ISSN: 2321-7782 (Online)
Volume 3, Issue 5, May 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Predicting Students' Academic Failure Using Data Mining Techniques

Lumbini P. Khobragade[1]
Student
Department of Computer Science
Deogiri institute of Engineering and Management studies
Aurangabad – India

Prof. Pravin Mahadik[2]
Assistant Professor
Department of Computer Science
Deogiri institute of Engineering and Management studies
Aurangabad – India

Abstract: ***This paper proposes predicting student's academic failure using Data Mining Techniques. Real time data of school or graduating students from an institute is taken and various data mining techniques (classification algorithms), such as induction rules, decision trees and naive bayes are applied on it. The results of these algorithms are being compared and optimized for foretelling which students might fail in future. We first consider all the available attributes of students, then select few best attributes and finally, rebalance the data using classification algorithms. This paper focuses on designing various methods that will help the teachers and the principal (Administrator) of the school to figure out the weak students and improve their educational standards and environment in which they learn. I propose the use of data mining procedures, because the complexity of the problem is high, data to be handled is very large and often highly unbalanced. The final objective of this paper is to detect the failure of students as early as possible to prevent them from dropping out and improve their academic performance. The outcomes are compared, the models with best results can be shown and the students who are at risk of failure can be provided with the guidance.***

Keywords: Data Mining, Educational Data Mining, Decision Trees, Induction Rules, Rebalancing Data, Classification Algorithms.

I. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing the results as useful information. The use of data mining concepts in the field of education is called as Educational Data Mining (EDM) [1]. The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research.

Many educational organizations and school administrations today, leave no stone unturned to improve their student's academic performance. In which the marks obtained by the student in the examination decide his/her future. They want to increase the number of student's getting passed in the yearly academics. The reason for this is to develop the best quality of the education process in their institute, to maintain the brand name of the organization and to educate students in a better way. In order to increase the number of students getting passed, the students that may get failed in that particular year in academics need to find firstly. This project basically aims to foretell the student's failure beforehand, so that some measures can be taken to avoid the student's failure in future.

To predict the failure of students is a complex task, as it requires large number of the data to be handled. For which the record of students, their each and every activities, academic related information need to maintain. Based on this information, it will be easier to predict the student's failure by applying data mining algorithms on it.

Data mining is the abstraction of needful data from large databases and ignoring the rest. Data mining tools predict future trends and behaviors, allowing the organizations to make proactive, knowledge-driven decision. Data mining helps the people to

make quick decisions on a situation as compared to statistical analysis. Data mining tools can easily handle large amount of data stored in datasets, then can pre-process the data, and can work on unbalanced data easily. Data mining basically uses more direct approach and does meta-heuristics search on data.

The scope of data mining is subjected to automated prediction of trends and behaviors. Artificial neural networks, decision trees, genetic algorithms, nearest neighbor method and induction rules are some of the most widely used methods of data mining. This project makes use of Classification algorithm based on two rules induction, two decision tree algorithms of data mining and naive bayes algorithm (which is also a classification algorithm used for prediction). Data mining techniques have been under development for decades and are of huge use in research areas like statistics, artificial intelligence and machine learning [2].

This study proposes to foretell the student's academic failure using the algorithms of data mining techniques. The algorithms are applied on huge collection of dataset and the results are obtained, through which the failure can be predicted. This information is more useful for the teachers and principal of the organization, so that they can make proper arrangements and facilities to increase the capability of students and reduce/prevent the failure of students in academics years. These experiments can show almost expected results in context with economic, educational or sociological characteristics that may be helpful in foretelling low academic performance.

The paper is organized as follows: Section II presents Related Work. Section III, present Classification and finally section IV, summarizes the Main conclusions.

II. Related Work

Bresfelean worked on the data collected through the surveys from senior undergraduate students at the faculty of economics & business administration in Cluj-Napoca [5]. Decision tree algorithms in the WEKA tool, ID3 and J48 were applied which students are likely to continue their education with the postgraduate degree. The model was applied on two different specializations students' data and an accuracy of 88.68% and 71.74% was achieved with C4.5.

P. Cortez and A.Silva [6] worked on secondary students' data to predict their grade in contact education system. Past performance as well as socio-economic information was collected and the results were obtained using different classification techniques. It was found that the tree based algorithm outperformed the methods like Neural Networks and SVM.

V.P. Bresfelean, M. Bresfelean and N. Ghisoiu [7] found that students' success depends on students choice in continuing their education with post university studies or other specialization attribute, students admittance grade and the fulfillment of their prior expectation regarding their present specialization.

Baradwaj and Pal [8] obtained the university students data like attendance, class test, seminar and assignment marks from the students' previous database, to predict student mark he is likely to achieve.

In Canada there is a failure rate of more than 30% after the first two years in the Faculty of Engineering [9]. Different data mining techniques such as clustering and classification approaches e.g. K-means and hierarchical clustering, and K-nearest neighbour and naïve Bayes classifiers can be used to predict the failure rate of students.

For avoiding extra, unofficial seats during engineering admissions, New Zealand Government has adopted policy of penalizing the institute [10]. So as a result, New Zealand Universities adopted entry policy i.e. student's performance can be evaluated during final year of secondary school. Based on this, performance prediction of success of particular student in engineering can be done. Students who do not perform well can be restricted from admission of engineering. Mathematics with calculus, Mathematics with statistics, Physics etc. subject's marks were considered for prediction and data mining algorithms were applied on it.

S.Anupama Kumar and Vijayalakshmi M.N concluded that Decision rule and One R rule algorithms can be used to predict the result of the fifth semester of student in higher education based on the marks obtained by the students in the previous four semesters [11]. Rule based algorithm can provide efficiency in predicting the student's performance in higher education using the previous historical data.

H. Bydovska and L. Popelinsky used student academics related data and social behavioural data of students to predict the performance [12]. They carried out different experiments using data mining techniques such as support vector machine (SMO), OneR rule, Naïve bayes and decision tree. Support vector machine technique gave most accurate result.

III. CLASSIFICATION

This projects aims in predicting the student's academic failure using data mining technique. The method proposed in this paper for predicting the academic failure of students belongs to the process of Knowledge Discovery and Data Mining.

A. Methods of the Project:

There are four main methods of the project. They are as follows:

1. Data Collection
2. Data Management
3. Data Mining
4. Implementation

Data Collection is a process where information about the students is collected. This information is nothing but the data that will be useful in predicting the academic failure of students. The data about students is collected in three different categories;

A. First category is specific survey: where personal and family related information of the student is collected.
B. Second category is general survey: where previous education information of the students is obtained. The data is the information that is required by various higher and secondary education institutions while admitting the students in their institutions.
C. Third category is departmental survey: Where the academic related factors of students i.e. mark obtained by the students in different class at the end of semester.

And Finally, the output variable /attribute (result) to be predicted is the final student performance or academic status that has two possible values PASS (Students who has passed the course) and FAIL (Students who has to repeat the course). And lastly, all data is to be integrated into single dataset.

Data Management refers to preparing the data for applying data mining techniques. In data management, data cleaning, transformation of variables, data reduction and data partitioning is carried out. One of the most important techniques of data management is the selection of features (attributes). The attribute selection algorithm tries to select those features of students which have greater impact on their academic status. There are a wide range of attribute selection algorithms that can be grouped in different ways. From which I have selected few attribute of it and those are CfsSubsetEval, Filtered-AttributeEval, FilteredSubsetEval, etc. Because of these attribute selection algorithms, the best attributes out of huge number of attributes of students that affect the student's performance can be selected.

Data Mining consists of certain algorithms that help in predicting the student's failure using classification algorithms. For doing this task, the classification algorithms based on two rules of induction algorithms, and two decision tree algorithms is proposed to use. And Naive Bayes Algorithm is also used to resolve the problem of high dimensionality [16] provided by Microsoft SQL Server Analysis Services. This algorithm is basically used for predictive modeling which is based on Bayesian

Techniques. Finally, all the algorithms have been executed, evaluated and compared in order to determine which one obtains the best results.

Implementation is the last phase of the project where the results obtained from DM techniques are interpreted into a model. For implementation, I am going to make use of .Net Technology.

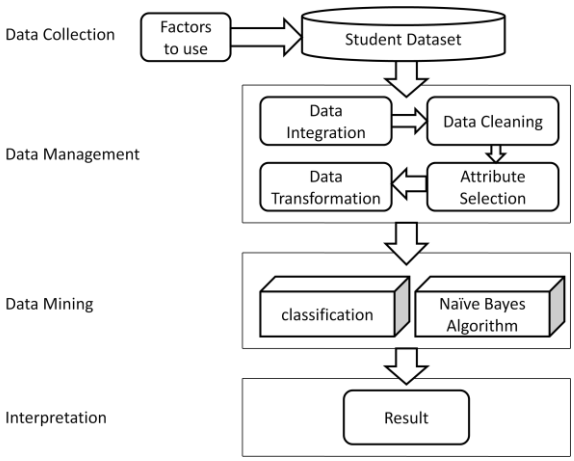

Figure 1: Methods used to predict student failure.

In my paper some of DM algorithms will be used to foretell the student's failure so that proper attention can be given to those students who may fail in future. This project will help the instructors as well as students to improve their performance by adapting certain changes in the standards of their teaching methodologies.

D. Modules of the project:

Educational data mining is basically used which focuses on development of methods to better understand students and the environment in which they learn. For this project, I am going to implement the spiral model of the software development models. Spiral model is a combination of prototyping model and waterfall model [15]. This model is basically used for large projects and the projects that require continuous up gradation. The spiral model consists of four phases named as planning phase, risk analysis phase, development and testing phase, and evaluation phase.

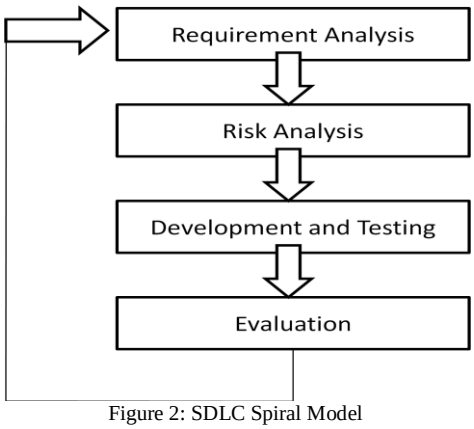

Figure 2: SDLC Spiral Model

One iteration (spiral) consists of these four activities and the output of this is a small prototype of the large software. This prototype is checked to see if it meets the required expectations and then all the four activities are again repeated for all the spirals until the entire project is built.

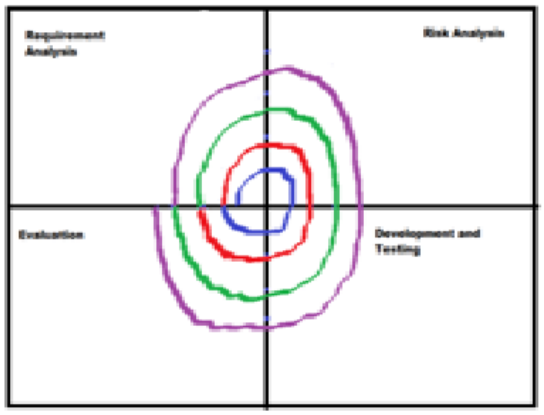

Figure 3: Spiral Model Design

In the (fig.3) each iteration is represented by different color. The first iteration (spiral) is shown by blue color which covers all the four phases of spiral model (Requirement Analysis, Risk Analysis, Development and Testing, Evaluation). Once the evaluation phase for the first iteration (spiral) is completed, the second iteration (spiral) is started which is represented by red color, here again from requirement analysis to evaluation phase and so on until the entire project is build. The advantage of using spiral model is that development of the project is fast, risk factor is evaluated, customer feedback is taken and changes are implemented faster and so on. The disadvantage is that it is not suitable for smaller projects; spiral may go infinitely.

E. Components of the Project:

The components of the project are mainly divided into two parts: Functional Components and Non- Functional Components.

Functional Components: are those components of the project whose actions/ results can be seen on screen. These are the entities whose actions can predict the failure of students in future. They are as follows:

1. Student
2. Teacher
3. Administrator
4. Prediction Tool

Student is the basic component of the project. The project mainly focuses on predicting the students' academic failure so that proper guidance can be provided to those students who may fail in future and help them from dropping out. Each student registers itself on the site, and can fill its information details. The information can be his/her personal, academic and department wise information. The students only have the authority to see their results and notices arranged for them by their teacher. Students can use them to identify their learning tasks, activities and resources to improve their learning.

Teacher has a very important role in this project. The teacher is the only person who has the right and authentication to access the prediction tool. The teacher can view the results calculated by the prediction tool and take appropriate decisions regarding that particular student. The teacher can view the details of all the students, manage the lecture, manage the practical batches of the students, add/update other skill-sets of students, short-list the students, arrange exam schedule for students, arrange notices regarding test, exams, results or any other departmental activity for the students and prepare a report of it. Teachers can use them to get more feedback, to identify the students at risk of failure and guide them to help them succeed, to identify most commonly made mistakes and to organize the contents of site in an efficient way.

Administrator can view the final class wise result of students and accordingly arranges the notices for the teachers. He can also use them to decide which course to offer. Basically here all the users have different rights and authentication to access the information.

Prediction Tool is a tool that calculates the number of students that may fail in future. The tool is basically based on data mining concept and consists of classification algorithm that calculates the failure of students. The classification algorithm is composed of two rules of induction algorithm, two decision tree algorithms and naive bayes Algorithm. The induction rule algorithms are NNge (it is a nearest neighbor approach) and OneR [1], which uses minimum-error attribute for class prediction and the decision tree algorithms are SimpleCart [3], which implements minimal cost-complexity feature and RandomTree [1], which considers K randomly chosen attributes at each node of the tree. Naive Bayes classification algorithm calculates the probability of every state of each input column, given each possible state of the predictable column [16]. The decision tree algorithms, induction rules and naive bayes algorithms can be easily implemented in the form of IF-THEN rules of object-oriented programming, which can be easily understood. These rules can show factors and relationships that influence student academic performance. It can be used to classify the frequent candidate sequences into different classes such as high performance students, low performance students. So implementing the if-then rules is mainly used for interpretation of the result. Students who are at the risk of failure can be provided with the guidance of the teacher. In this way, even a normal user who doesn't have any deep knowledge about data mining, for e.g. teacher and administrator can easily understand the results obtained using these algorithms.

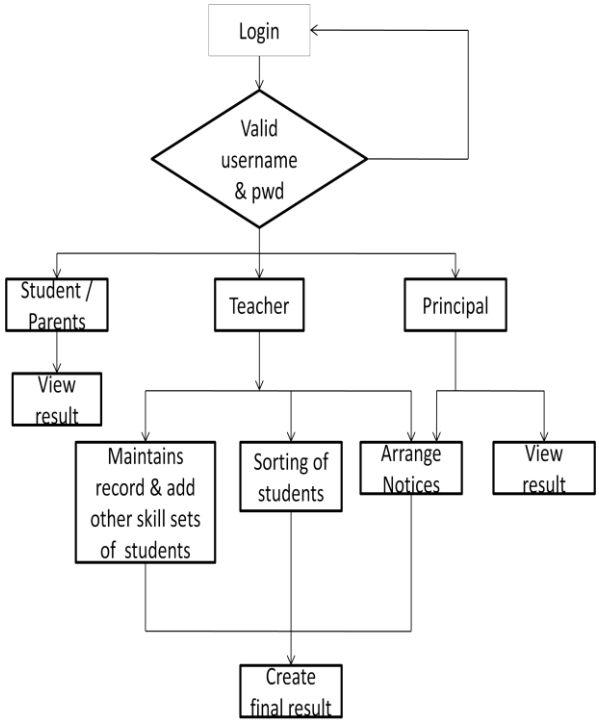

Figure 4: Flow of the System

The Fig. 4 shows the actual flow of the whole system. There are three types of users; student, teacher and principal (Administration). Each user has different user name, user id, user type and password.

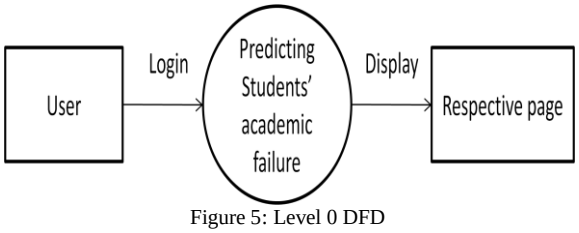

Figure 5: Level 0 DFD

A data flow diagram (DFD) is a graphical representation of the "flow" of data through an information system. DFDs can also be used for the visualization of data processing. Fig. 5 shows Level 0 DFD. It shows the entire system as a single process.

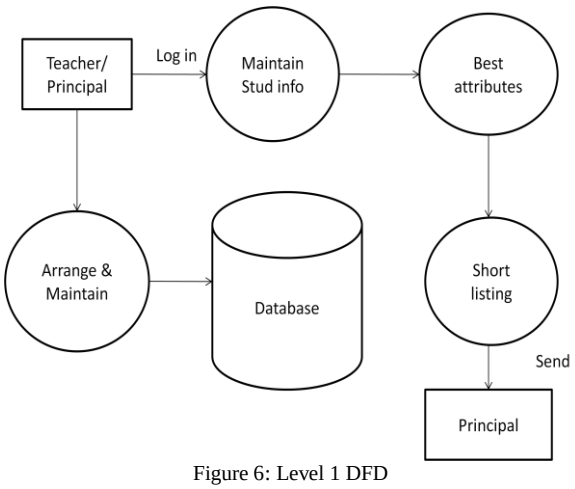

Figure 6: Level 1 DFD

This level 0 DFD is next "exploded" to produce a level 1 DFD that shows some of the details of the system being modeled.

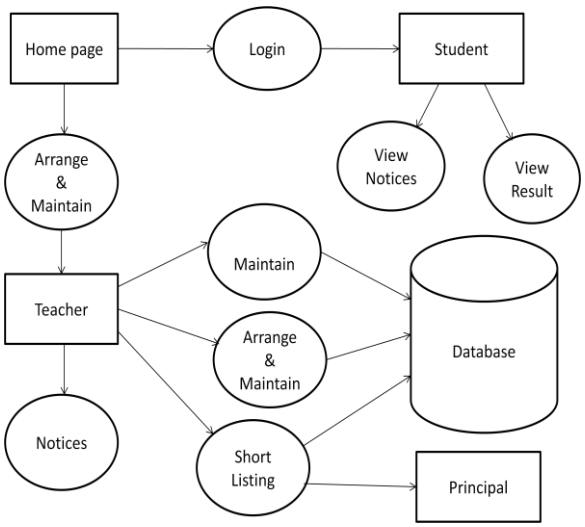

Figure 7: Level 2 DFD

Next, Fig. 7 shows Level 2 DFD, it shows how the system is divided into sub-systems (processes), each of which deals with one or more of the data flows to or from an external agent, and which provide all of the functionality of the system as a whole.

Non-Functional Components: are those components of the project that run in background and whose actions can't be seen on screen. The non-functional components support the functional components and together they produce the final result of student's failure report.

They are as follows:

1. Data Collection Techniques
2. Attribute Selection
3. Dataset Management

Data Collection Techniques basically deal with gathering student related information that will be useful in predicting their failure in future. The information is provided by the student itself. There are three categories in which the data is collected as shown previously. All the information is collected through the survey/ from college reports which include the personal and family related information of the students like number of members in a family, occupation of father and mother, living with

one's parents, their living location, whether suffering from critical illness, etc. Then students' information of about their Past education institutions that is required while admitting the students in the next school/ institutions are also used. For example, age, gender, previous school information, type of school, marks obtained in previous class, extra activity, sports etc as past performance of a student is indicative of his present/future performance in most of the cases. And lastly the final score obtained in the present education institution is also collected. Finally all this information is then stored in the dataset.

Attribute Selection basically deals with selecting the best attributes out of huge collection of data, based on which the results can be calculated. Practically, the information provided by each student is more than sufficient for the prediction. Instead of making use of each and every information, here we can select few best attributes out of the huge collection of data for prediction and precede the further process of prediction. This simplifies the complexity of the programmer and also the program. There will not be much difference in the results obtained. This step of attribute selection is only to ease the functionality.

Dataset Management deals with the management of data that is stored in the dataset. The information provided by the students may not be accurate or may not be precise. Also it is difficult to obtain the social data. For example, students are reluctant to reveal the information like parents' income and may provide incorrect data. Dataset Management involves Data Cleaning, Integration and Discretization, and Variable Transformation. It also involves data redundancy, spelling mistakes, invalid data, etc. For example, "N" is to be transformed into "N ". Also in case of student's age, it should be set in the dd/mm/yy format. Another case is that numerical values of the marks obtained by students in each class should be changed to categorical values [1]. For e.g. if marks obtained between 70-100% then "Topper" , if marks obtained between 55-69% then "Average" and if marks obtained between 40-54% then "Below Average" and so on. And lastly, all the cleaned data is to be integrated into a dataset. It will be enable to identify the students in advance who are likely to fail and allow the teacher to provide appropriate inputs.

IV. CONCLUSION

Prior work on predicting student's academic failure was based on Weka tool. All the algorithms required for obtaining results were just outsourced by the previous system. Also the existing system implement five rules of induction and five decision tree algorithms which increased the complexity and overhead of the system. In this paper, I implemented the algorithms in the system on my own. We did not outsource the algorithms from Weka tool. Also I implemented only two rules of induction, two decision tree algorithms and naive bayes algorithm which decreased the complexity and overhead of the system. The selection of the features attributes of the student can be done manually or automatically using algorithms. I made this project a real-time application which can be used in any educational organization for predicting the failure of students and reducing it.

References

1. Carlos Marquez-Vera, Cristobal Romero Morales, and Sebastian Ventura Soto, "Predicting school failure and dropout by using data mining techniques". IEEE Journal of Latin-American Learning Technologies, Vol. 8, No. 1, February 2013.
2. C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Syst. Appl., vol. 33, no. 1, pp. 135-146, 2007.
3. L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. New York , USA: Chapman & Hall, 1984.
4. C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans.Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 6, pp. 601-618, Nov. 2010.
5. V.P. Bresfelean, "Analysis and Predictions on Students' Behaviour using Decision trees in WEKA Environment", Proceedings of the ITI 2007 29th Int Conf. on Information Technology Interfaces, June 25-28, 2007.
6. P.Cortez and A.Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeria (Eds.), pp.5-12.
7. P. Bresfelean , M. Bresfelean, N. Ghisoiu, " Determining Students' Academic Failure Profile Founded on Data Mining Methods", Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces, June 23-26 2008.
8. B.K. Baradwaj, S. Pal, "Mining Educational Data to Analyze Students' Performance", (IJACSA) International Journal of Advanced Computer Science and Application, Vol. 2, No. 6, 2011.

9. Kin Fun Li, D. Rusk, F. Song,—Predicting Student Academic Performance‖, Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, IEEE conference publication, pp 27-33, July 2013.
10. Dale A. Carnegie, Craig Watterson, Peter Andreae, Will N. Browne —Prediction of Success in Engineering Study‖, Global Engineering Education Conference (EDUCON), IEEE, pp 1-9,Apr 2012.
11. S.Anupama Kumar, Vijayalakshmi M.N., —Mining of Student Academic Evaluation Records in Higher Education‖, International Conference on Recent Advances in Computing and Software Systems (RACSS), IEEE conference publication, pp 67–70, Apr,2012.
12. H. Bydovska, L. Popelinsky, -Predicting student performance in higher education‖, 24th International Workshop on Database and Expert Systems Applications (DEXA), IEEE conference publication, pp 141-145 Aug 2013.
13. http://www.thearling.com/text/dmwhite/dmwhite.htm
14. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
15. http://www.softwaretestinghelp.com/spiral-model-what-is-sdlc-spiral-model/
16. https://msdn.microsoft.com/en-us/library/ms174806.aspx

AUTHOR(S) PROFILE

Lumbini P. Khobragade, received the B.E. degree in Information Technology from P.E.S. College of Engineering, Dr. BAMU University, in 2011. She has worked as a Lecturer, in Department of Computer Science, of P.E.S College of Engineering, Dr. BAMU University. She is currently pursuing the M.E. degree in Department of Computer Science from Deogiri institute of Engineering & Management Studies, Dr. BAMU University, Aurangabad. Her research interest includes Data Mining and Software Testing.

Pravin Mahadik, Assistant Professor at Deogiri institute of Engineering & Management Studies, Dr. BAMU University, Aurangabad. He has received the B.E. degree and M Tech. degree in Computer Science from JNEC College of Engineering, Aurangabad and Dr. Babasaheb Ambedkar Technological University, Raigad in 2010 and 2013, respectively. His area of research includes Network Programming, Java, and Software Engineering. He has published two papers in Siber Times Internal Journals.

