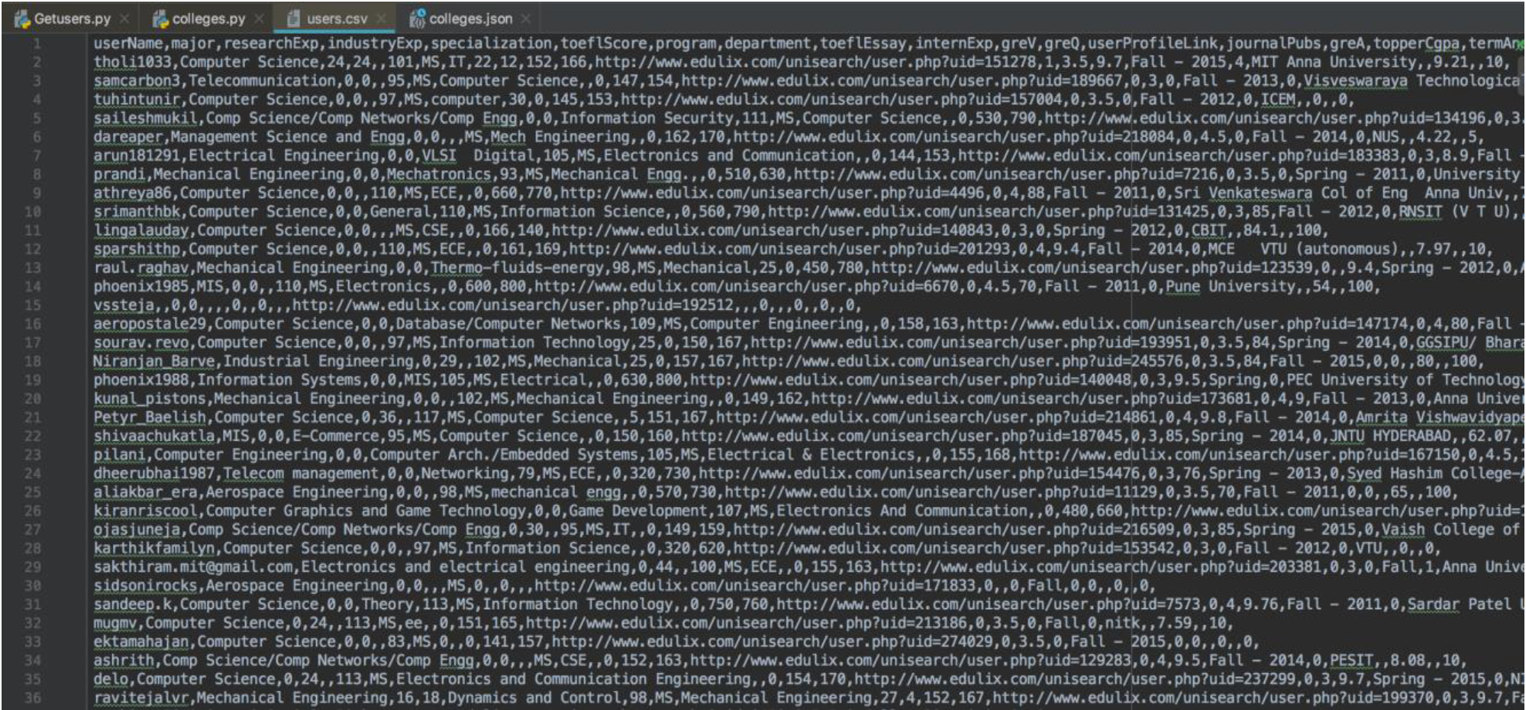
the attributes – username, major, research exp, industry exp, specialisation, TOEFL,

graduate program, department, GREV, GREQ etc.

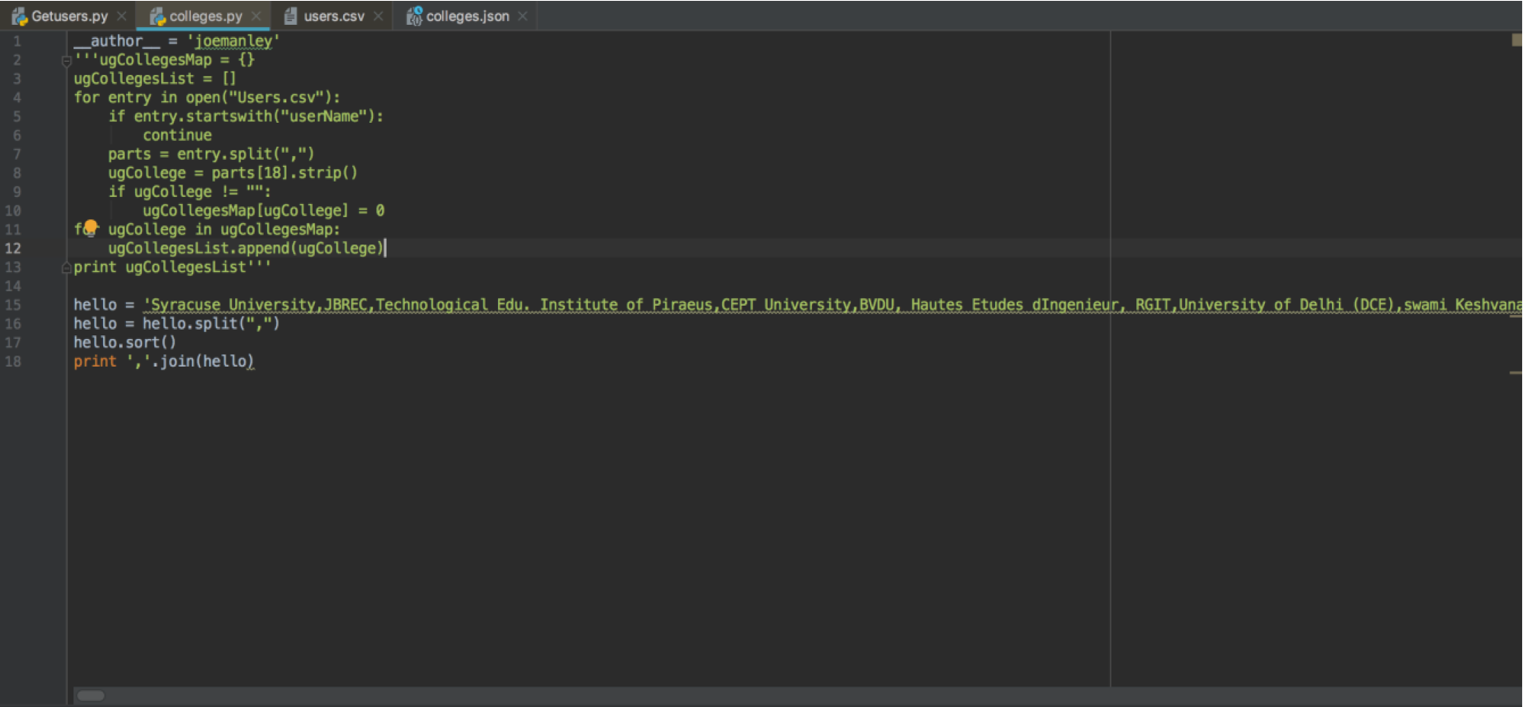


****

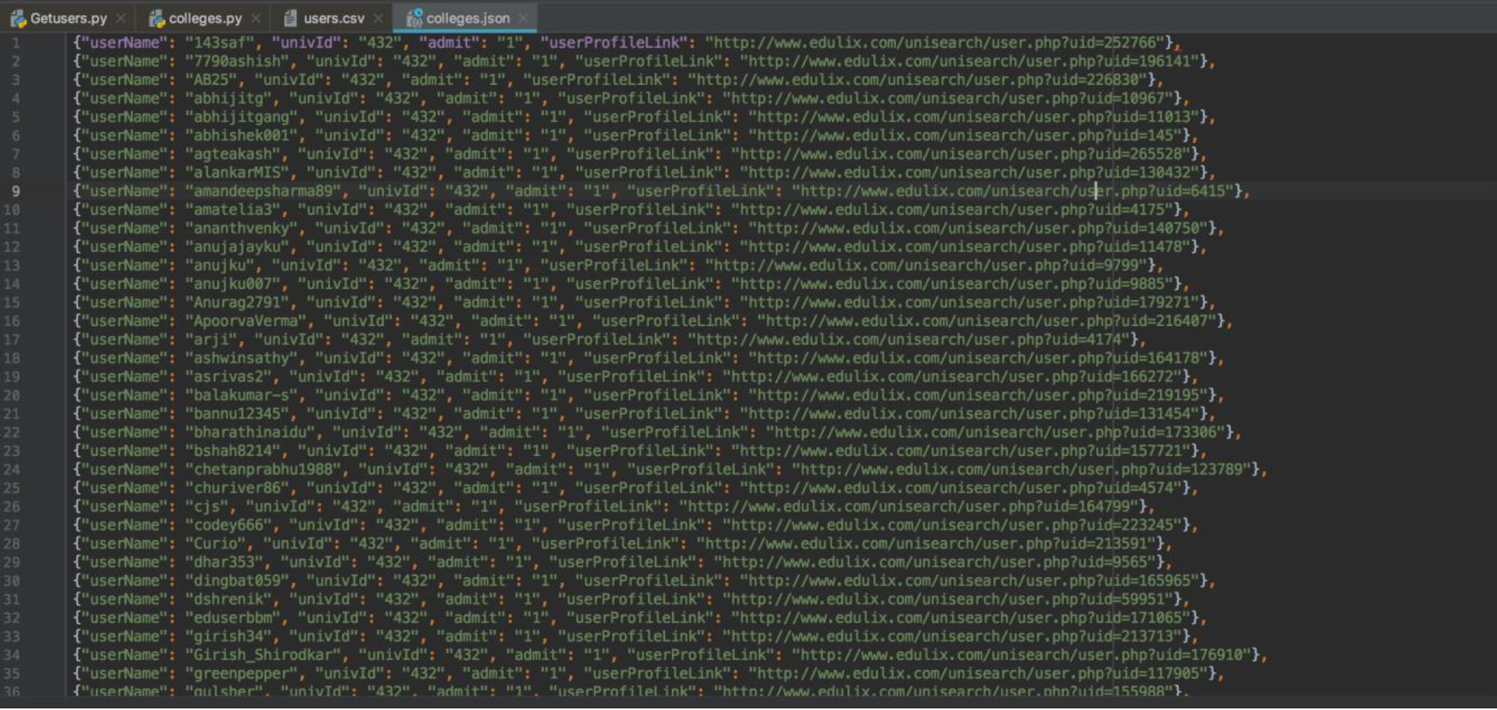
* The below screenshot is the CSV file containing the users extracted data



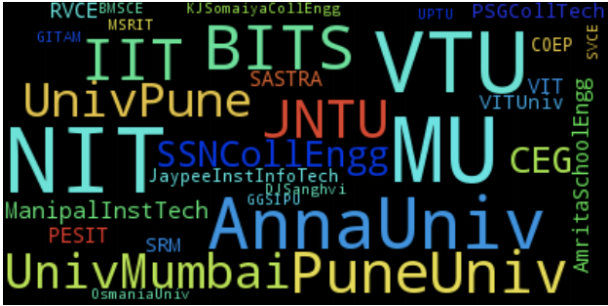
* The below screenshot shows python code for the extracting the colleges data(I.e. those who got admitted previously) separately from the users.csv:



The following is the colleges data in json:



The first step in building any recommendation system is the identification of the data set. For this particular problem, academic details and background information which are provided during the application process, forms the core data. In order to build the classification model for the recommender system, this data has to be organized with appropriate labels. This core data for the application process is not readily available on the internet for direct consumption. Though there were few forums which had some vital information regarding the same in terms of scores, the distinguishing information regarding the students’ research interest and knowledge in a particular topic remains unknown. However, this whole approach is based on making maximum use of the available information. Variation of the number of admissions to any graduate program based on undergraduate universities is represented by Figure 2. It was found that Mumbai University(1587), National Institute of Technology(1467), Visvesvaraya Technological University(1426) and Anna University(1032) were some of the undergraduate universities with highest number of admits

****

**visualization of some of the undergraduate universities**

The ’Edulix’ forum is one of the most popular forums for students planning to pursue graduate studies. This is the hub for students who wish to take part in discussions and queries regarding any information about graduate studies. This forum basically collects the academic details of its users to evaluate their profile against past experiences. Out of all these data, some data like the candidate’s undergraduate university, CGPA, GRE and TOEFL scores, number of research publications, work experience etc. were identified as prospective features. By writing a web crawler script, relevant data necessary for this model was scraped off from their website, cleaned and then transformed into appropriate forms to be used as input data for the models.

The columns present in the dataset and the information provided by them is provided in the table below:

Various Columns present in the user-dataset and corresponding information provided by them.

|  |  |
| --- | --- |
| **Column Name** | **Information** |
| UserName | Provides the name of the user |
| Gre\_Verbal | provides the verbal score in GRE |
| Gre\_Quant | provides the Quant score in GRE |
| ToeflScore | provide the TOEFL score |
| Gre\_AWA | provides the AWA score in GRE |
| CGPA | provides the CGPA of the user |
| Industry\_exp | provides the number of years of experience |
| major | Major in which the user wants to pursue |
| UGcollege | provides the UG college of the user |
| Internship\_experience | Provides the internship years of experience |

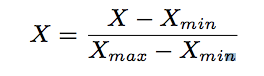
The Universities data set is prepared by first finding all the users who are admitted into universities and taking the students belonging to each university and averaging their scores in order to find the data for each university. Finally, all these records are combined together to form a university data set.

1. **Data Cleansing and Transformation:**

About 800 samples of raw data was obtained by as a result of scraping. Each sample corresponds to the profile of a student. The data points extracted included GPA, undergraduate university, GRE verbal score, GRE quantitative score, GRE analytical writing score, industry experience, research experience, internship experience and pursuing major. Cleansing the data of undergraduate universities had to be done, since this field was just a text box and not a select field. So input from different students created anomalies and this was corrected by trimming the string and removing spaces found in them. The GRE scores(Verbal, Quantitative and AWA) were also cleansed since they contained scores of both old and new versions of the examination. Similarly the GPA scores available were based on different point systems, so all the GPA scores were uniformly scaled to 4 point scale. Also, certain categorical features like the student’s undergraduate university and department to which they apply were considered as separate features. A total of 1435 distinct undergraduate universities and 53 distinct majors were obtained after filtering and each of these were used as binary features.

1. **Feature Extraction:**

The most important property of a feature is its correlation with the predicted output. Exploratory analysis was done by plotting the feature values for two different universities and observing their variation. Variation of features CGPA and GRE for two different universities(Purdue and NJIT), has been shown in Figures below. Initially, when all the features in the data set were considered the accuracy was comparatively low(40%). The forward selection algorithm was used to select the best set of features for the model. In the first iteration of the algorithm, the single best feature was identified that best describes the variance in the data. In the second iteration, the best feature was fixed and the next best feature was found. This process was repeated till the accuracy no longer improved. Based on this method, undergraduate university, research experience, GRE and GPA were found to be the most effective features. After using forward selection algorithm, the accuracy improved. During this process, a situation arose, when the accuracy did not show any improvement, even though the best features were chosen. This was because, the numerical features like CGPA and GRE score were based on different scales, and so had an adverse implication on the model. However, when scaled from 0 to 1, there was a significant improvement in the accuracy. Hence, all the numerical variables were then normalized to a scale of 0 to 1 by using the following formula

****

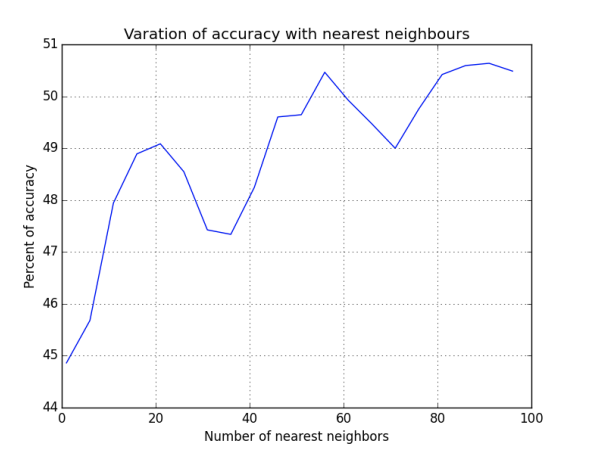
where X is value of any feature.

1. **Recommendation Models:**

The baseline model is one in which it randomly predict 10 universities out of a total of 45 universities for each user. The accuracy of this model was found to be 22%. Two different models, K-Nearest Neighbors and Random Forest, were built using a combination of all the features mentioned above, to classify a student profile to the best university that they must apply to, among the available 45 universities. Once the best university was found for the student, the 9 most similar universities in terms of the selected features was found by computing Euclidean distances to give a total of 10 universities, that the student must apply to. The data was split as 80:20 for training and testing. The model classified the training data with good accuracy but had a high error rate for test data. This problem was due to over-fitting and can be avoided by techniques like Cross validation to test the model on more datasets or by techniques like Principal Component Analysis to reduce the dimension (number of features used) of the model or more datasets can be used. The first technique has been employed in this project. k-fold cross validation mainly prevents overfitting as it reduces the variance by averaging over k different partitions, so the performance estimate is less sensitive to the partitioning of the data4. The entire data set is divided into 5 sets and each time 4 sets are used as the training data and the model is tested on the remaining 1 set which is used as the test data. The accuracy of the model is determined and this process is repeated 5 times. Each time a different set is used as the test data. The error rates obtained are all averaged to obtain the final error rate. The following subsections describe the models that we tried

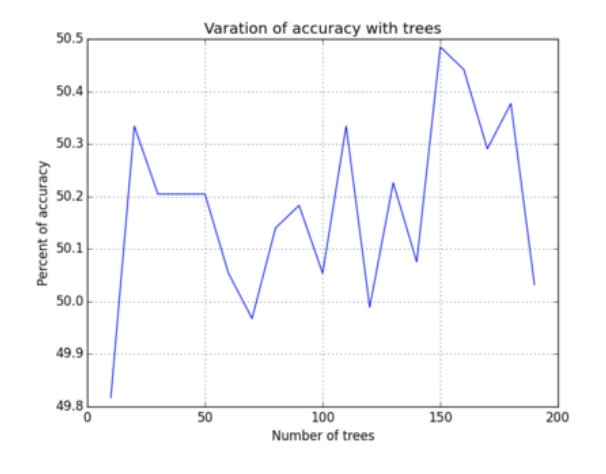
* 1. **K-nearest neighbors:**

K-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. In K-NN classification, the output is a class membership. An object is classified by majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. The K-NN model was run by varying the number of neighbors that were used and it was found that the best accuracy of 50.6% was obtained when the number of neighbors was equal to 56. The variation of accuracy with the number of trees constructed is shown in figure below

****

* 1. **Random Forest:**

Random Forest is an ensemble of decision trees. Unlike single decision trees which are likely to suffer from high Variance or high bias (depending on how they are tuned). Random Forests use averaging to find a natural balance between the two extremes. Since they have very few parameters to tune and can be used quite efficiently with default parameter settings (i.e. they are effectively non-parametric). Random Forests are good to use as a first cut when you don’t know the underlying model, or when you need to produce a decent model in a short time. The Random forest model was run by varying the number of trees that were used and it was found that the best accuracy of 50.5% was obtained when the number of trees was equal to 150. The variation of accuracy with the number of trees constructed is shown in figure below:

****

1. **Results:**

In this work, K-Nearest Neighbor, Random Forest were considered for recommending the 10 best universities for aspiring graduate students and their performances are summarized below:

****

Accuracy of the models

The features - Undergraduate university, GPA, GRE Score and Research experience were found to explain the maximum variance in the data and were used to build the final model. The features - number of journal publications, number of conference publications, industry experience, internship experience and pursuing major did not provide any new information about the data and hence did not contribute to the model.

1. **Conclusion:**

Random Forest, K-Nearest Neighbor have been successfully used for the building the university recommendation system. Both the models are found to be accurate. New features like Statement of Purpose, Letter of Recommendation etc. can be analyzed using text mining techniques and could be incorporated if found to improve accuracy. Also, as an extension to this work, recommendation of university with respect to research interest can be made with further study.

1. **References:**
2. Huiyi Tan, Junfei Guo and Yong Li, "E-learning Recommendation System," csse, vol. 5, pp.430-433, 2008 International Conference on Computer Science and Software Engineering, 2008.
3. Osmar R. Zaíane, “Building a Recommender Agent for e-Learning Systems”, Proceedings of the International Conference on Computers in Education, p.55, December 03-06, 2002.
4. Ahmad A. Kardan, Solmaz Abbaspour and Fatemeh Hendijanifard, “A Hybrid Recommender System for E-learning Environments Based on Concept Maps and Collaborative Tagging”, IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
5. Badrul Sarwar, George Karypis, Joseph Konstan and John Riedl, “Analysis of Recommendation Algorithms for E-Commerce“, Proceedings of the 2nd ACM Conference on Electronic Commerce, pp. 158-167, 2000
6. Gediminas Adomavicius and Alexander Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-theArt and Possible Extensions," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, June 2005, doi:10.1109/TKDE.2005.99