ISSN: 0974-5823 Vol. 7 (Special Issue, Jan.-Mar. 2022)

International Journal of Mechanical Engineering

Comparison of K-Means Algorithm and SVM in College Recommendation System

M.Vinod kumar[1], R. Sabitha[2]

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105

[2]Project Guide, Corresponding Author, Department of Computer Science and Engineering,

Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences,

Saveetha University, Chennai, Tamilnadu. India. Pincode: 602105.

ABSTRACT

Aim: The proposed work targets to evaluate the precision and accuracy in predicting the College Recommendation system for Students using K-Means and Support vector machine(SVM)and classification algorithms. **Materials and Methods:** SVM applied on a college dataset that consists of 778 records. A structure for the College Recommendation system in the educational sector comparing K-Means and Support vector machines has been suggested and expanded. The sample size was calculated as 55 in each group using G power. The precision and accuracy classifiers had been assessed and noted. **Results:** The K-Mean group generates (50%) in predicting the College Recommendation System on the data set used whereas the Support vector machine produces(58.1%). The significant value is 0.0. Hence the Support vector machine seems to be better than the K-Means. **Conclusion :** In terms of precision and accuracy, the results show that the SVM exhibits higher accuracy than K-Means in measuring the efficiency of recommending the Colleges for students.

Keywords: College Recommendation, Support Vector Machine, K-Mean, Innovative Cluster method, kernel based approach, Parameter value.

INTRODUCTION:

Great quantities of techniques are utilized by humans in this technological environment for various purposes. There are numerous software and applications made for humans. We can build A wide variety of lists of universities that can be found that a student is eligible for using this software. In today’s educational and business systems, data mining techniques are critical (Liu 2021). A data mining task can be specified or explained using a data model. The difficult challenge with a college recommendation system is to compile a database of all college students (Ng and Linn 2017). To create a list of colleges from all of the colleges, the candidate must first exclude those colleges where he is not eligible. While going through the admission process, a student must enter a minimum number of colleges to which he may be admitted. As a result, an applicant must compile a list of colleges to which he wants to apply. This system is widely used in colleges, universities, education sectors to recommend the college for students easily (Shin, Lee, and Kim 2005).

Around 39 related articles published in IEEE Xplore and 23 related articles were published related to this work in google scholar. The proper planning is the key to success. Every person has his own objectives and dreams. Every student would believe that we have to work hard when we first start college (Rathnavel et al. 2017). Study hard and finish your courses on time, then finish your degree three are no backlogs. The issue here is that many people are having problems after joining the college into university students are having problems with times and assignments. Several works have demonstrated that the performance of K-Means is poor and provides less accuracy in prediction of college recommendation systems. A study provides an SVM algorithm used widely to improve the college recommendation system variation of students to recommend the colleges easily (Wang 2020). Recommendation system will rank the colleges according to the placements, courses,facilities in the college. The student once registered in the recommendation system he can get the unique id and password (Chen and Yu 2020). Then students can select the best college according to the ratings and rankings. Then he can also produce the rating for that college. (Harries et al. 2021). It is important to analyse and compare the various classification algorithms that provide better accuracy.Previously our team has a rich experience in working on various research projects across multiple disciplines(Ezhilarasan et al. 2021; Balachandar et al. 2020; Muthukrishnan et al. 2020; Kavarthapu and Gurumoorthy 2021; Sarode et al. 2021; Hannah R et al. 2021; Sekar, Nallaswamy, and Lakshmanan 2020; Appavu et al. 2021; Menon et al. 2020; Gopalakrishnan et al. 2020; Arun Prakash et al. 2020)

From the survey, the K-Means has subsequent limitations that require more research in selecting kernel function and also its performance lags with noisy dataset and with the size of dataset (Shin, Lee, and Kim 2005). Hence, the work aims at comparing the accuracy of K-Means and LR algorithms in predicting the college recommendation system.

MATERIALS AND METHODS

The research work was performed in the Department of Computer Science and Engineering Saveetha School of Engineering, SIMATS. The work was carried out of 778 records taken from the college dataset The precision in predicting the college recommendation system was performed by evaluating the two groups. A total of 10 iterations were performed on each group to achieve better precision. The dataset was downloaded from Kaggle website. The dataset contains 778 rows and 18 columns. Some of the important attributes taken for experiment setup are Room Board,Books,Personal,Terminal,Expand,Outstate,SF Ratio,etc(Richardson et al. 2021)

The sample size was calculated as 55 in each group using G Power. The College dataset has been used with a sample size of 303 students,76 features and some missing values. Sample size was calculated using clinical analysis, With alpha and beta values 0.05 and 0.5, 95% confidence, pretest power 80% and enrolment ratio 1.(Gibson and Elrod 2018)

K-Means

Clustering is a strategy for uncovering commonalities and insights about the structure of data in exploratory data analysis. It can be described as a strategy for determining where data points in a dataset can be divided into smaller groups. Clusters are smaller groups with data points that are similar in key ways, whereas data points from distinct clusters are dissimulated. Clustering is classified as an unsupervised learning method because no target classes are supplied against which the clusters' output may be compared in order to assess their performance. This method can be used to keep track of a student's academic progress.The k-means algorithm is an iterative technique that attempts to split datasets into K pre-defined, unique, non-overlapping subgroups (clusters), each of which contains just one data point.

```
Input: College dataset
Output: Accuracy
1. Initialize K centroids randomly
2. Associate each data point in D with the nearest centroid .This will divide the data data points into K clusters.
```

3. Recalculate the positions of centroids.

Repeat steps2 and 3 until there are no more changes in the membership of the data points.

4. Data points with cluster memberships.

Select the number K to decide the number of clusters. Select random K points or centroids; it can be other from the input dataset. Assign each data point to their closest centroid,which will form the predefined K clusters. Calculate the variance and place a new centroid of each cluster. Repeat the third steps, which means assign each datapoint to the new closest centroid of each cluster. If any reassignment occurs, then go to another to FINISH.

Support Vector Machine (SVM)

Support Vector Machine” may be a supervised machine learning algorithmic rule which can be used for each classification of regression challenges. However, within the SVM algorithm, we tend to plot every data item to some extent in an n-dimensional area (where n is the variety of options you have) with the value of every feature of a specific coordinate. Then, we tend to perform classification by finding the hyper-plane that differentiates the two categories alright.

In the SVM classifier, it’s easy to own a linear hyper-plane between these two categories. The SVM algorithmic rule includes a technique known as the kernel trick. The SVM kernel is a perform that takes a low dimensional input area and transforms it to a higher dimensional area. It’s principally helpful in non-linear separation problems.

The Pseudocode for SVM is as follows:

```
Inputs: College dataset
Output: Selected features and Accuracy.
1. Load the dataset
2. Split the dataset randomly into training (80%) and testing (20%) dataset
3. Set the target variable
4. Generate the SVM classifier based on the training set
5. Train the classifier using rbf kernel parameter
6. Predict the testing set based on training dataset
7. Evaluate the classifier.
8. Return Accuracy.
```

Support vector machine(SVM) is a regulated machine learning algorithm which can be utilized for both classification and regression challenges. In this study, to train the SVM the svc class of scikit learn library was used. Import the college.csv dataset and load the dataset. The dataset is split randomly into training (80%) and testing (20%) sets.The target variable is selected. Then, the SVM classifier based on the training set is generated. Rbf is used as the parameter value for this kernel based approach. The proposed method implements an innovative cluster method that incorporates a kernel based approach. The testing set is predicted based on the training set. The SVM classifier is evaluated and the accuracy is calculated.

The proposed work was experimented in Google Colab, The Hardware and Software requirements for experimenting the work includes i3 processor, 50GB HDD, 4GB RAM,Windows OS, Python: Colab/Jupyter.

Initially, the dataset was divided into two parts: training and testing sets. Then the algorithm is experimented on the training and testing sets. The training and testing sets are varied 10 times based on test set size. Table 1 depicts the comparison of accuracy and precision of K- mean and Decision Tree for 10 iterations.

The various parameters for the analysis can be calculated as follows:

Equation (1) - Accuracy : It identifies the number of instances that were correctly classified.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (1)$$

Precision is used to calculate which part of prediction data is positive using equation (2).

$$Precision = \frac{Tp}{TP + Fp} \quad (2)$$

Recall is also called sensitivity which calculates the relevant instances that are selected, which is calculated using equation (3).

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Here “TN” means True Negative, “TP” means True Positive, “FP” means False Positive and “FN” means False Negative.

F-measure measures model accuracy on a dataset using equation (4).

$$F - measure = 2 \times (\frac{precision \times Recall}{Precision + Recall}) \quad (4)$$

Statistical Analysis

Besides experimental analysis, the work was evaluated statistically using Statistical Package for Social Sciences (SPSS). The analysis was done to obtain Mean, Standard Deviation and Standard Error Mean. An independent variable T Test was carried out to compare the parameters on both the groups. The analysis uses several independent variables such as apps, accept, top 10 percentage, top 25 percentage, f.undergrade, p.undergrade, out state, room, board, books, terminal, ph.d, expand. The dependent variables used are accuracy and precision.

RESULTS

Table 1 shows the comparison of accuracy and precision of both the groups for 10 iterations. Table 2 depicts the various parameters of both groups. The accuracy, Precision, Recall, F1 Score and F2 Score has been calculated for K-Mean and decision tree.The analysis of two groups shows that SVM has higher accuracy (85.1%) and Precision (68.4%) compared to K-Mean. From Fig. 1 and Fig.2 , it is inferred that the ROC graph shows the performance of the K-Mean and SVM classification model at various classification thresholds. Table 3 shows the statistical analysis of K-Mean and Support Vector Machine with different test datasets. An innovative cluster method that incorporates a kernel based approach is applied. The mean accuracy of the Support Vector Machine model appears to be higher than the K-Mean model. Also, the precision of the Support Vector Machine is much higher than the K-Mean. The performance of the Support Vector Machine algorithm is superior to the K-Mean algorithm.The Table 4 depicts the statistical analysis of Significant levels for both groups. There is no Significant difference among the two groups. Hence the decision tree is better than K-Mean. Fig. 3 inferred the mean accuracy and mean precision of K-Mean and Support Vector Machine. The statistical analysis of two independent groups shows that Support Vector Machines have higher accuracy mean (85.1%) and Precision mean (68.4%) compared to K-Mean. The mean error of K-Mean is a little lesser than Support Vector Machine.

DISCUSSION

Prediction of college students is a major issue in the college recommendation system. Experimental work was done among two groups K-Mean and Support Vector Machine by varying the test size. From the experimental results (Fig. 3 and Fig. 4) done in Google colab, the accuracy and precision of the Support Vector Machine by applying an innovative cluster method which uses a kernel based approach is 89.40% and 85.20%, whereas K-Mean provides the accuracy (47.00%) and precision to be (47.10%). This depicts that Support Vector Machine is better than K-Mean. The various parameters like TP rate, FP rate, Precision, Recall, F-measures are also compared From the SPSS graph, the proposed Support Vector Machine Classifier which uses a kernel based approach performs better in terms of accuracy (85.1%) and precision (68.4%) compared with the K-Mean algorithm. Fig. 3 depicts that the mean error of the Support Vector Machine is found to be little higher than K-Mean, which has to be minimised.

The most important aspect in predicting college recommendation is accuracy and precision. In the study by a machine-learning-based diagnosis system for college recommendation prediction by using a college student dataset was proposed. Popular machine learning algorithms(Corker et al. 2017), three feature selection algorithms, the cross-validation method, and seven classifiers performance evaluation metrics such as classification accuracy, specificity, sensitivity, Matthews' correlation coefficient, and execution time were used by the study (Banik 2018). In the study a scalable solution was proposed for predicting college recommendation systems. The Support Vector Machine algorithm was used on spark framework for predicting college recommendation and demonstrated that even with a dataset of 600 documents, achieving a higher accuracy rate by the study (Chung et al. 2020)

In the study, attribute filtering, frequent item mining and a variety of data mining techniques such as Support vector machine and KNN classifications are used for predicting recommendation systems at early stages (Aggarwal 2016). When it comes to predicting college recommendations, the accuracy was superior to that of other algorithms.

The accuracy of the Support Vector Machine classification algorithm depends on the training and testing dataset size (Ricci, Rokach, and Shapira 2015). In our study, the accuracy and precision appears to be better than the K-Mean. However, the mean error seems to be higher in our proposed work which has to be minimized.

Although the results of the study are better in both experimental and statistical analysis, there are certain limitations in the work. The evaluation of accuracy cannot provide a better outcome on larger data sets. Moreover in K Means, selecting the initial starting number of cluster centers is difficult. The mean error also appears to be higher than Decision Tree. It would be better if the mean error can be reduced to a considerable extent. In future, the work can be enhanced by applying optimization algorithm techniques, to achieve better accuracy and less mean error. Feature selection algorithms can be used before classification to improve the classification accuracy of classifiers (Ng and Linn 2017). Hence, through Support Vector Machine algorithms, we can reduce the computation time and improve the classification accuracy of classifiers.

CONCLUSION

The work shows that the accuracy and precision for college recommendation prediction using Support vector machine (SVM) by applying an innovative cluster method which uses a kernel based approach appears to be better than the K-Means. The mean error is found to be little higher than K-Mean. Hence, it is concluded that SVM results in acceptable accuracy and precision than K-Mean.

DECLARATIONS

Conflict of Interests

No conflict of interest in this manuscript.

Author Contribution

Author VK was involved in data collection, data analysis, algorithm framing, implementation and manuscript writing. Author RS was involved in designing the work flow, guidance and review of manuscript.

Acknowledgements

We would like to acknowledge Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing facilities and constant help to carry out this study.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. ZONIAC Software Pvt. Ltd.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Aggarwal, Charu C. 2016. *Recommender Systems: The Textbook*. Springer.
2. Appavu, Prabhu, Venkata Ramanan M, Jayaprabakar Jayaraman, and Harish Venu. 2021. “NOx Emission Reduction Techniques in Biodiesel-Fuelled CI Engine: A Review.” *Australian Journal of Mechanical Engineering* 19 (2): 210–20.
3. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. “Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite.” *Biomass Conversion and Biorefinery*, August. https://doi.org/10.1007/s13399-020-00938-0.
4. Balachandar, Ramalingam, Logalakshmanan Baskaran, Ananthanarayanan Yuvaraj, Ramasundaram Thangaraj, Ramasamy Subbaiya, Balasubramani Ravindran, Soon Woong Chang, and Natchimuthu Karmegam. 2020. “Enriched Pressmud Vermicompost Production with Green Manure Plants Using Eudrilus Eugeniae.” *Bioresource Technology* 299 (March): 122578.
5. Banik, Rounak. 2018. *Hands-On Recommendation Systems with Python: Start Building Powerful and Personalized, Recommendation Engines with Python*. Packt Publishing Ltd.
6. Chen, Zhen, and Xiaoxuan Yu. 2020. “Adoption of Human Personality Development Theory Combined With Deep Neural Network in Entrepreneurship Education of College Students.” *Frontiers in Psychology* 11 (July): 1346.
7. Chung, Kyungmi, Jin Young Park, Kiwan Park, and Yaeri Kim. 2020. “Which Visual Modality Is Important When Judging the Naturalness of the Agent (Artificial Versus Human Intelligence) Providing Recommendations in the Symbolic Consumption Context?” *Sensors* 20 (17). https://doi.org/10.3390/s20175016.
8. Corker, Katherine S., M. Brent Donnellan, Su Yeong Kim, Seth J. Schwartz, and Byron L. Zamboanga. 2017. “College Student Samples Are Not Always Equivalent: The Magnitude of Personality Differences Across Colleges and Universities.” *Journal of Personality* 85 (2): 123–35.
9. Ezhilarasan, Devaraj, Thangavelu Lakshmi, Manoharan Subha, Veeraiyan Deepak Nallasamy, and Subramanian Raghunandhakumar. 2021. “The Ambiguous Role of Sirtuins in Head and Neck Squamous Cell Carcinoma.” *Oral Diseases*, February. https://doi.org/10.1111/odi.13798.
10. Gibson, Caitlin M., and Shara Elrod. 2018. “Students’ versus Residency Programs' Perceptions of a High-Quality PGY1 Residency Applicant.” *Currents in Pharmacy Teaching & Learning* 10 (2): 137–45.
11. Gopalakrishnan, R., V. M. Sounthararajan, A. Mohan, and M. Tholkapiyan. 2020. “The Strength and Durability of Fly Ash and Quarry Dust Light Weight Foam Concrete.” *Materials Today: Proceedings* 22 (January): 1117–24.
12. Hannah R, Pratibha Ramani, WM Tilakaratne, Gheena Sukumaran, Abilasha Ramasubramanian, and Reshma Poothakulath Krishnan. 2021. “Author Response for ‘Critical Appraisal of Different Triggering Pathways for the Pathobiology of Pemphigus vulgaris—A Review.’” Wiley. https://doi.org/10.1111/odi.13937/v2/response1.

13. Harries, Aaron J., Carmen Lee, Lee Jones, Robert M. Rodriguez, John A. Davis, Megan Boysen-Osborn, Kathleen J. Kashima, et al. 2021. “Effects of the COVID-19 Pandemic on Medical Students: A Multicenter Quantitative Study.” *BMC Medical Education* 21 (1): 14.
14. Kavarthapu, Avinash, and Kaarthikeyan Gurumoorthy. 2021. “Linking Chronic Periodontitis and Oral Cancer: A Review.” *Oral Oncology*, June, 105375.
15. Liu, Fang. 2021. “Personalized Recommendation System of Resource Database for College Students’ Innovation and Entrepreneurship.” *2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. https://doi.org/10.1109/icmtma52658.2021.00140.
16. Menon, Soumya, Happy Agarwal, S. Rajeshkumar, P. Jacquline Rosy, and Venkat Kumar Shanmugam. 2020. “Investigating the Antimicrobial Activities of the Biosynthesized Selenium Nanoparticles and Its Statistical Analysis.” *BioNanoScience* 10 (1): 122–35.
17. Muthukrishnan, Sivaprakash, Haribabu Krishnaswamy, Sathish Thanikodi, Dinesh Sundaresan, and Vijayan Venkatraman. 2020. “Support Vector Machine for Modelling and Simulation of Heat Exchangers.” *Thermal Science* 24 (1 Part B): 499–503.
18. Ng, Yiu-Kai, and Jane Linn. 2017. “CrsRecs: A Personalized Course Recommendation System for College Students.” *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*. https://doi.org/10.1109/iisa.2017.8316368.
19. Rathnavel, Jayanti, Student MTech, Department of Computer Engineering, K J Somaiya College of Engineering, and Mumbai. 2017. “Personalized Book Recommendation System.” *International Journal Of Engineering And Computer Science*. https://doi.org/10.18535/ijecs/v6i4.61.
20. Ricci, Francesco, Lior Rokach, and Bracha Shapira. 2015. *Recommender Systems Handbook*. Springer.
21. Richardson, Eric, Kathleen A. Ryan, Robert M. Lawrence, Christopher A. Harle, Alyson Young, Melvin D. Livingston, Amit Rawal, and Stephanie A. S. Staras. 2021. “Perceptions and Knowledge About the MenB Vaccine Among Parents of High School Students.” *Journal of Community Health*, January. https://doi.org/10.1007/s10900-020-00954-1.
22. Sarode, Sachin C., Shailesh Gondivkar, Gargi S. Sarode, Amol Gadbail, and Monal Yuwanati. 2021. “Hybrid Oral Potentially Malignant Disorder: A Neglected Fact in Oral Submucous Fibrosis.” *Oral Oncology*, June, 105390.
23. Sekar, Durairaj, Deepak Nallaswamy, and Ganesh Lakshmanan. 2020. “Decoding the Functional Role of Long Noncoding RNAs (lncRNAs) in Hypertension Progression.” *Hypertension Research: Official Journal of the Japanese Society of Hypertension*.
24. Shin, Kyung-Shik, Taik Soo Lee, and Hyun-Jung Kim. 2005. “An Application of Support Vector Machines in Bankruptcy Prediction Model.” *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2004.08.009.
25. Wang, Xiaokui. 2020. “College Students’ Innovation and Entrepreneurship Resources Recommendation Based on Collaborative Filtering and Recommendation Technology.” *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1533/2/022013.

TABLES AND FIGURES

Table 1. Accuracy and Precision achieved during evaluation of College Student prediction using test dataset with K-MEAN algorithm and Support vector machine technique for different iterations

ITERATIONS	**ACCURACY**		**PRECISION**	
	K-Mean	**SVM**	**SVM**	**K-MEAN**
1	46.00	88.40	84.10	46.10
2	44.30	91.21	86.05	47.02
3	45.75	87.70	84.30	45.73
4	46.04	87.05	86.78	47.31
5	44.76	87.36	84.84	46.27
6	45.03	90.32	86.53	47.10
7	46.98	87.43	87.43	46.32
8	44.86	89.23	82.54	44.29
9	45.79	86.43	85.63	44.82
10	46.34	88.34	87.34	45.21

Table 2. Experimental analysis in Google Colab for Accuracy, Precision,Recall,F1 Score and F2 Score for K-Mean and SVM. Support vector machine provides better Accuracy (86.66%) and Precision (76.66%) than K-Mean

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**F2 Score**
SVM	86.6667	76.6667	1.00000	0.6444 4	0.819209
K-Mean	46.8889	47.6667	0.965517	0.88889	0.93333

Table 3. Statistical Analysis of Mean, Standard Deviation and Standard Error of Precision and Accuracy of K-Mean and SVM algorithms. There is a statistically significant difference in precision and accuracy values between the algorithms. Support vector machine has higher precision (68.4%) and accuracy (85.10%) than K-Mean.

GROUP	N	Mean	Std.Deviation	Std.Error Mean
ACCURACY K-MEAN SVM	10 10	50.2000 85.1000	38.68620 2.55821	12.23365 .80898
PRECISION K-MEAN SVM	10 10	51.1000 68.4000	39.12501 20.25504	12.37242 6.40521

Table 4. Comparison of the Significance level for K-Mean and SVM algorithms with value p < 0.05. Both K-Mean and SVM have a significance level less than 0.05 with a 95 % confidence interval

	Levene's Test for Equality of Variance		T-test for Equality of Means						
								95% Confidence Interval of the Difference	
	F	Sig	t	df	Sig(2-tailed)	Mean Difference	Std.Error Difference	Lower	Upper
Accuracy	119.505	.000	-2.847 -2.847	18 9.079	.011 .019	-34.90000 -34.90000	12.26037 12.26037	-60.65808 -62.59827	-9.14192 -7.20173
Precision	24.782	.000	-1.242 -1.242	18 13.50	.230 0.235	-17.30000 -17.30000	13.93210 13.93210	-46.57025 -47.28534	-11.97025 12.68534

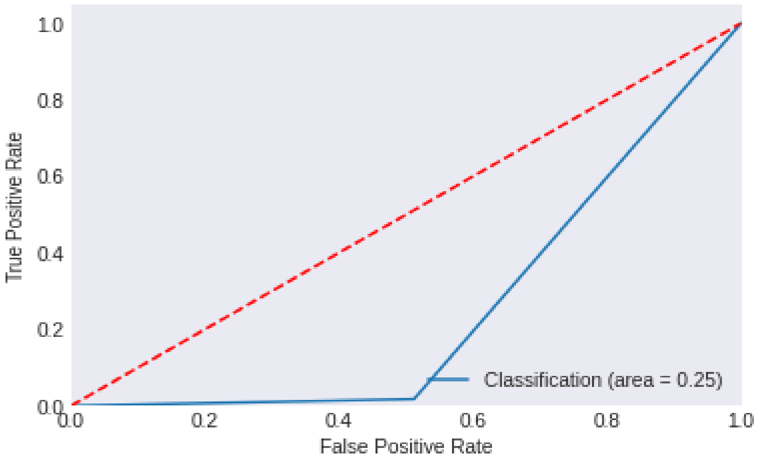

Fig. 1. Receiving Operating characteristic (ROC) Curve for SVM

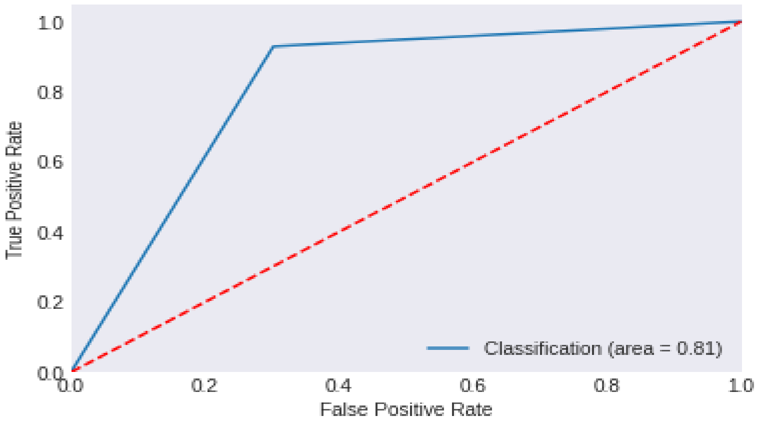

Fig. 2. Receiving Operating characteristic (ROC) Curve for K-Means

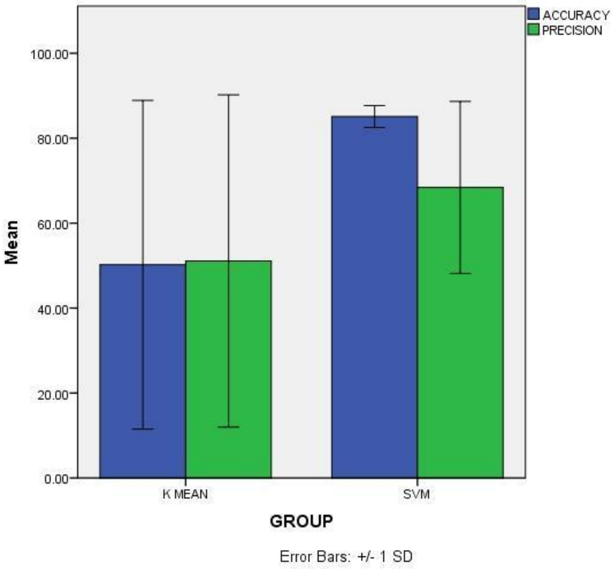

Fig. 3. Bar Chart representing the comparison of mean accuracy of College Recommendation system prediction using K-Mean and Support vector machine algorithms. SVM produces better accuracy and more consistent results X-axis: K-Mean vs SVM. Y-axis: Mean Accuracy ± 1 SD.