College Recommender System using Neural Networks and Data Mining Techniques

Abstract:
This paper proposes a recommendation system for Admission in Colleges, based on GPA, Quants, Verbal and AWA Scores. Sometimes, it can be really quite tricky for a person to decide what college to apply, with his scores. With the help of Neural Networks and Data Mining Techniques, the recommendation system has been developed, so that it reduces the effort of a person to decide what colleges to apply. The Clustering of various colleges has been done using K-means Clustering. The dataset consists of the following: College names, GPA, Quants score, Verbal Score and AWA Score. Even with the given dataset, some pre-processing of data has to be done. The values of K, in the 'K-Means Algorithm' helps indicate the number of tiers that are taken into consideration to search for colleges. MLP (Multi-Layer Perceptron) is then used to predict the probability of a person getting into colleges at a tier level. Based on the probability value and the scores of the individual, we find a number of colleges close to the scores of the person, at a certain tier level. Scores are used as the input for predicting the colleges into which he can get admission. There are a number of colleges that are found for each tier, similar to tier 1. This method provides a simple interface for the user to enter his scores, which then provides a list of colleges he can get admission to. This is a relatively simple and effective implementation for a College recommendation system, compared to others that may or may not be available.

Keywords: Recommender System, Multilayer Perceptron, Data Preprocessing, K-Means Clustering, College Prediction, Nearest Neighbours.

1. Introduction:

For anyone pursuing their postgraduate studies, it would be difficult for them to find out what college they may join, based on their GPA, Quants, Verbal, TOEFL and AWA Scores. People may apply to many colleges that look for candidates with a higher score set, instead of applying to colleges at which they have a chance of getting into. This would be detrimental to their future. It is very important that a candidate should apply to colleges that he/she has a good chance of getting into, instead of applying to colleges that they may never get into.

There aren't many efficient ways to find out the colleges that one can get into, relatively quickly. The Education Based Recommender System helps a person decide what colleges they can apply to with their scores. The dataset that is used for processing consists of the following parameters: College names, Quants and Verbal Scores (GRE) TOEFL and AWA Scores. The GRE Test (Graduate Record Examinations) is a standardized test used by many universities and graduate schools around the world as part of the graduate admissions process.

Other factors are also taken into consideration while applying to colleges, such as Letter of Recommendation (LoR), Statement of Purpose (SoP), Co-curricular activities and Research papers as well (research papers from journals that are not well known or have a high percentage of plagiarism are not taken into consideration for this case).

When a person has completed their undergraduate degree and wants to pursue a Postgraduate degree in a field of their choice, more often than not, it is very confusing for the person to figure out what colleges they should apply to with the scores that they have obtained in GRE and TOEFL, along with their GPA at the time of their graduation. Many candidates may apply to colleges that do not fall under their score requirements and hence waste a lot of time. Applying to many colleges with scores also increases the cost. There are not many efficient methods that are available to help address this issue and hence an Education Recommender System has been developed.

In the system proposed, a person can enter their scores in the respective fields provided. The system then processes the data entered and produces an output of the list of colleges that a person could get into, with their scores. This is relatively quick and helps conserve time and money. In order to achieved this we have proposed a novel method utilising Machine Learning and Data mining algorithms.

2. Literature Review:

The recommendation systems collect the information regarding items. They gather preferences and profiles and analyse the same to advise the user to make right decisions regarding products, people, policies, and services.[1]. As day after day, the availability of electronic and web content is growing fast, researchers are relying more on content to extract the vital information for better recommendations. So, recommendation systems became popular in assisting numerous decision-making contexts. The user is an entity to which the recommendation is provided, and an item is a product/service being is recommended. Recommendation analysis predicts the future preferences by analysing the previous interaction between users and items because past behaviours are often good indicators of future choices. It is the toughest job to design and implement a large-scale online service that can find what is to be recommended to the customers based on the past purchase history. For example, Amazon gives product recommendation, yahoo makes web page recommendation. The process of constructing an efficient and effective recommendations system is a challenging task. The underlying reason is the large size of the product (object) space and context space. The main goal of recommender systems is to assist its users in finding their preferred objects from the large set of available objects. The voting of a particular customer on a particular object is learned through a random payoff and this payoff is received by the recommender system based on the response details of the customer to the recommendation system. Zhibo Wang, et al. [2] proposed a unique similarity-based metric to find the similarity details of users in terms of their lifestyles and they have constructed a Friend book system to recommend friends based on their lifestyles. Recommendation systems have developed in parallel with the web technology J. Bobadilla et al. [3]. At the initial time of their existence, they were based on demographic, content-based and collaborative filtering. Now they are in a position to incorporate social information also. A knowledge-based recommendation system considers user centric requirements rather than his/her past history in order to make recommendations. Hector Nunez, et al. [4] discussed the comparison of different similarity measures for improving the classification process. Authors said that automatic knowledge acquisition and management methods are needed to build consistent, robust, reliable, fault tolerant, and effective decision support systems. Fazeli Soude, et al. [7] said that recommender systems are being used (have been using) in many real-world applications such as e-commerce-based applications – Amazon and eBay. Recommender systems must be accurate and useful to as many numbers of users as possible. The fundamental goal of the educational recommender systems is to satisfy many quality features such as accuracy, usefulness, effectiveness, novelty, completeness, and diversity. Recommender systems must satisfy user-centric requirements. User-centric based recommender systems are more useful than data-centric recommender systems. The methods existing for the recommendation are content-based filtering, collaborative filtering, and rule mining approaches. Content-based filtering approach recommends an item to a user by clustering the items and the user pairs into groups. This clustering is used to gain similarity between user and item. Personal information of the user is not considered here. Queen Esther Booker creates a prototype of a system for course recommendations [10]. The system accepts user requirements as keywords and recommends courses for students. Collaborative filtering (CF) approach recommends an item to a user by grouping similar users based on user profiles and predicts the user interests towards the items. Hana introduces a system based on CF approach to recommend courses for a student by analysing and matching the student's academic records [8]. Then the system analyses and recommends a course that meets the student’s profile. Elham S. Khorasani et al. proposed a Collaborative Filtering model based on Markov Chain to recommend courses based on historical data [7]. Rule mining approach focuses on recommending a series of items to a user by discovering the association rules. Itmazi and Megias developed a recommendation system based on rule mining to recommend learning objects [9]. Akrivi Vlachou, et al. [11] said that finding the most influential database tuples from a given database of tuples is very useful in real-world applications such as market data analysis and decision making. Authors proposed two algorithms for finding most influential database objects. The first one uses properties of the sky-band (SB) set for limiting the maximum number of resultant candidate objects and the second one follows branch and bound (BB) algorithm paradigm and it uses upper bound on influence score. Amit Singh, et al. [14] proposed an approximate solution to answer reverse nearest neighbour queries in high dimensional spaces. Authors said that the approach is mainly based on a feature called strong co-relation between k-nearest neighbour (k-NN) and reverse the nearest neighbour (RNN) in connection with Boolean range query (BRQ). Elke Achtert, et al. [6] said that all the existing generalized reverse k-nearest neighbour (RkNN) search methods are only applicable to Euclidian distances but not for general metric objects. As a result, authors proposed first approach for efficient reverse k-nearest neighbour search in arbitrary metric spaces (RkNNSAMS) and k value will be given at query run time. Duc Thang et al. [5] said that fast, usability, simplicity and with reasonably good performance features are always better than the best performing algorithm only in some cases and rare usage of the algorithm because of high complexity. DINO IENCO, et al. [15] said that the process of clustering data objects containing only categorical attributes is a tedious task because defining a distance value between pairs of categorical attributes is difficult. Authors proposed a framework to find a distance measure between categorical attributes. Madhavi et al. [3] formulated measures on the data containing categorical attributes. They categorized existing measures as context free and context-sensitive measures for categorical data. Usue Mori et al. [12] said that the most famous Euclidian distance and the common measures used for non-temporal data are not always the best methods for finding similarity between time series data because they do not deal with noise and misalignments in the time series data. Authors said that Euclidian distance suffers from noise and outlier problems. Yung-Shen Lin et al. [13] said that similarity measures are being used extensively in text classification and clustering. In the literature, various methods used for similarity comparison are - Euclidian distance, Manhattan distance, taxicab distance, cosine similarity measure, city-block distance, Bray-Curties measure, Jaccard coefficient, extended Jaccard coefficient, Hamming distance, Dice coefficient, IT-Sim and so on. Bouzekri E, et al. [20] in their work the highlight is on engineering issues related to the development of a recommender system in a critical context. They propose a generic architecture to decompose Recommender Systems. This generic architecture integrates existing proposals from the Recommender Systems community with current knowledge in interactive systems architectures. In order to engineer Recommender Systems compliant with the entire list of requirements identified, they propose to use a set of complementary integrated modelbased approaches from the literature. Zhang, et al. [22] go against the classic way of implicitly assuming the underlying classification is binary, that is, a candidate item is either recommended or not. Here they propose an alternate framework that integrates three-way decision and random forests to build recommender systems by considering both misclassification cost and teacher cost. With these costs, a threeway decision model is built, and rational settings for positive and negative threshold values α^* and β^* are computed. Next, they construct a random forest to compute the probability P that a user will like an item. Experimental results show that the (α^*, β^*)-pair determined by threeway decision is optimal. Y. Subba Reddy, et al. [16] proposed a recommendation system for college/course selection. The experimental results showed that applying WCLUSTER in this domain is superior to traditional and previous approaches. To speed up the query execution process a multidimensional indexing structure called R-Tree was used. Deokate Monali, et al. [17] present a framework to choose a desired college using a recommendation system on the basis of college NAAC grade, NBA grade, campus placement and review from alumni student and also add Semantic analysis algorithm which will capture positive and negative sentiments, combining Naive Bayes and Adaboost algorithm which was used to rank the branches as well as colleges. Tulasi K.Paradarami, et al. [18] show that a set of content and collaborative features allows for the development of a neural network model with the goal of minimizing logloss and rating misclassification error using stochastic gradient descent optimization algorithm; and that the hybrid approach is a very promising solution when compared to standalone memory-based collaborative filtering method. K Qazanfari, et al. [19] present a novel text-content-based recommender system as a valuable tool to predict user interests. It developed a specific procedure to create user models and item feature-vectors by soliciting from a user a few keywords and expanding those keywords into a list of weighted near-synonyms, resulting in higher precision and accuracy in comparison to well-known feature-vector-generating methods like Glove and Word2Vec. Kanoje S, et al. [21] address the issue of very less precise and exact data available about Universities/Institutes/Colleges as well as users and hence present a one stop portal where this information could be placed in a systematic manner and can be accessed by the users for better decision making. They have used a novel Profile extraction model to extract data from various web sources. Also, they have profiled users implicitly based on their social networking website. After getting this data we have converted this unstructured data into structured keyword-based profile and have managed to find out change in the ranks of the institutes with respect to user interest, which was analysed by calculating the ranks of each institute with changing criteria weights

When a soft computing model is developed for data clustering, it first calls for an understanding of the soft computing paradigm of computing. Soft computing provides a common platform for genetic algorithms, neural networks and fuzzy logic to come together. This association has been tapped to evolve a method for data clustering. After an introduction to genetic algorithms and fuzzy logic, a literature survey on the data clustering algorithms available has been carried out. Following the survey of conventional algorithms, it is seen how soft computing techniques have, of late, been successfully applied to data clustering problems. A comparison of the various techniques of clustering is done. Though the long term aim of the algorithm is general clustering, the algorithm developed in its present form, is suited only for sequential data clustering. As this comes under the purview of grouping, the general problem of grouping ordered data is discussed. The conclusion sums up the discussion, making it easy to study the new algorithm developed in this work in the light of the existing algorithms.

3. Methodology:

The following steps are involved in our method to predict colleges when we are provided a set of scores. We have utilised a predefined dataset that contains college admit data.

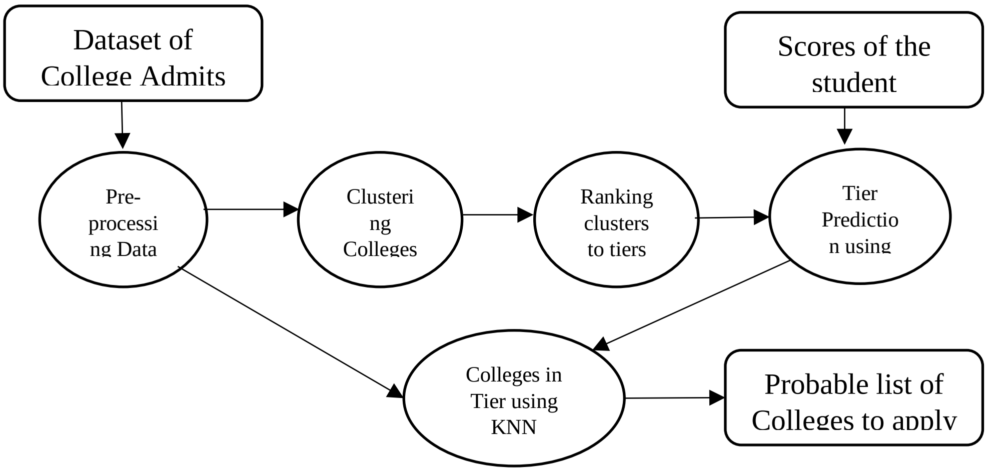

Dataset of College Admits
Scores of the student
Pre-processing Data
Clustering Colleges
Ranking clusters to tiers
Tier Prediction using
Colleges in Tier using KNN
Probable list of Colleges to apply

Figure 1. Flow of algorithm

3.1. Clean and pre-process the data to usable form.

There may arise some discrepancies while entering the data of a person. The data may either get corrupted or may not be entered at all. In the case of missing or corrupted data, we use the mean of the remaining data as a replacement for this.

Data Pre-processing steps was taken on the dataset owing to the fact that it was noisy and unclean. Initially we filter the data by selecting only data corresponding to MS and Accepted Admit records. Following which, we remove columns and retaining only GRE scores, GPA and college Name. Next, we remove fields that have NULL data for Scores. This step is followed by normalizing GPA from 10-point scale to 4-point scale and also normalizing GRE scores from 800 scale(old) to 170 scale(new).

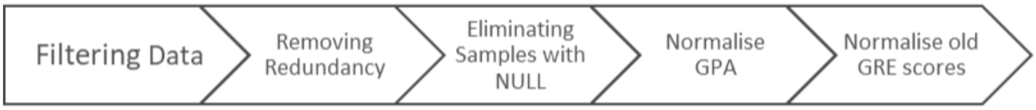


Figure 1. Pre-processing steps

3.2. Allot each college to corresponding tiers based on clustering them.

Colleges are clustered using the K means Algorithm. K means algorithm involves splitting up of the data into k different groups. To calculate that similarity, we will use the Euclidean distance as measurement. The algorithm works by initially finding k different centres randomly. We then categorize each item to its closest mean, and we update the mean's coordinates, which are the averages of the items categorized in that mean so far. We repeat the process for a given number of iterations and at the end, we have our clusters. The clusters were then evaluated using Silhouette Score. It determines the similarity between clusters based on inter-cluster and intra-cluster distances. The clusters then obtained were ranked into tiers based on the average scores for each of the tiers.

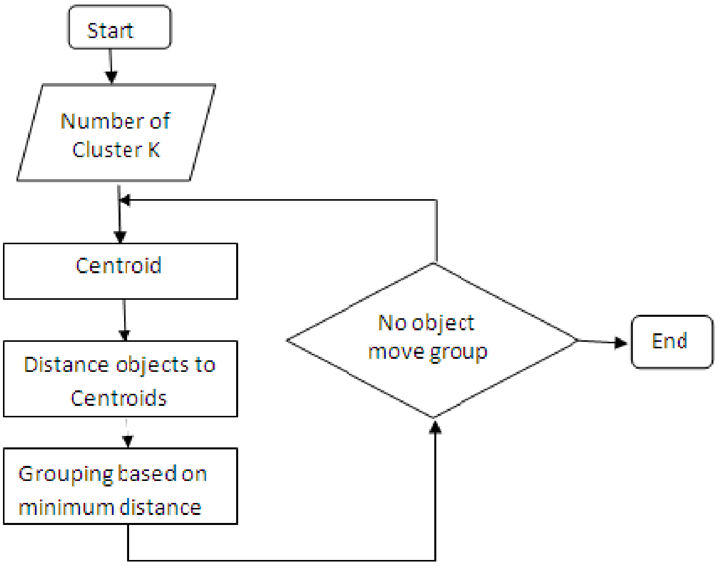

Start
Number of Cluster K
Centroid
Distance objects to Centroids
Grouping based on minimum distance
No object move group
End

Figure 3. K – means

3.3. Create a Machine Learning model to predict the tier for a given set of Input Scores

As inputs, we provide the following the GRE scores for verbal and Quants, the GPA and the written essay scores. We then adjust the weights by means of a backpropagation with gradient descent method. The hidden layers will be 2 in number and output will have 5 neurons each telling the probability of a tier. Unlike other algorithm, this increases efficiency with an increase in the amount of data even though there is skewness in the data, with high concentration of data towards the mid tiers.

A multilayer perceptron (MLP) is an example of feedforward neural network. An MLP consists of, a minimum of three layers of nodes: an output layer, a hidden layer and an input layer. MLP takes the help of a supervised learning technique called backpropagation for training the weights of the network. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result.

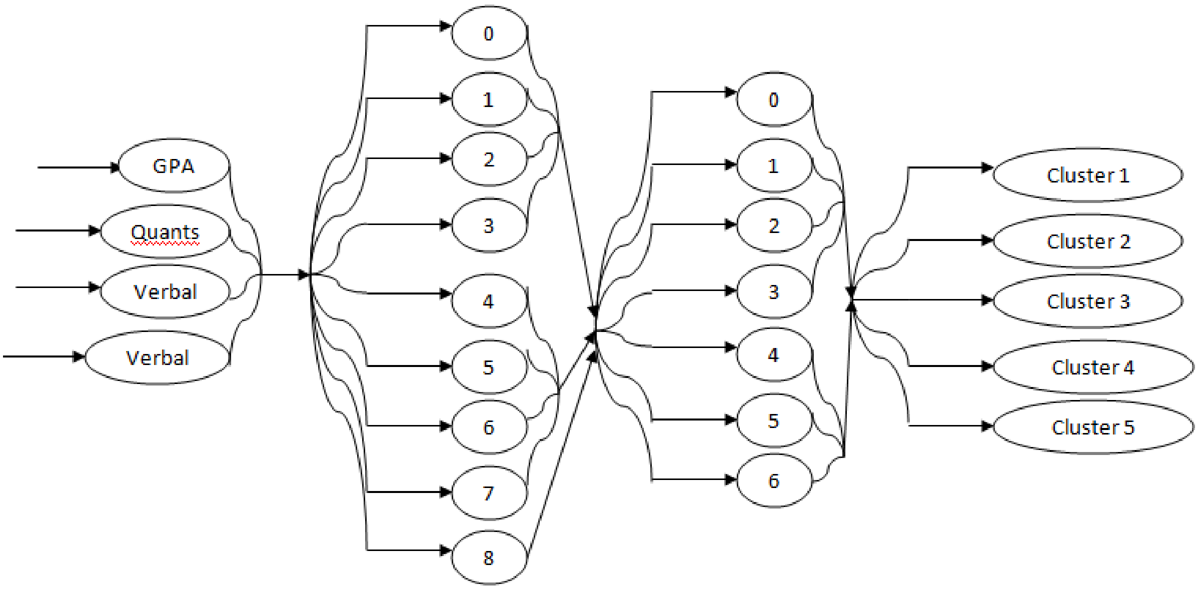


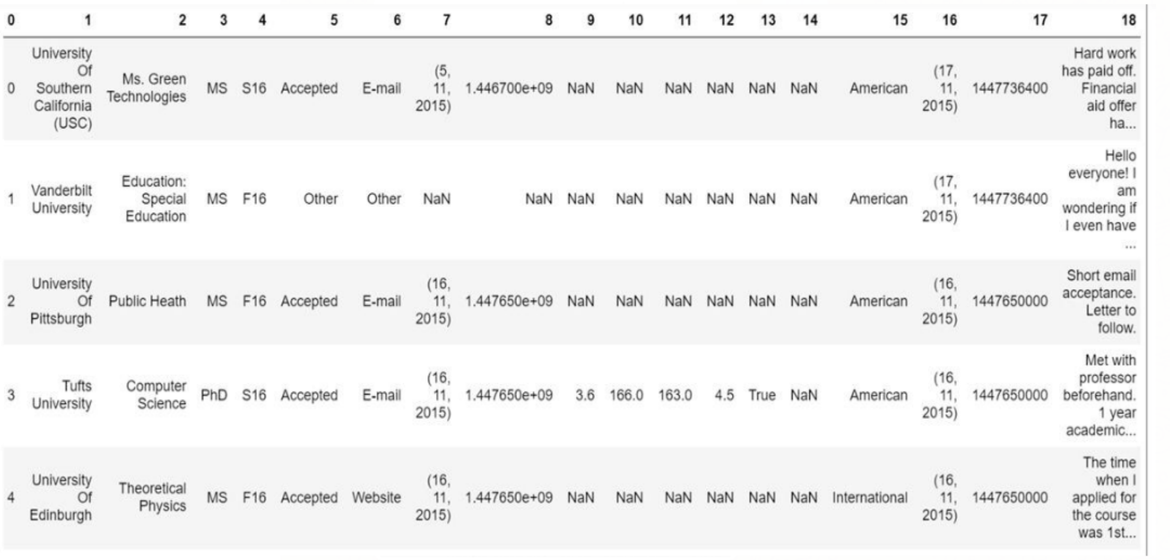
Figure 4. MLP layout

3.4. From the particular tier predicted provide a list of colleges closest in score.

In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. Based on the probability of a person we utilise the ciel value of probability times 10 as the number of colleges ‘k’ from that tier. We use the k-NN algorithm to find a certain number of colleges in that respective tier, which are closer to the person’s scores.

4. Results and Discussion:

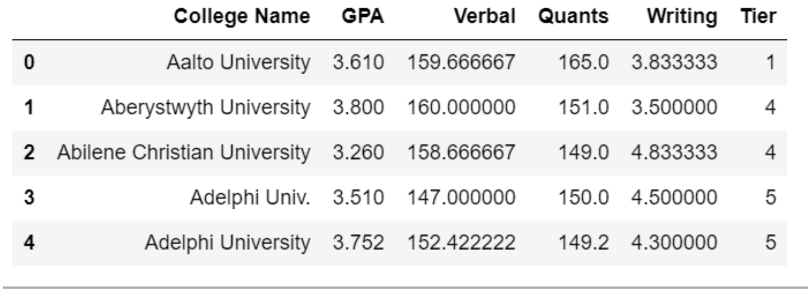
The dataset which we took was filled with impurities as previously said. A sample shot of this data is given below in Table 1. Once we processed the data we obtained a dataset of approximately thirteen thousand entries. The pre-processed data only contained the scores and the college names.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0	University Of Southern California (USC)	Ms. Green Technologies	MS	S16	Accepted	E-mail	(5, 11, 2015)	1.446700e+09	NaN	NaN	NaN	NaN	NaN	NaN	American	(17, 11, 2015)	1447736400	Hard work has paid off. Financial aid offer ha...
1	Vanderbilt University	Education: Special Education	MS	F16	Other	Other	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	American	(17, 11, 2015)	1447736400	Hello everyone! I am wondering if I even have ...
2	University Of Pittsburgh	Public Heath	MS	F16	Accepted	E-mail	(16, 11, 2015)	1.447650e+09	NaN	NaN	NaN	NaN	NaN	NaN	American	(16, 11, 2015)	1447650000	Short email acceptance. Letter to follow.
3	Tufts University	Computer Science	PhD	S16	Accepted	E-mail	(16, 11, 2015)	1.447650e+09	3.6	166.0	163.0	4.5	True	NaN	American	(16, 11, 2015)	1447650000	Met with professor beforehand. 1 year academic...
4	University Of Edinburgh	Theoretical Physics	MS	F16	Accepted	Website	(16, 11, 2015)	1.447650e+09	NaN	NaN	NaN	NaN	NaN	NaN	International	(16, 11, 2015)	1447650000	The time when I applied for the course was 1st...

Table 1. Initial Dataset

Following which we added the Tiers associated with each college and that is shown below. Table 2. contains a sample of the data in which the colleges are clustered into tiers based on the mean scores in each cluster. The evaluation metric for the k-means clustering used was silhouette score and we obtained an acceptable value of around 0.34.



	College Name	GPA	Verbal	Quants	Writing	Tier
0	Aalto University	3.610	159.666667	165.0	3.833333	1
1	Aberystwyth University	3.800	160.000000	151.0	3.500000	4
2	Abilene Christian University	3.260	158.666667	149.0	4.833333	4
3	Adelphi Univ.	3.510	147.000000	150.0	4.500000	5
4	Adelphi University	3.752	152.422222	149.2	4.300000	5

Table 2. Tiers of Colleges

In our training of the Multilayer perceptron we realised by general convention that 2 hidden layers would be required. However, we were not sure of the number of neurons in each layer. In order to determine we ran a loop for each layer and determined that for 9 and 7 neurons in the first and second hidden layers we obtained the maximum accuracy of 70.2% on a test set derived from the same distribution. We also experimented with the random forest classifier involving 1000 sub trees and got a accuracy much lower of around 65%. Comparing the two methods we chose the multilayer perceptron as we can feed it more data and its accuracy would significantly increase.

5. Conclusion:

We have focused on making the process of college admission more convenient and help students to choose colleges which fits best for them based on their GRE scores. We have used various machine learning algorithms like K Means Clustering and Multi-layer perceptron and many more optimization techniques for the college recommendation process. For more accuracy and optimality in recommendations we have used metrics like Silhouette Score to enhance the clustering of colleges into tiers. Finally, the usage Nearest Neighbour Algorithm helps in increasing the closeness of colleges to a score from within a tier, thereby providing a complete solution for the same.

References

[1] Subba Reddy.Y and Prof. P. Govindarajulu," A survey on data mining and machine learning techniques for internet voting and product/service selection", IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.9, September 2017.

[2] Zhibo Wang, Jilong Liao, Qing Cao, Hairong Qi, and Zhi Wang, "Friend book: A Semanticbased Friend Recommendation System for Social Networks", IEEE Transactions on Mobile Computing.

[3] J. Bobadilla et al. "Knowledge-Based System" 2013 Elsevier B.V.

[4] Hector Nunez, Miquel sanchez-Marre, Ulises Cortes, Joaquim Comas, Montse Martinez, Ignasi Rodriguez-Roda, Manel Poch, "A Comaprative study on the use of similarity measure in case based reasoning to improve the classification of environmental system situations,", ELSEVIER, Environmental Modeling and Software XX (2003) xxx-xxx.

[5] DINO IENCO, RUGGERO G. PENSA and ROSA MEO, "From Context to Distance: Learning Dissimilarity for categorical Data Clustering," ACM Journal Vol. X. 10 2009, pages 1- 0??.

[6] Duc Thang Nguyen, Lihui Chen, Chee keong Chan, "Clustering with Multi viewpoint Based Similarity Measure," IEEE Transactions on Knowledge and Data Engineering. Vol. 24. No. 6. June 2012.

[7] Elham S.Khorasani, Zhao Zhenge, and John Champaign. AMarkov Chain Collaborative Filtering Model for Course Enrollment Recommendations: 2016, "IEEE International Conference on Big Data (Big Data)", P. 3484 – 3490.

[8] Hana Bydžovská. Course Enrollment Recommender System: Proceeding of the 9th International Conference on Educational Data Mining, P. 312 – 317.

[9] Jamil Itmazi and Miguel Megias (2008), Using recommendation Systems in Course Management Systems to Recommend Learning Objects, P. 234 – 240.

[10] Queen Esther Booker (2009). A Student Program Recommendation System Prototype: Issues in Information Systems, P. 544 - 551.

[11] Akrivi Vlachou, Christos Doulkerids, Kjetil Norvag, and Yannis Kotidis, "Identifying the Most Influential Data Objects with Reverse Top-k Queries," Proceedings of the VLDB Endowment, Vol. 3, No. 1, Copy right 2010 VLDB Endowment 2150-8097/10/09.

[12] Usue Mori, Alexander Mendiburu, and Jose A.Lozano, "Similarity Measure Selection for Clustering Time Series databases," IEEE Transactions on Knowledge and Data Engineering. Vol. 28. No. 1. January 2016.

[13] Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee, "A Similarity Measure for Text Classification and Clustering," IEEE Transactions on Knowledge and Data Engineering. Vol. 26. No. 7. July 2014.

[14] Amit Singh, Hakan Ferhatosmanoglu, and Ali Saman Tosun, "High Dimensional Reverse Nearest Neighbor Queires," CIKM'03, November 3-8, 2003, New Orleans, Louisiana, USA, copyright 2003 ACM 1-58113-723-0/03/0011.

[15] Charif Haydar, Anne Boyer, "A New Statistical Density Clustering Algorithm based on Mutual Vote and Subjective Logic Applied to Recommender Systems", UMAP 2017 Full Paper UMAP'17, July 9- 12, 2017, Bratislava, Slovakia.

[16] Reddy, M. Y. S., & Govindarajulu, P. (2018). College Recommender system using student'preferences/voting: A system development with empirical study. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, 18(1), 87-98.

[17] Deokate monali, Gholave Dhanashri, Jarad Dipali, Khomane Tejaswini (2018). College Recommendation System for Admission. International Research Journal of Engineering and Technology, 9(3), 187-175.

[18] Paradarami, T. K., Bastian, N. D., & Wightman, J. L. (2017). A hybrid recommender system using artificial neural networks. Expert Systems with Applications, 83, 300-313.

[19] Qazanfari, K., Youssef, A., Keane, K., & Nelson, J. (2017, October). A novel recommendation system to match college events and groups to students. In IOP Conference Series: Materials Science and Engineering (Vol. 261, No. 1, p. 012017). IOP Publishing.

[20] Bouzekri, E., Canny, A., Fayollas, C., Martinie, C., Palanque, P., Barboni, E., ... & Gris, C. (2019). Engineering issues related to the development of a recommender system in a critical context: Application to interactive cockpits. International Journal of Human-Computer Studies, 121, 122-141.

[21] Kanoje, S., Mukhopadhyay, D., & Girase, S. (2016). User Profiling for University Recommender System Using Automatic Information Retrieval. Procedia Computer Science, 78, 5-12.

[22] Zhang, H. R., & Min, F. (2016). Three-way recommender systems based on random forests. Knowledge-Based Systems, 91, 275-286.