# MINING EDUCATIONAL DATA USING MKNN CLASSIFICATION TO DECREASE THE DROPOUT RATE IN TECHNICAL EDUCATION

*Revathi.P*[1], *M. Jaikumar*[2]

**ABSTRACT**

To increase the student graduation rates and decrease the student dropout rate some predictive measures should be taken. This can be achieved through data mining methods. Educational Data mining concerns with developing methods for discovering knowledge from data that come from educational domain. Attribute selection is done by MATLAB; the data processing tool mainly works in prediction and classification of knowledge. Here Modified K-Nearest Neighbor, k means, Gaussian Mixture Model, fuzzy c means, classification is used to classify and neural network is used to compare the MKNN, GMM, Fuzzy C means classification algorithms and predict the algorithm which is highly accuracy and time computing process. The proposed KNN classification is called Modified K-Nearest Neighbor (MKNN). The main idea is to classify an input query according to the most frequent tag in set of neighbor tags. MKNN can be considered a kind of weighted KNN, so that the query label is estimated by weighting the neighbors of the query. The procedure computes the frequencies of the same labeled neighbors to the total number of neighbors.

*Keywords* - Data Mining, MKNN

[1]M.Phil Research Scholar, Sri Ramakrishna Mission Vidyalaya College of Arts and Science,Coimbatore – 641 020, Tamilnadu, India. reva.yellow@gmail.com

[2]Assistant Professor & Head, Dept of Computer Applications (UG), Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Coimbatore – 641 020, Tamilnadu, India jai2911@gmail.com

## I. INTRODUCTION

Education is a key factor for achieving a long-term economic progress. Nowadays, research interests have been increasing in the field of Education using Data mining. We used educational data mining to predict student's dropout.

Student dropout is a demanding task in higher education [1] and it is reported that about one fourth of students dropped college after their first year.

Dropout rate of student is a big risk for both the educational institution as well as for the upcoming career of student. It is the need of today's era to find out reason behind increasing dropout rate of students from the courses.

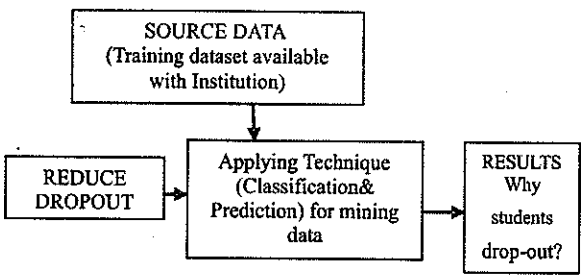


**Figure :1 Requirement and Result Model**

Data Mining is a technology that elaborates on knowledge discovery. To analysis of those relationships can be evaluated analytical, artificial intelligence and machine intelligence to enable users to quote and classify greater enlighten and successive power than simple query and study approaches.[9].

Many Polytechnic students drop the course after first year. Based on review of the literature, some of the details are finding the student failure. These motives were then generally collected into three major categories:

1. Result
2. Attendance
3. Family Income

The cause of student drop out is often termed as the antecedent of dropout because it refers to the pivotal event which indicate to student failure.. during this, for all that, is the conclusion of a much longer process of leaving college that began long before the date that a student actually discontinues attendance. Student dropout has become an indication of academic performance and enrolment management. In our study, we apply machine learning algorithm to analyze and extract information from existing student data to establish predictive model. The predictive model is then used to determine among new incoming first year students those who will dropout from a college.

## II. Background And Relatedwork

Tinto [7] developed the most popular model of retention studies. According to Tinto's Model, withdrawal process depends on how students interact with the social and academic environment of the institution.

To understand the factors influencing university student retention, Superby et. al. [2] used questionnaires to collect data including personal history of the student, implication of student behavior and perceptions of the student. The authors applied different approaches such as decision tree, random forests, neural networks, and linear discriminate analysis to their questionnaires. However, possibly because of the small sample size, the prediction accuracy is not very good.

A number of Open Distance Learning institutions have carried out dropout studies. Some notable studies have been undertaken by the British Open University (Ashby [3]; Kennedy & Powell [4]). Different models have been used by these researchers to describe the factors found to influence student achievement, course completion rates, and withdrawal, along with the relationships between variable factors.

Yadav, Bharadwaj and Pal [5] conducted study on the student retention based by selecting 398 students from MCA course of VBS Purvanchal University, Jaunpur, India. By means of classification they show that student's graduation stream and grade in graduation play important role in retention. Khan [6] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-

economic status had relatively higher academic achievement in science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Ayesha, Mustafa, Sattar and Khan [10] describe the use of k-means clustering algorithm to predict student's learning activities. The information generated after the implementation of data mining technique may be helpful for instructor as well as for students.

### III. Modified K-Nearest Neighbour

The main plan of the given technique is assignment the category label of the information consistent with K valid data points of the training set. In additional, first, possible values of all knowledge samples within the training set is calculated. Then, a weighted KNN is performed on any existing look at samples. Fig. 1 It shows the pseudo code of the MKNN algorithmic rule.

```
Output_label := MKNN ( plaything , test_sample )
Begin
For i := one to train_size
Validity(i) := work out Validity of i-th sample;
End for;
Output_label:= Weighted_KNN (Validity,test_sample);
Return Output_label ;
End.
```

Fig. 1. Pseudo-code of the MKNN Algorithm

In the remains of this section the MKNN technique is delineate well, respondent the queries, the way to work out the validity of the points and the way to work out the ultimate category label of take a look at samples.

A. Validity of the Train Samples

In the MKNN algorithmic rule, each sample in plaything should be valid at the primary step. The validity of every intention is calculated regular with its neighbors. The validation method is carrying out for all train samples once. When assignment the validity of every train sample, it's used as a lot of data regarding the points.

To validate a sample purpose within the training set, the H nearest neighbors of the aim is taken into detail and between the H nearest neighbors of a training sample x, validity(x) calculate the size of points with a similar label to the label of x. The method that is prepared to work out the validity of each point in plaything is (1).

$$\text{Validity(x)}=\frac{1}{H}\sum_{i=1}^{H} S(lbl(x), lbl(Ni(x))) \quad (1)$$

Where H is that the variety of thought -about neighbors and lbl(x) returns verity category label of the train sample x. also, Ni(x) stands for the ith nearest neighbor of the intent x. The perform S takes into consideration the similarity between the purpose x and also the ith nearest neighbor. The (2), defines this perform.

$$\text{S(a,b)}= \begin{cases} 1 & a=b \\ 0 & a\neq b \end{cases} \quad (2)$$

*B. Applying Weighted KNN*

Weighted KNN is one amongst the variations of KNN technique that uses the K nearest neighbors, in spite of their categories, then again uses weighted votes from every sample instead of a straightforward majority or plurality option rule. Every of the K samples is given a weighted vote that's typically adequate to some decreasing perform of its distance from the unknown sample. for instance, the vote may set be adequate to 1/(de+1), wherever Delaware is geometer distance. These weighted votes are then summed for every category, and also the category with the biggest total vote is selected. This distance weighted KNN technique is suspect almost like the window technique for estimating density functions. for instance ,employing a weighted of 1/(de+1) is admire the window technique with a window perform of 1/(de+1) if K is chosen adequate to the entire variety of coaching samples [8].

In the MKNN technique, initial the burden of every neighbor is computed mistreatment the 1/(de+0.5). Then, the validity of that coaching sample is increased on its raw weight that is predicated on the geometer distance. With in the MKNN technique, the burden of every neighbor sample comes consistent with (3).

$$W(i) = \text{Validity}(i) \times \frac{1}{d_e+0.5} \qquad (3)$$

Wherever W(i) and Validity(i) represent the burden and also the validity of the ith nearest sample within the plaything. This method has the result of giving bigger importance to the reference samples that have bigger validity and closeness to the take a look at sample. So, the choice is a smaller amount laid low within the reference samples that aren't terribly stable within the feature are compared with alternative samples. In alternative hand, the multiplication of the validity live on live will overcome the weakness of any distance based weights that have several issues within the case of outliers. So, the planned MKNN algorithmic rule is considerably stronger than the standard KNN technique that is predicated simply on distance.

**Gmmalgorithm**

**Model-Based clustering**

There's another way to deal with clustering problems: a model-based approach, which contains using particular models for clusters and attempting to optimize the fit between the data and the model. In exercise, every cluster can be mathematically represented by a parametric distribution, like a Gaussian (continuous) or a Poisson (discrete). The entire data set is then modeled by a mixture of the particular distributions. An individual distribution practice to model a specific cluster is frequently referred to as a component distribution.

**FUZZY C MEANS CLUSTERING**

In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, reasonably than apply entirely just in one cluster. Thus, percentage on the end of a cluster can be in the cluster to a lesser degree than points in the center of cluster.

## IV. Data Mining Process

### Data Preparations

The data used in this project contains polytechnic students information collected from the Nanjiah Lingammal Polytechnic College, Mettupalayam for a period of three years

in period from 2011-12 to 2013-14. The polytechnic student's data set consists of 700 records.

### Data selection and Transformation

The important field only selected which were required for Data Mining. Questionnaire form filled by the student. The student enter their data (category, gender etc), past performance data (SSC or 10th marks, HSC or 10 + 2 exam marks etc.), address and mobile number. A few variables are chosen. While some of the data for the variables was extracted from the database.

The values for certain variables were defined for the present investigation as follows:

**Branch** – The courses offered by Nanjiah Lingammal Polytechnic College are Computer Science (*CSE*), Civil Engineering(CE), Mechanical Engineering (*ME*), Electrical Engineering (*EE*), Electronics and Communication Engineering (*ECE*) and Automobile Engineering (AE).

**HSC** - Students grade in Higher Secondary School . Students the one in state board occur for six subjects each carry 100 marks. Grade are allocated to all students usage for the following mapping O – 90% to 100%, A – 80% - 89%, B – 70% - 79%, C –60% - 69%, D – 50% - 59%, E – 40% - 49%, and F - < 40%.

**SSLC** - Students grade in Senior Secondary education. Students appear for SSLC exam that contains five subjects each carry 100 marks. Grade are allotted to all students using following grades O – 90% to 100%, A – 80% - 89%, B – 70% -79%, C – 60% - 69%, D – 50% - 59%, E – 40% -49%, and F - < 40%.

**Atype** - The admission type which may be through Direct First Year or Lateral Entry.

**Med** – This paper study covers only the Polytechnic colleges of Tamil Nadu state of India. Here, medium of instructionsre Tamil or English.

**TABLE I**

**Student Related Variables**

| Variables | Description | Possible Values |
|---|---|---|
| Branch | Students Branch | {CE, ME,AE,EEE, ECE,CT} |
| Sex | Students Sex | {Male, Female} |
| HSC | Students Marks inHigh School | {O – 90% -100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%} |
| SSLC | Students Marks in Higher Secondary | {O – 90% -100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, |

| | | |
|---|---|---|
| | | D – 50% - 59%, E – 40% - 49%, F - < 40% } |
| Adtype | Admission Type | {Lateral Entry (LE), Direct} |
| Med | Medium of Teaching | {Tamil, English} |
| LLoc | Living Location of a Student | {Village,Town, Taluk,District} |
| Hostel | Staying in Hostel | {yes,No} |
| Attpercent | Attendance Percentage | {Below 80%,75%-80%,80%-90%, 90%-100%} |
| Fsize | Family Size | {1,2,3,4} |
| FAIn | Family annual income Status | {very poor, poor,medium, high} |
| FQual | Fathers Qualification | {not educated, elementary, Secondary, Degree or UG, PG,Ph.D,NA} |
| MQual | Mother's Qualification | {not educated, elementary, Secondary, Degree or UG, PG,Ph.D,NA} |
| FOccupation | Father's Occupation | {Service, Business, Agriculture, Retired,NA, Govt, Private} |
| MOccupation | Mother's Occupation | {House-wife (HW),Service, Retired, NA,Govt,Private} |
| DOB | Date of Birth | Student date of birth |
| ExtraCurr | Extra Curricular Activities | NSS,Sports |

The data is collected from the Polytechnic Institution. Data was analyzed for students entering Polytechnic colleges from the academic year 2012-2013 to 2013-2014.

Academic performance of an Institution can be improved knowing the reasons for student dropout. A number of fields are considered for predicting the academic outcome of the student.

## V. Results

MATLAB was used for evaluation of information in directory. The data set was gathered and examined using k means clustering, Fuzzy c means Clustering, Modified K Nearest Neighbor Algorithm and Gaussian Mixture Model, through this algorithm we categorize the important attributes based on that we generate result. It point out Modified K-Nearest Neighbor algorithm is the best classifier as compare to Fuzzy C Means, Gaussian Mixture Model, K Means Clustering for present data set to predict the 'student's dropout status.

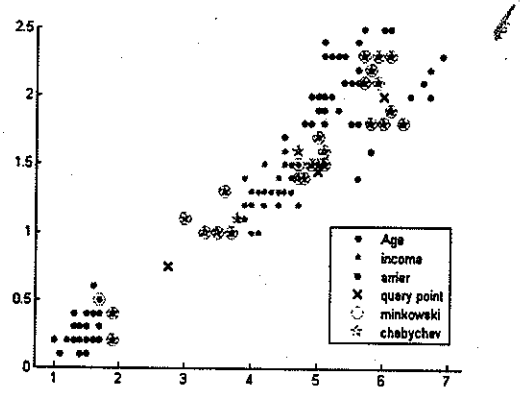


**Figure : 2 Modified K-Nearest Neighbor**

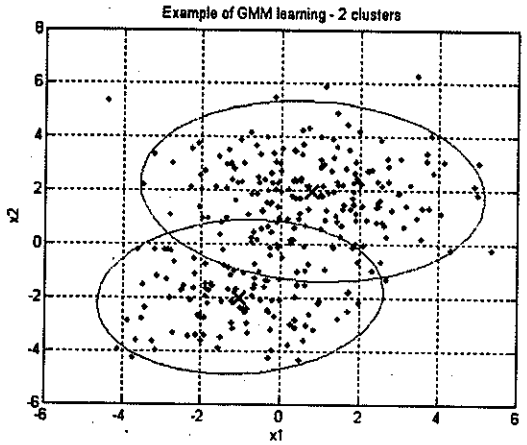


**Figure : 3 Gaussian Mixture Model**

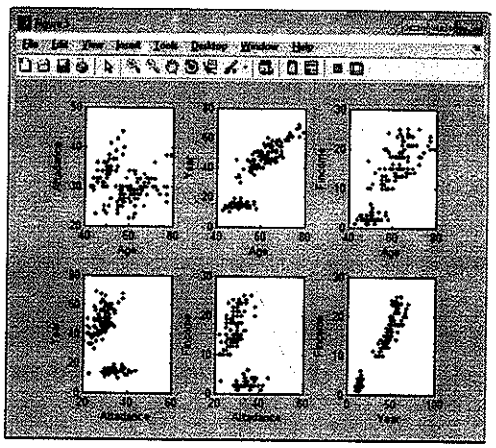

**Figure : 4 Fuzzy C means Algorithm**

## COMPARISON OF ALGORITHM USING NEURAL NETWORKS

To compare the Fuzzy C Means, Gaussian Mixture Model, and MKNN Classification to predict the accuracy of best algorithm and to predict the shortest time taken for the execution. Compare with those algorithms using Neural network to predict the MKNN classification is the best algorithm with high accuracy and less time taken to execute the result.

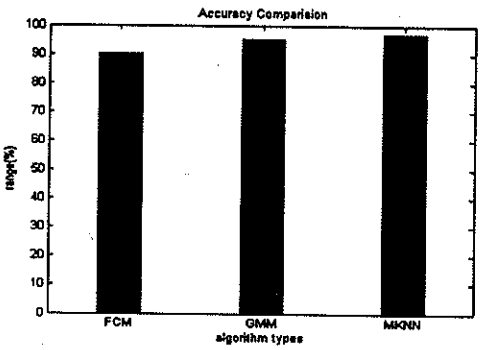


**Figure : 5 Accuracy Calculation for Classification Algorithm**

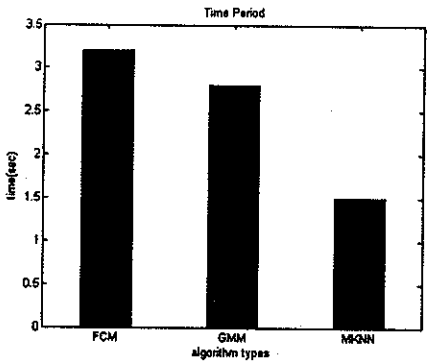


**Figure : 6 Classification Algorithm for Time Period**

## VI. Conclusion

College dropout deeply causes harm to a successful future. College dropout only continues to the cycle of poverty. A good education is key to ending poverty within our youth. Implementation of effective early prevention is essential to end college dropout.

During finding of dropout students, many attributes have been tested, and some of them are found efficient on the prediction.

Predicting dropout is great concern to the technical education. In recent times, data mining can be avail in a technical education system.

The result indicates that MKNN algorithm is best classifier compare to Fuzzy C Means, GMM. This algorithm is to find those students which require special care to decrease dropout rate. The data produced will be useful for better preparation and execution of educational program and organization under computable state to increase the admission of students in Technical Education Courses.

Neural network is used to compare the MKNN algorithm with the other algorithms such as Fuzzy C Means, Gaussian Mixture Model and predicted the result accuracy wise and time basis. The MKNN algorithm is taking less time to process the data set and highly accuracy.

## References


[1] Tinto, V., *"Research and practice of student retention: What next, College Student Retention: Research"*, Theory, and Practice, 8(1), 1-20, 2006.

[2] Superby, J.F., Vandamme, J-P., Meskens, N., *"Determination of factors influencing the achievement of the first-year university students using data mining methods."* Workshop on Educationa, 2006.

[3] Ashby, A., *"Monitoring Student Retention in the Open University: Detritions, measurement, interpretation and action"*. Open Learning, 19(1), 65-78, 2004.

[4] Kennedy, D., & Powell, R., *"Student progress and withdrawal in the Open University"*. Teaching at a Distance, 7, 61-78, 1976.

[5] Yadav S. K., Bharadwaj B. K. and Pal S., *"Mining Educational Data to Predict Student's Retention: A Comparative Study"*, International Journal of Computer Science and Information Security (IJCSIS), Vol. 10, No. 2, Feb 2012, pp 113-117.

[6] Z. N. Khan, *"Scholastic achievement of higher secondary students in science stream"*, Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[7] Tinto, V., *"Leaving College: Rethinking the cause and cure of student attrition"*. Chicago: University of Chicago Press, 1993.

[8] E. Gose, R. Johnsonbaugh and S. Jost, Pattern Recognition and Image Analysis, Prentice Hall, Inc., Upper Saddle River, NJ 07458, 1996.

[9] Saurabh Pal, *"Mining Educational Data to Reduce Dropout Rates of Engineering Students"*, MECS PRESS, 2012.

[10] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M.Inayat Khan, *"Data mining model for higher education system"*, Europen Journal of Scientific Research, Vol.43,No.1, pp.24-29, 2010.

**AUTHOR'S PROFILE**

**P.Revathi** completed MCA and currently Pursuing M.Phil in Computer Science at Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Coimbatore-20. She working as a Lecturer in Nanjiah Lingammal Polytechnic College, Mettuapalayam. Her Teaching Experience is 8 years. Her area of interest is Data mining Web mining

**M.Jaikumar** completed MCA at Bharathiar University, M.Phil at Bharathiar University and currently pursuing Ph.D in Sri Ramakrishna Mission Vidyalaya College of Arts and Scienc, Coimbatore-20. He working as Assistant Professor & Head of the Dept. of Computer Applications Department (UG) in Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Coimbatore-20. His teaching experience is 10 years. His area of interest is Software Engineering and Data Mining.