International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Educational Data Mining: A Survey of Analyzing Student Academic Performance Methods

K.D. Purani[1*], M.B. Chaudhary[2]

[1,2] Department of Computer Engineering & Information Technology, Government Engineering Collage, Modasa, India

DOI: https://doi.org/10.26438/ijcse/v7i2.832838 | Available online at: www.ijcseonline.org

Accepted: 16/Feb/2019, Published: 28/Feb/2019

***Abstract*—** Over the past decade, there has been a fast development in the advanced education system which prompts an enormous amount of data. Predicting students' performance turns out to be all the more difficult because of this enormous measure of information in educational databases. However, this data from the educational department acts as a gold mine for institutions and also encourages the analysts and researchers to make a framework that can improve the general educating and learning process. Analysts and researchers apply Data mining techniques on educational data to explore it. Educational Data Mining helps in a big way to answer the issues of predictions and grouping of not only students but also the other stakeholders of education sectors. This paper talks about the utilization of different Data Mining techniques and tools that can be adequately utilized in noting the issues of predictions of students' performance and their grouping.

***Keywords*—** Data mining, Data Mining Techniques, Educational Data Mining, Student Performance Prediction.

I. INTRODUCTION

Data mining is the extraction of interesting, non-trivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amount of data [1]. An application of DM in education is called Educational Data Mining (EDM). EDM can be defined as "An emerging discipline concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students, and the settings which they learn in" [2]. EDM works by extracting and interpreting the raw data coming from the educational system that could potentially have a great impact on student performance, student retention rate, university success rate. Impact of EDM in education research is great in recent times, as education is delivered in various ways like E-learning, Intelligent tutorial systems (ITS), Offline Education.
The organization of the paper is as follows:

Section I contains the introduction of Educational data mining, objective, component, environment of Educational data mining, contains information about educational data.
Section II talks about Educational data mining methods.
Section III contains information about tools which are useful for EDM.
Section V contains the related work of data mining methods for the prediction of students' performance.
Section VI is the conclusion of the survey.

A. OBJECTIVE OF EDM

The fundamental Objective of using EDM is to enhance the processes of education and learning. Research objectives, such as gaining a deeper understanding of the teaching and learning phenomena, identifying weaker students at an early stage, recommending them extra Module/Course, identifying the factors affected to student success the most, and analysing student behaviour in e-learning system. Using EDM, we can identify the factors which need improvement to increase the university success rate as well as student success rate, which is a major objective of any educational institutions.

B. COMPONENTS OF EDM

Educational data mining touches and influences various aspects of the education industry. The major components of EDM are as follow.

1. Stakeholders of the education system.
2. Educational environment.
3. Educational data.
4. Data mining tools and techniques for EDM [3].

Stakeholders: Stakeholders of the education system can be majorly categorized into the following:

Learners: Learners/Students are involved directly in the process of learning; they are the most important and impacted stakeholders.

Faculties: Educators/Teachers are benefitted as they can figure out which student requires extra support. Educators can analyse the data and decide the most commonly made errors.

Parents: Parents are part of the secondary group. They are liable for helping their kids to get them to enrol in the most suitable courses for them.

Course Researchers and Educational Developers: They are the people who structure and change the course. They are responsible for the development of education.

Administrators: They can also be called as hybrid users. They are in charge of different authoritative choices; for example, infrastructure development and employing the expert faculty. EDM is valuable for compelling usage of assets.

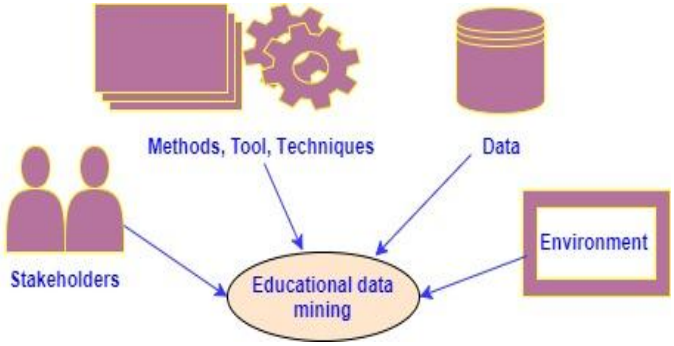

Figure 1. EDM Components

C. EDM ENVIRONMENT

In today's era, majorly two types of educational environments exist which can be explored and analysed by Educational Data Mining**:** 1. Traditional Education and 2. Computer-based Education**.** Each provides data from different sources that must be pre-processed in particular ways depending on the nature of each of them, the problems and the specific tasks to be solved.

D. EDUCATIONAL DATA

Analysis of the enormous amount of educational data is involved in decision-making and future planning. Data gathered from different sources, such as in the form of responses given by large numbers of students to various questions, LMS (Learning Management System) Data Collection, Intelligent tutors, and learning Applications. As we know, data can be generated from online and offline sources.

Offline Data: Offline data is produced through real-time situations and settings. It is also generated from traditional education, is where knowledge transfer to learners is based on face to face contact [5]. For example, traditional classroom tests, teacher-student interactions, student-to-student interactions, participation in different activities by the students, students' attendance, behaviour and attitude, data derived from different courses and various departments of any institute.

Online Data: Online data is derived from Weblogs, E-mails, E-learning and Learning Management System (LMS), Intelligent Tutoring System (ITS) and Adaptive Educational Hypermedia System (AEHS), and Publication Databases. Online data is not dependent upon any kind of geographical location [5].

II. EDM METHODS

Educational Data Mining methods come from different literature sources including data mining, machine learning, and information visualization. A point of view, proposed by Baker [6], classifies the work in EDM as follows:

A. Prediction
- Classification
- Regression

B. Clustering

C. Relationship Mining.
- Association rule mining.

Prediction is to identify unknown variables purely on the description of historical data for the same variable. Prediction derives the relationship between labelled data and output variable you need to predict for the future reference. There are general types of predictions like classification and regression.

Classification is a supervised learning technique. It helps in classifying data based on the training set and then uses that pattern to introduce the new data set, which is also known as the test set. Some popular Classification methods used in EDM are Decision Tree, Support Vector Machine, Naïve Bayes Classifier, KNN etc.

Regression is different from classification as regression predicts continuous variables. Different methods of regression are linear regression and neural networks. For example, in EDM to predict the final result of student, various parameters like Age, Gender, Attendance, Family Income, Occupation, Family Qualification can be used as predictors.

Clustering is the method that helps to group similar records together. This is an unsupervised learning approach which mainly focuses on high-dimensional data. K-Means is the most commonly used method for Clustering. For example, in EDM, clustering can be used to group students based on their learning styles such as Visual, Aural, and Kinaesthetic.

Relationship Mining expects to discover relationships between various variables in data sets with a large number of variables. This entails finding out which variables are most strongly associated with a specific variable of a particular area. Association rule mining is the most commonly used EDM method. Association rules in Educational Data Mining are used to determine remarkable and strong association rules from educational databases using support and confidence as the predefined measures [7]. Apriori algorithm is one of the most commonly and widely used methods in Association Mining [5].

III. POWERFUL TOOLS FOR EDUCATIONAL DATA MINING

Some open source tools which are used in applications of EDM are as follow:

R: R is a statistical computing software. It is an open source language used for statistical and data analysis. It can run on multiple platforms (e.g. Windows, MacOS or Linux). As of January 2019, R ranks 12th in the TIOBE index, a measure of the popularity of programming languages [8].

WEKA*:* WEKA stands for Waikato Environment for Knowledge Analysis. It is a non-propriety, freely available, and application-neutral standard for data mining projects [5]. It has several tools, algorithms and graphics methods which lead to the analysis and predictions. Most of the algorithms are inbuilt in these tools. It is widely adopted in academic and business and have an active community.

RapidMiner (YALE): Rapid Miner is a software platform which uses a client/server model with the server offered as Software as a Service or on cloud infrastructures [9].

Orange*:* The Orange is an open source tool for educational data mining researcher and it is Python-based. Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis, visualization, and Python bindings and libraries for scripting [10].

KNIME*:* Konstanz Information Miner, is an open source data analytics, reporting and integration platform. This tool is based on Eclipse and written in java [9].

IV. RELATED WORK

C. Romero and S. Ventura (2007) [11], conduct a survey of EDM research from 1995 to 2015 consisting of more than 50 papers. Up to 2005 most of EDM research papers involved Relationship Mining Methods. After 2005 prediction slowly became the most popular research area as it replaced Relationship Mining Methods.

C. Romero and S. Ventura (2007) [12], said that the decision tree models are easily understood because of their reasoning process and can be directly converted into set of IF-THEN rules.

G. Dekker (2009) [13], a research was done to predict the Electrical Engineering (EE) students drop out after the first semester of their studies or even before they enter the study program as well as identifying success-factors specific to the EE program. Data collected over the period 2000 – 2009 of 648 students. They compared the two decision tree algorithms CART and C4.5, a Bayesian classifier (BayesNet), a logistic model, a rule-based learner and the Random Forest using weka tool. The results proved that decision tree gives good accuracy.

S. Sembering and M.Zarlis (2011) [14], predict final performance of student in higher education. They collect the data from student by using questionnaire and find relationships between behavioural of student and their academic performance. Generate prediction rules using decision tree and implement the rules into SSVM algorithm to predict the students' final grade. Clustered the students into groups using kernel k-means clustering. Shows the strong correlation between mental condition of students and their final academic performance. For graph they used rapid miner.

Saurabh Pal (2012) [15], implemented certain data mining methodologies to find students who are likely to be dropped out from their first year B.E program. The author used Naïve Bayes classification algorithm for the accurate prediction of attributes from the existing data sets. The results exhibited that the machine learning algorithm compared the new data set with the existing and was able to establish effective dropout predictive model. The system produced almost accurate prediction in identifying the students who needed special attention to reduce drop-out rate. But the result was too short and tested only with the small dataset.

Suhem Parack, Zain Zahid and Fatima Merchant (2012) [16], discussed the application of data mining in education for student profiling and grouping. They applied Apriori algorithm to the database containing academic records of various students and tried to extract association rules in order to profile students based on various parameters like exam marks, result grades, attendance, and practical exams. They also applied K-means clustering to the same data set to group the students. The conclusion stated the use of data mining techniques easily clustered the students, identified hidden patterns about their learning styles, found undesirable student behaviour and performed student profiling. The results enormously deduced the manual work involved in identifying the students' tendencies, conduct and the system

in which they studied. But, this paper considered only academic parameters of the students', not the personal characteristics such as past histories, family background in order to predict if a student is prone to violence.

D. Kabakchieva (2013) [17], used the CRISP-DM (Cross-industry standard process for data mining) model to predict student academic performance. He applied a Decision Tree (J48) classifier, Bayesian classifiers such as Naïve Bayes, BayesNet, Nearest Neighbor method (IBk) and rule learners namely OneR and JRip. Decision Tree had better accuracy than the rule learner JRip and IBk. Bayesian Classifiers had the least accuracy.

D. Ahmed and I. Sayed (2014) [18], describe the most used and useful data mining technique "classification". They used Decision Tree classifier, a type of classification technique to predict the final grade of students. The data set used in this study was obtained from the educational institutions, on the sampling method of the Information system department from session 2005 to 2010. The size of the data was 1547 records.

M. Mayilvaganan and D. Kalpanadevi (2014) [19], focused on the improvement of Prediction/ classification techniques which are used to analyse the skilled expertise based on their academic performance by the scope of knowledge. The paper shows the comparative performance of C4.5 algorithm, AODE, Naïve Bayesian classifier algorithm, Multi-Label K-Nearest Neighbor algorithm to find the well-suited accuracy of classification algorithm and decision tree algorithm to analyse the performance of the students which can be experimented in Weka tool. Experimentation result concluded that Multi-labeled K Nearest Neighbor has the best accuracy of time taken in classification when compared to other techniques. It had taken less time to identify the students' performance as a slow learner, average learner, good learner, and excellent learner.

A. M. Shahiri, W. Husain, and N. A. Rashid (2015) [20], they said Neural Network, Decision Tree, SVM, K.NN, Naïve Bayes with Neural Network having the highest accuracy and Classification method the most frequent in EDM.

Elaf Abu Amrieh1, Thair Hamtini and Ibrahim Aljarah (2016) [21], proposed model evaluated the impact of student's learning behavioural features on the student's academic performance. They used filter-based feature selection methods. Classification is performed by using one of the data mining techniques i.e. Decision tree, Naïve Bayesian, ANN. For improving the performance of these classifiers, ensemble methods- Bagging, Boosting and Voting are used.

Ashwin Satyanarayana, Mariusz Nuckowski (2016) [22], they compare a single filters with ensemble filters and show that using ensemble filter works better for identifying and eliminating noisy instances. They use multiple classifiers like J48, Naïve Bayes and Random Forest. They identify association rules that influence student outcomes using a combination of rule-based techniques like Apriori, Filtered Associator and Tertius.

Febrianti Widyahastuti and Viany Utami Tjhin (2017) [23], they have done an analysis of linear regression and multilayer perceptron models using WEKA tool kit. They have used undergraduate student data in information system management. The dataset consists of 50 data records. Linear regression uses the concept of the effect of one dependent variable on one or more independent variable, here final grades are chosen as the dependent variable and posting and attendance as independent variables. The relations between the variables are plotted using a scatter plot graph. In the multilayer perceptron model, the concept of neural networks is being used, which is based on weighted connections. Each node have a weight associated with it which is then multiplied by the input node to generate output prediction. They have concluded that the multilayer perceptron has more accuracy than the linear regression model. But it takes more time for processing larger datasets.

Xiaofeng Ma and Zhurong Zhou (2018) [24], they have used a decision tree based model and SVM. The dataset was taken from the UCI Machine learning Repository. They propose a new concept: features dependencies, and use the grid search algorithm to optimize DT and SVM, in order to improve the accuracy of the algorithm. In the decision tree, Information gain is used for the splitting process of the node. The information gain is calculated using the concept of entropy. The partition continues until there are no data left the leaf node contains the label of the data. After the initial model construction, test data is used to test the model.

Mrinal Pandey and S. Taruna (2018) [25], proposed an ensemble-based Decision Support System (DSS) for the prediction of students' Performance using ensemble methods. Classifiers Naïve Bayes, K-Nearest Neighbor (IBK), and Decision Tree are combined and a novel hybrid ensemble of classifiers is proposed to predict the performance of the students at different stages during their engineering. The gain ratio measure is used for attribute selection. The WEKA toolkit is used for the analysis.

Pooja Kumari, Praphula Kumar Jain and Rajendra Pamula (2018) [26], they evaluate the impact of student's learning behavioural features on the student's academic performance. They have used four classifiers: ID3, Naïve Bayes, K-Nearest Neighbour (KNN), Support vector machine (SVM). For improving the performance of classifiers, they have used

ensemble methods- Bagging, Boosting and voting to improve the accuracy of the student performance model.

Olugbenga Wilson Adejo and Thomas Connolly (2018) [27], Propose multi-model heterogeneous ensemble approach in which three data sources were used to develop seven different models with three different classification algorithms: DT, ANN, and SVM. In addition, the stacking ensemble of the models was done. The result implies that the proposed multi-data source and the use of an ensemble of classifiers had achieved statistically better performance than any other methods, creating diversity by using different combinations of the three data sources as well as classifiers.

Table 1. Comparative Study of Different Techniques.

No	Ref	Technique	Tool	Algorithm	Dataset
1	12	Classification	Java	Statistical Classifier, DecisionTree, RuleInductio, Fuzzy Rule Learning, Neural Networks.	Collected data from seven Moodle courses with Cordoba University students.
2	13	Classification	WEKA	Decision Tree, Bayesian classifier, logistic model, Rule-based learner (JRip) and Random Forest.	Data collected over the period 2000 –2009 of 648 students.
3	14	Clustering, Classification	Rapid miner	Decision Tree, SSVM, and kernel k-means.	Data collected from student by using questionnaire.
4	15	Classification	WEKA	Naïve Bayes	Data obtained from VBS Purvanchal University, for Institute of Engineering and Technology
5	16	Association rule mining, Clustering	WEKA	Apriori algorithm, K-means algorithm.	Use student academic record file.
6	17	Classification	WEKA	DecisionTree, Naïve Bayes, BayesNet, Nearest Neighbor and Rule learners OneR & JRip.	Data provided by university technical staff.
7	18	Classification	WEKA	Decision Tree.	Data was obtained from the educational institutions from Information system department.
8	19	Classification	WEKA	C4.5, Naïve Bayes, AODE, and Multi-labeled K-Nearest Neighbor.	Over all Semester exam Percentage were collected.
9	20	Classification	-	Neural Netwrok, DecisionTree, SVM, K.NN, Naïve Bayes with Neural Network.	-
10	21	Classification, Feature selection	WEKA	DecisionTree, Naïve Bayesian, ANN, Bagging, Boosting and Voting.	Student's Academic Performance Dataset From kaggle repository.
11	22	Classification, Association rule mining.	-	J48, Naïve Bayes and Random Forest, Apriori, Filtered Associator.	(a) UCI Student Performance dataset (b) New York City College of Technology CST Introductory course dataset.
12	23	Classification.	WEKA	Linear regression, Multilayer perceptron with Neural Networks.	The data from e-learning logged-post in discussion forum and attendance.
13	24	Classification	WEKA	Decision Tree and Support Vector Machine.	Data from UCI repository.
14	25	Classification	WEKA	Naïve Bayes, K-Nearest Neighbor, and Decision Tree,	The data from an engineering college from India.

				Ensemble methods	
15	26	Classification	WEKA	ID3, Naïve Bayes, KNN, Support Vector Machine.	Data from UCI repository.
16	27	Classification	Rapid miner	DT, ANN and SVM, Stacking Ensemble.	Data was collected from questioners.

V. CONCLUSION

Educational data mining affects numerous parts of the education industry and bound to be extremely beneficial in the visualization of facts, forecasting student performance, predicting students profiling, planning, and scheduling. Predicting students' performance is mostly useful to help the educators and learners improving their learning and teaching process. This paper has reviewed previous studies on EDM with various analytical methods. Most of the researchers have used prediction techniques, the classification method is frequently used in an educational data mining area. Under the classification techniques, Decision Tree, Naïve Bayes, KNN are highly used by the researchers for predicting students' performance. To enhance accuracy further ensemble methods applied. Future Research will concentrate on Composing an effective predictive model and Study of Parameters that influence the Teaching-Learning process most. By EDM, We can improve the quality of Education.

ACKNOWLEDGMENT

We would like to show our gratitude to Gujarat Technological University for supporting this research. I would like to thank my guide, Prof. M.B Chaudhary, Head of Computer Engineering Department, Government Engineering College, Modasa for providing continuous encouragement through a relaxed approach and support and proper guidance for holding me to the higher standard.

REFERENCES

[1] Sethunya R Joseph, Hlomani Hlomani, Keletso Letsholo, *"Data Mining Algorithms: An Overview"*, International journal of Computers and Technology, Vol.15, Issue.6, pp.6806-6813, **2016.**

[2] S. A. Kumar and M. Vijayalakshmi, "A Novel Approach in Data Mining Techniques for Educational Data" (ICMLC 2011), pp.152–154, **2011.**

[3] Vandna Dahiya, *"A Survey on Educational Data Mining"*, International Journal of Research in Humanities, Arts and Literature, Vol. 6, Issue.5, pp. 23-30, **2018.**

[4] Iti Burman, Subhranil Som, Mayank Sharma, **"**Enhancing Student Learning Behavior Using EDM And Psychometric Analysis" In International Conference on Reliability, Infocom Technologies and Optimization (ICRITO 2017), pp.20-22, **2017.**

[5] Abdulmohsen Algarni, *"Data Mining in Education"* International Journal of Advanced Computer Science and Applications, Vol. 7, Issue. 6, pp. 456-461, **2016.**

[6] Ryan Baker, "Mining Data for Student Models." Advances in Intelligent Tutoring Systems, pp. 323-338, **2010.**

[7] Zailani Abdullaha , Tutut Herawanb , Noraziah Ahmadb , Mustafa Mat Derisc "Mining significant association rules from educational data using critical relative support approach" Procedia - Social and Behavioral Sciences, Vol. 28 , pp.97-101, **2011**.

[8] https://en.wikipedia.org/wiki/R_(programming_language)

[9] T.Thilagaraj, Dr.N Sengottaiyan "A Review of Educational Data Mining in Higher Education System" In Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering, pp. 349-358, **2017.**

[10] Kalpana Rangra, Dr. K. L. Bansal, *"Comparative Study of Data Mining Tools"* , International Journal of Advanced Research in Computer Science and Software Engineering , Vol.4, Issue.6, pp. 216-223, **2014**.

[11] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications, vol. 33, Issue. 1, pp. 135–146, **2007.**

[12] C. Romero, S. Ventura, P. G. Espejo, C. Herv "Data mining algorithms to classify students, in: Educational Data Mining 2008", EDM 2008, **2008.**

[13] G. Dekker, M. Pechenizkiy and J. Vleeshouwers. "Predicting students drop out: A case study." In Educational Data Mining 2009,pp.41-50,**2009.**

[14] S. Sembering, M.Zarlis, "Prediction of student academic performance by an application of data mining techniques", International conference on management and Artificial Intelligence (IPEDR 2011), Indonesia, Vol.6, pp.110-114, **2011.**

[15] Dr. Saurabh Pal, *"Mining Educational Data Using Classification to Decrease Dropout Rate of Students",* International Journal Of Multidisciplinary Sciences And Engineering, Vol. 3, Issue. 5, pp.35-39, **2012.**

[16] Suhem Parack, Zain Zahid, Fatima Merchant, " Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns", 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), India ,**2012.**

[17] D. Kabakchieva, *"Predicting Student Performance by Using Data Mining Methods for Classification",* Cybernetics and Information Technologies- The Journal of Institute of Information and Communication Technologies of Bulgarian Academy of Sciences, Vol.13, Isse.1, pp.61-72, **2013.**

[18] D. Ahmed, I. Sayed, *"Data Mining: A Prediction for student's performance using classification method",* World journal of Computer Application and Technology, Vol.2, Issue.2, pp.43-47, **2014.**

[19] M. Mayilvaganan, D. Kalpanadevi, *"Comparison of classification techniques for predicting the performance of students academic environment", in: Communication and Network Technologies*, International Conference on, IEEE, pp. 113–118, **2014.**

[20] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," Procedia Computer Science, Vol. 72, pp. 414–422, **2015.**

[21] Elaf Abu Amrieh1, Thair Hamtini and Ibrahim Aljarah, *"Mining Educational Data to Predict Student's academic Performance using Ensemble Methods"*, International Journal of Database Theory and Application, Vol.9, Issue.8, **2016.**

[22] Ashwin Satyanarayana, Mariusz Nuckowski, *"Data Mining using Ensemble classifiers for improved Prediction of student AcademicPerformance",S*pring'2016'Mid.Atlantic'ASEE'Conference, **2016.**

[23] Febrianti Widyahastuti and Viany Utami Tjhin. "Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron", IEEE 10th International Conference on Human System Interactions (HSI), South Korea, PP.188-192, **2017.**

[24] Xiaofeng Ma and Zhurong Zhou. "Student Pass Rates Prediction Using Optimized Support Vector Machine and Decision Tree", IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), pp.209-215, **2018.**

[25] Mrinal Pandey and S. Taruna **"**An Ensemble-Based Decision Support System for the Students' Academic Performance Prediction**"** Springer, Advances in Intelligent Systems and Computing, Vol.653, pp.163-169, Singapore, **2018.**

[26] Pooja Kumari, Praphula Kumar Jain, Rajendra Pamula "An Efficient use of Ensemble Methods to predict Student Academic Performance**"** IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), USA, **2018.**

[27] Olugbenga Wilson Adejo, Thomas Connolly, "Predicting student academic performance using multi-model heterogeneous ensemble approach", Journal of Applied Research in Higher Education, Vol.10, Issue.1, PP.61-75, **2018.**

Authors Profile

Ms. Kruti D. Purani pursued Bachelor of Engineering in Information Technology from Gujarat Technological University. Currently pursuing Masters of Engineering in Computer Engineering from Government Engineering College, Modasa, India

Prof. M. B Chaudhari pursued Bachelor of Engineering and Master of Engineering in Computer Engineering. He has 28 years of teaching experience. Currently working as Head of Computer Engineering & Information Technology department in Government Engineering Collage, Modasa, India.

