WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

CGPA-Based University Recommendation System

Haresh Gayakhe[1], Dipesh Ghag[2], Prof. Dr. Pooja Raundale[3]

[1]Master of Computer Applications, Sardar Patel Institute of Technology, Mumbai, India dipesh.ghag@spit.ac.in
[2]Master of Computer Application, Sardar Patel Institute of Technology, Mumbai, India haresh.gayakhe@spit.ac.in
[3]Master of Computer Application, Sardar Patel Institute of Technology, Mumbai, India pooja@spit.ac.in

ABSTRACT—

A prospective graduate student faces a dilemma while deciding which university to apply to. Students frequently question whether their profile is strong enough for a particular university. This issue has been dealt with in this research by modeling a recommendation system based on different regression and classification techniques. The input elements dealt with the academic background of the student which includes CGPA, SGPA score for the last two semesters, Department, and Degree for which admission is taken. There are various supervised classification machine learning algorithms for prediction and recommendation purposes like Decision Tree, Random Forest, KNN, and naive Bayes, but since our dataset is small it is not possible to achieve intended accuracy with them that is why we are using a KNN classification algorithm which calculates Euclidian distance between two points and at the end sort all the points by ascending order, and then can recommend top N values closer to the given point. This university recommendation system takes the various input data from the user and recommends the N universities where the chances of admission are high as compared to recommending only one university.

Keywords—Recommendation system, classification algorithm, CGPA, SGPA, Department, Supervised.

I. INTRODUCTION

When students graduate from college most of them aim towards higher education usually in the best educational institutions in the world so that they can have better futures for themselves and their families this mere thought can put them under pressure as it can be seen as choosing the right university is key to their future.

Choosing the right university is very important but, in this case, it does not depend on the student it depends on the student's educational achievements and his motive behind applying for this university and so it can be said that it is always favorable if a student can stick with 3-4 universities and apply for them only rather than applying for every university, he/she finds interesting.

Normal university/college recommendation system uses different exam scores and other aspects like GRE score, TOEFL score, previous university rank, LOR strength, Research, etc., but we are only using the data that can be available after passing your degree like CGPA, SGPA, Department and the degree you are hoping to do. That's what makes our recommender system different from other systems. Also, in this case, the normal classification algorithm does not work as we are working on a relatively small dataset so it will give us very little accuracy to avoid this problem, we can use the KNN classification algorithm that will calculate the Euclidian distance between two points and sort all the distances in ascending order so it will recommend N universities at a time and then we won't have to worry about accuracy as we are getting multiple recommendations at a time.

II. LITERATURE REVIEW

"College Recommendation System" by "Vinit Jain, Mohak Gupta, Jenish Kevadia", in this paper they had proposed a college recommendation system using Data Mining and Query Optimization techniques which generates the list of colleges in which the candidate is most probable to be eligible. This system is for students who have passed SSC and are preparing to take admission into junior college. They need to do the tedious work of selecting streams and filtering out selective colleges to fill out from the hundreds of colleges, this system aims to automate this process [1].

The authors of the study "College Recommendation System" by "Leena Deshpande, Nilesh Dikhale, and Himanshu Shrivastav" had suggested various data analysis and data mining approaches that might be employed for the system. This approach was developed with students, parents, and educationalists looking for engineering colleges in mind. Recommendation systems use extensive data searches to address the issue of information overload. Different prediction methods are available to assist recommendation systems in gathering data. It uses data mining and machine learning methods to filter data and deliver the necessary information. The list of colleges can be obtained using similar data mining approaches [2].

To address challenges with predicting college acceptance rates, Abdul Hamid M. Ragab, Abdul Fatah S. Mashat, and Ahmed M. Khedra suggested a unique design for a "Hybrid Recommender System for College Admissions" based on data mining techniques and knowledge discovery principles. For excellent performance, this system comprises two cascade hybrid recommenders functioning in tandem with a college predictor. For students in the preparatory year, the first recommender allocates tracks to the students. While the second recommender places students who did well on their preparatory year exams in the specialized college. This predictor system leverages GPA data from prior kids' college admissions to forecast the most likely institutions. It examines a student's academic achievements, history, academic records, and college entrance requirements [3].

The research paper "CAPSLG: College Admission Predictor and Smart List Generator" authored by "Kiran Kumari, Meet Kataria, Viral Limbani, and Rahul Soni" deal with the problem of admission into engineering after HSC students face the dilemma of choosing right engineering college and branch, also there are various factors that affect the admission like HSC Marks, CET score, JEE score, University type, and Reserved category. So, this research paper proposed a system to generate a smart list of colleges with a high probability of getting admission based on past data, they also take the distance between home and college into account. They have used AdaBoost classification algorithms for this task after comparing it with other similar algorithms such as Random Forest and Decision Tree [4].

"A Novel Approach for Colleges Recommendation for Admission Seekers Using Decision Tree" by "Rohit Gharge, Nilesh Nerelekar, Aditya Shinde, and Prof. Rakesh Suryawanshi" have created a college recommendation system for students seeking admission into engineering they have used Random Forest classifier to compile preference list best suited for you based on the CET score, branch University type. The list will be generated using the past data [5].

"Zhe Wang, Hao Xu, Pan Zhou, and Gang Xiao" in their Research Paper "An Improved Multilabel k-Nearest Neighbor Algorithm Based on Value and Weight" have proposed a new improved ML-KNN Algorithm for multilabel classification as normal ML-KNN has poor performance on imbalanced multilabel data [6].

"Ensemble of Networks for Multilabel Classification" By "Loris Nanni, Luca Trambaiollo, Sheryl Brahnam, Xiang Guo, and Chancellor Woolsey" proposes merging ensembles and deep learners' methods by combining a set of gated recurrent units, temporal convolutional neural networks, and long short-term memory networks trained with variants of the Adam optimization approach [7].

"College Admission Predictor" Students can enter their grades and personal information into the College Admission Predictor System, a web-based application system. This aids in predicting their college admissions. An administrator can enter batches and college data. The entry seat allocation process is made easier and more efficient by using this application [8].

This Research paper proposes a system to predict chances of admission into foreign universities using various ML algorithms based on GRE score, TOEFL score, LOR strength, Research, and University rank [9].

"A Recommendation System for Selecting the Appropriate Undergraduate Program at Higher Education Institutions Using Graduate Student Data" By "Yara Zayed, Yasmeen Salman, and Ahmad Hasasneh" proposes an intelligent recommendation system that helps students choose the best university major based on prior knowledge and information, including an applicant's gender, the applicant's past performance, labor market statistics, the applicant's marks, the applicant's behavior, and the expected salary after completing their studies. Different ML techniques, such as the decision tree classifier (DTC), support vector machine (SVM), and random forest (RF) classifiers, were used and researched to achieve this [10].

By considering the nonlinear and higher-order interactions among the elements, this research paper proposes an ICF solution that is more expressive. Using non-linear neural networks, we consider the interaction between all interacting item pairs in addition to modeling just the second-order interaction (such as the similarity between two things). This allows us to precisely imitate the higher-order connection. among the items and capture more profound consequences in user decision-making [11].

III. METHODOLOGY

A. Data Collection

Collecting data such as CGPA, SGPA of each Sem (if possible), Department, Degree, and university for which student got an admission, from various sources and compiling that data into one excel/CSV file for further processing.

sr.no	name	year	department	program	Sem7	Sem8	cgpi	university
1	Fernandes Lance Vincent	2021	ETRX	MS, Electrical and Electronics	10	9.83	9.92	Purdue University
2	Pote Sameep Vijay	2021	ETRX	Master of Engineering, Robot	9.65	9.17	8.46	University of Maryland
3	Vishnani Vinay Bharat	2021	ETRX	MS, Electrical and Computer E	10	9.83	9.94	University of Texas
4	Bhise Rugved Mandar	2021	ETRX	MS, Computer Science	9.82	9.39	8.34	University of California
5	Kumar Siddhant Rajeev	2021	ETRX	M.S. Data Science	10	9.58	8.98	Columbia University
6	Mane Arthav Ashok	2021	ETRX	MS, Computer Science	9.65	9.13	8.96	University at Buffalo
7	Pahalwan Sparsh Sunil	2021	ETRX	MSc Management Student at	10	9.17	7.45	University college Dublin
8	Aditi Kandoi	2021	COMPS	M.S. computer science	10	10	8.89	Stony Brook University
9	Manasi Beldar	2021	COMPS	M.S. computer science	9.71	10	8.1	Oklaham state university
10	Rahul Ramteke	2021	COMPS	M. Tech cybersecurity	9.57	9.58	7.35	National Forensics Science University
11	Meet Bawankar	2021	IT	Master of Science Manageme	9.76	8.54	6.64	University of Illinois
12	Vrinda Bhatu	2021	IT	Master of Science program in	10	9.58	8.81	University of Maryland
13	Yash Gandhi	2021	IT	Master of Science Computer	10	10	9.54	University of Southern California

B. Data Cleaning and Transformation

Dropping records containing NULL values from the dataset as each record holds its value and cannot be replaced by any other method as each attribute is important then feature encoding can be done for continuous values into discrete categorical values.

C. Feature Engineering

Feature engineering is a crucial step in this proposed methodology as it involves the selection of the most relevant features of the machine learning model. The total features include name, year, department, program, sem7, sem8, cgpa, and university of higher education. Drop the features or attributes that do not contribute to the final output and keep only the features essential for prediction which will give more and more accurate predictions.

D. Model Training and Evaluation

After we have a final dataset, we will split the dataset into training and testing and also convert data into one standardized form using sklearn's standard Scaler.

Feature scaling

```
[16]: from sklearn.preprocessing import StandardScaler

[17]: sc = StandardScaler()

[18]: x_train = sc.fit_transform(x_train)
      x_test = sc.transform(x_test)

[19]: x_train[5]

[19]: array([0.02228261, 0.9359573 , 0.80340342])
```

In this study, we have used various classification algorithms like Decision Tree, Naïve Bayes, SVM, and KNN algorithm as a multiclass algorithm but since the size of the dataset is small and the unique values to predict are around 57 all the above algorithms give accuracy close to none, and this can be of concern.

Decision Tree

```
]: from sklearn.tree import DecisionTreeClassifier
   # Build a Decision tree Classifier
   DTC = DecisionTreeClassifier(max_depth = 5)

   # Model training
   DTC.fit(x_train, y_train)

   # Predict Output
   predicted = DTC.predict(x_test)
   # print("Predicted Value:", predicted);
   print(accuracy_score(y_test, predicted));

   0.041666666666666664
```

Naive Bayes

```
# Build a Gaussian Classifier
GNB = GaussianNB()
# Model training
GNB.fit(x_train, y_train)
# Predict Output
predicted = GNB.predict(x_test)
# print("Actual Value:", y_test[:5]);
# print("Predicted Value:", predicted[:5]);
print(accuracy_score(y_test, predicted))
```

```
0.08333333333333333
```

SVM

```
svm = svm.SVC()
svm.fit(x_train, y_train)
y_pred2 = svm.predict(x_test)
# print("Actual Value:", y_test[:5]);
# print("Predicted Value:", predicted[:5]);
print(accuracy_score(y_test, y_pred2))
```

```
0.041666666666666664
```

To tackle this problem, we have used KNN classification algorithm and instead of showing one value we have shown N (1 < N < 5) values by calculating Euclidian distance for each value for input and sorted it into ascending order in this way accuracy score can be improved and since it is recommendation system recommending N universities instead of 1 will be useful in a way because while applying student can apply for that N universities since they have highest Euclidian distance compared to other options so he/she can have better chance of getting admission into one of them.

```
def euclideanDistance(data1, data2, length):
    distance = 0
    for x in range(length):
        distance += np.square(data1[x] - data2[x])
    return np.sqrt(distance)

def knn(trainingSet, testInstance, k):
    print(k)
    distances = {}
    sort = {}
    length = testInstance.shape[1]

    for x in range(len(trainingSet)):

        dist = euclideanDistance(testInstance, trainingSet.iloc[x], length)

        distances[x] = dist[0]

    sorted_d = sorted(distances.items(), key=lambda x: x[1])


    neighbors = []

    for x in range(k):
        neighbors.append(sorted_d[x][0])

    classVotes = {}
```

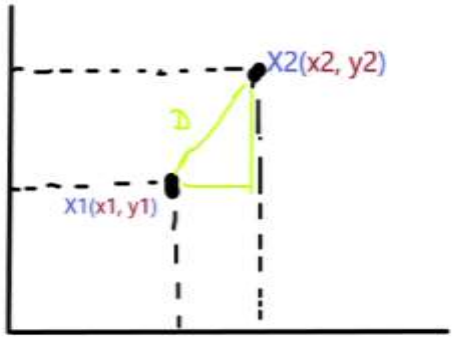
```
1]: k = 4
result,neigh= knn(data, test, k)

list1 = []
list2 = []
for i in result:
    list1.append(i[0])
    list2.append(i[1])
for i in list1:
    print(i)
```

```
4
Stony Brook University
University of Florida
Virginia Tech
Georgia Institute of Technology
```

Euclidean Distance

We mostly use this distance measurement technique to find the distance between consecutive points. It is generally used to find the distance between two real-valued vectors. Euclidean distance is used when we must calculate the distance of real values like integers, float, etc. One issue with this distance measurement technique is that we must normalize or standardize the data into a balance scale otherwise the outcome will not be preferable.

Let's take an example of a 2-Dimensional vector and take a geometrical intuition of distance measures on 2-Dim data for a better understanding.



From the above image, you can see that there are 2-Dim data X1 and X2 placed at certain coordinates in 2 dimensions, suppose X1 is at (x1, y1) coordinates and X2 is at (x2, y2) coordinates. We have 2-dim data so we considered F1 and F2 two features and D is considered as the shortest line from X1 and X2, if we find the distance between these data points that can be found by the Pythagoras theorem and can be written as:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

IV. FUTURE SCOPE AND LIMITATIONS

A. Future Scope

a) The department can be encoded and then recommendations based on the department can be done.

b) If we encode department and degree, then recommendations can be more accurate.

c) Deep learning algorithms can be applied and after finding suitable evaluation matrix algorithms can be compared.

d) More data needs to be collected so that way accuracy of normal classification algorithms can be improved.

e) A proper Web App with other functions besides university recommendation can be developed instead of a simple GUI.

B. Limitations

a) The dataset is very small, which is why normal classification algorithms are ineffective.

b) Using the last two semesters' SGPA instead of all Eight semesters' SGPA affects the accuracy.

V. CONCLUSION

"CGPA Based University Recommendation System" has explored the possibility of using data related to students' academics to recommend N universities for higher education instead of mainstream aspects such as GRE score, TOEFL score, etc. The study tried classification algorithms first to predict and evaluate the results, but it was not successful as envisioned then the KNN classification algorithm was used but this time instead of 1 recommendation N recommendations were given as it was more suitable for this problem statement. KNN algorithm calculates Euclidian distance. Between each point and then stored them in ascending order as the distance is less the similarity is greater and gives us top N results. The results were satisfactory enough, but more algorithms need to be applied to this data and for that more data is needed as time passes that data needs to be collected so that in the near future appropriate accuracy can be achieved using the algorithms that did not give higher accuracy before and also other algorithms besides them can be applied.

REFERENCES

[1] Vinit Jain, Mohak Gupta, Jenish Kevadia, Prof. Krishnanjali Shinde, 2017, College Recommendation System, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICIATE – 2017 (Volume 5 – Issue 01),

[2] Leena Deshpande, Nilesh Dikhale, Himanshu Srivastava, Apurva Dudhane, Umesh Gholap "College recommendation system", ISSN: 2321-9637, NCPCI-2016, 19 March 2016, in press.

[3] Abdul Hamid M. Ragab, Abdul Fatah S. Mashat, Ahmed M. Khedra, "Hybrid Recommender System for Predicting College Admission", 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 107-113.

[4] Kumari, Kiran and Kataria, Meet and Limbani, Viral and Soni, Rahul, CAPSLG: College Admission Predictor and Smart List Generator (April 9, 2019). 2nd International Conference on Advances in Science & Technology (ICAST) 2019 on 8th, 9th April 2019.

[5] Rohit Gharge, Nilesh Nerelekar, Aditya Shinde, Prof. Rakesh Suryawanshi. "A Novel Approach for Colleges Recommendation for Admission Seekers Using Decision Tree," IJARSCT Volume 5, Issue 5, May 2020.

[6] Wang, Zhe & Xu, Hao & Zhou, Pan & Xiao, Gang. (2023). An Improved Multilabel k-Nearest Neighbor Algorithm Based on Value and Weight. Computation. 11. 32. 10.3390/computation11020032.

[7] Nanni, L.; Trambaiollo, L.; Brahnam, S.; Guo, X.; Woolsey, C. Ensemble of Networks for Multilabel Classification. Signals 2022, 3, 911-931.

[8] Annam Mallikharjuna Roa, Nagineni Dharani, A. Satya Raghava, J. Buvanambigai [Students], K. Sathish(Assistant Professor), Computer Science and Engineering," College Admission Predictor (Volume 8, No. 4, April 2018)", SRM Institute of Science and Technology, Chennai, Tamil Nadu, India, (JNCET) www.jncet.org Volume 8, Issue 4, April (2018)

[9] B. Uday Kiran, B. Simon Paul, T. Pavan Kumar, Mr. M. Sathya Narayana, "ADMISSION PREDICTION FOR MS IN FOREIGN UNIVERSITIES USING MACHINE LEARNING", Volume 4, Issue 12, December 2022.

[10] Zayed, Y.; Salman, Y.; Hasasneh, A. A Recommendation System for Selecting the Appropriate Undergraduate Program at Higher Education Institutions Using Graduate Student Data. Appl. Sci. 2022, 12, 12525.

[11] Xue, Feng & He, Xiangnan & Wang, Xiang & Xu, Jiandong & Liu, Kai & Hong, Richang. (2019). Deep Item-based Collaborative Filtering for Top-N Recommendation. ACM Transactions on Information Systems. 37. 1-25. 10.1145/3314578.