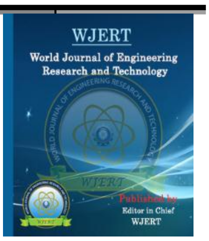
World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 5.924



IMPLEMENTATION AND ANALYSIS OF MACHINE LEARNING APPROACHES AND TECHNIQUES FOR STUDENT DROPOUT PREDICTION

[1]*Manpreet Singh and [2]Er. Harjeet Singh

[1]M.Tech Scholar, St. Soldier Inst. of Engg. & Technology, Jalandhar.
[2]Assistant Professor, Computer Science & Engg. St. Soldier Inst. of Engg. & Technology, Jalandhar.

Article Received on 25/08/2019 **Article Revised on 15/09/2019** **Article Accepted on 05/10/2019**

***Corresponding Author**
Manpreet Singh
M.Tech Scholar, St. Soldier Inst. of Engg. & Technology, Jalandhar.

ABSTRACT

The aversion of understudies dropping out is viewed as very significant in numerous instructive organizations. In this paper we portray the aftereffects of an instructive information examination contextual analysis concentrated on discovery of dropout of Systems Engineering (SE) college understudies following 6 years of enlistment in a Colombian college. Unique information is expanded and improved utilizing a component building process. Our test results demonstratedthat straightforward calculations accomplish dependable levels of exactness to recognize indicators of dropout. Data mining techniques outcomes were thought about in request to propose the best alternative. Additionally. Also, we present a few discoveries identified with information quality to improve the understudies information gathering process.

KEYWORDS: Machine Learning, Decision Tress, Predicting Dropout Student.

I. INTRODUCTION

Student dropout in tertiary education is a salient problem for asian countries as it is associated with social distress, losses in efficiency for public resources and increasing financial costs for private education. The problem has a broadly accepted theoretical background proposed by researchers, which models the process of student attrition as a sociopsychological interplay between the characteristics of the student and the experience she faces in University.

Researchers have reported thoroughly on improved retention rates when improving students' experiences.

Educational Data Mining (EDM) researchers have been supporting institutional interventions through students' dropout prediction. Identifying students prone to abandon and the factors that are more relevant to the prediction models, specialized professionals can generate action plans to retain them.

This study incorporates student data from different factors with no credits and evaluates their effect on the performance of dropout prediction using decision trees. Education Data Mining (EDM) is the research area which focuses on the use of Data Mining techniques in educational contexts.

2. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Some machine learning methods

Machine learning algorithms are often categorized as supervised or unsupervised

- Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
- In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system

doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

- Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiringunlabeled data generally doesn't require additional resources.
- Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

Machine learning enables analysis of massive quantities of data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with AI and cognitive technologies can make it even more effective in processing large volumes of information.

Data mining is used on an existing dataset (like a data warehouse) to find patterns. Machine learning, on the other hand, is trained on a 'training' data set, which teaches the computer how to make sense of data, and then to make predictions about new data sets.

3. Proposed methodology and algorithm

The traditional methodology for predicting student dropout uses all the data available at the end of the course together with classical and well-known classification algorithms. Next, we propose both a new methodology and a specific algorithm that attempts to detect students' dropout as early as possible.

Method The method proposed in this paper for predicting the academic failure of students belongs to the process of Knowledge Discovery and Data Mining.

The main stages of the method are:

– Data gathering. This stage consists in gathering all available information on students. To do this, the set of factors that can affect the students' performance must be identified and collected from the different sources of data available. Finally, all the information should be integrated into a dataset.

– Pre-processing. At this stage the dataset is prepared to apply the data mining techniques. To do this, traditional preprocessing methods such as data cleaning, transformation of variables, and data partitioning have to be applied. Other techniques such as the selection of attributes and the re-balancing of data have also been applied in order to solve the problems of high dimensionality and imbalanced data that are typically presented in these datasets.

– Data mining. At this stage, DM algorithms are applied to predict student failure like a classification problem. To do this, a new programming classification algorithm based on genetic programming is proposed and compared with other classical classification algorithms based on classification rules and decision trees. In addition, a cost sensitive classification approach is also used in order to solve the imbalanced data problem.

– Interpretation. In this last stage, the obtained models are analyzed to detect student failure. To achieve this, the factors that appear (in the rules and decision trees) and how they are related are considered and interpreted.

4. Algorithms

4.1 PART: PART is a separate-and-conquer rule learner. The algorithm producing sets of rules called „decision lists" which are planned set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the class of the first matching rule. PART builds a partial C4.5 decision tree. in each iteration and makes the "best" leaf into a rule each iteration and makes the "best" leaf into a rule. each iteration and makes the "best" leaf into a rule.

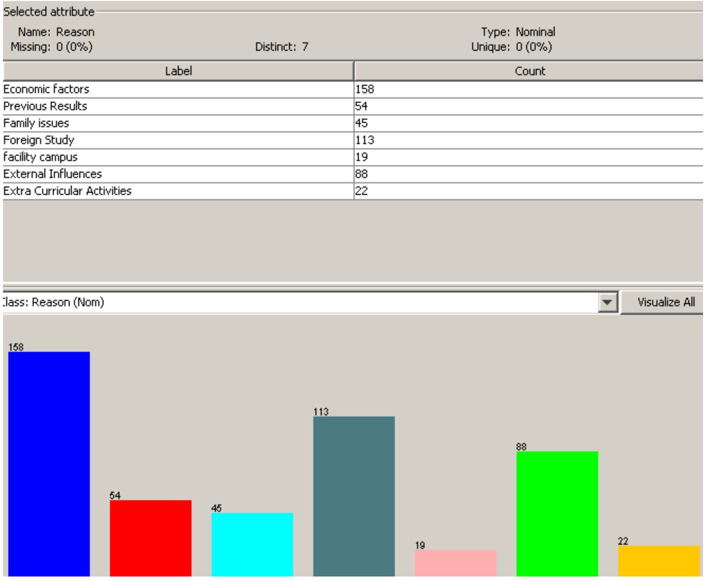
4.2 JRIP: JRip (RIPPER) is one of the basic and most popular algorithms. Classes are examined in growing size and an initial set of rules for the class is generated using incremental reduced error JRip (RIPPER) proceeds by treating all the examples of a particular decision in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

4.3 ECA: A ECA is a structure that shows the conditional dependencies between domain variables and may also be used to illustrate graphically the probabilistic underlying relationships among domain variables. A ECA consists of a directed acyclic graph and probability tables. The nodes of the net-ork represent the domain variables and an arc between two nodes indicates the existence of a underlying relationship or dependency among these two nodes.

5. Dataset: The data set used in this study comes from 500 students enrolled in the Engineering Program at a private college. The data is organized in single table, and include the following features:

Student name: Describe the name of the student
ID: Student ID/ Enrollment Number
Fathers Name: Describe the students father name
Branch: Name of the Course
Reason of Drop Out: describe the reason for drop out

6. RESULTS



Selected attribute

Name: Reason
Missing: 0 (0%)
Distinct: 7
Type: Nominal
Unique: 0 (0%)

Label	Count
Economic factors	158
Previous Results	54
Family issues	45
Foreign Study	113
facility campus	19
External Influences	88
Extra Curricular Activities	22

Fig. 1: Shows the graph of total dropout students.

Table 1: Shows the analysis between CCI, ICI & ER of PART, JRIP & ECA.

	PART	JRIP	ECA
CCI	32.66	32.06	36.47
Of ICI	67.33	67.93	63.52
ER	97.50	99.38	79.90

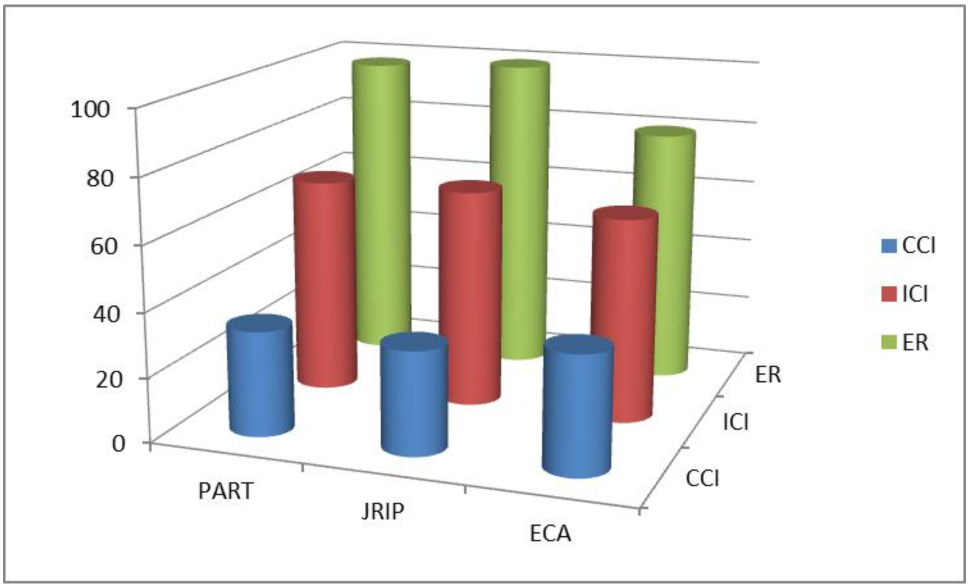


Fig. 2: Shows the analysis between CCI, ICI & ER of PART, JRIP & ECA.

7. CONCLUSIONS

As we have seen, predicting dropout students at engineering college can be a difficult task not only because it is a multifactor problem (in which there are a lot of personal, family, social, and economic factors that can be influential). To resolve these problems, we have shown the use of different DM algorithms and approaches for predicting student dropout. We have carried out several experiments using real data from engineering college. We have applied different classification approaches for predicting the academic status or final student performance at the end of the course. We have proposed a genetic programming model to obtain accurate and comprehensible classification rules. Furthermore, we have shown that some approaches such as selecting the best attributes, cost-sensitive classification, and data balancing can also be very useful for improving accuracy. Finally, as the next step in our research, we aim to carry out more experiments using more data and also from different educational levels (primary, secondary, and higher) to test whether the same performance results are obtained with different DM approaches (feature selection, data balancing, and cost-sensitive classification) and our ECA algorithm.

8. REFERENCES

1. M. N. Quadri and N. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques," G. J. of Computer Science and Technology, 2010; 10: 1-4.

2. M. A. Mac Iver and D. J. Mac Iver, "Beyond the indicators: An integrated school-level approach to dropout prevention," The G. W. U. C. for Equity and Excellence in Education, Arlington, VA., 2009.
3. C. Marquez-Vera, C. Romero and S. Ventura, "Predicting School Failure using Data Mining," in Proceedings of the 4th international conference on educational data mining, 2011.
4. C. K. Fox et. al., "Physical Activity and Sports Team Participation: Associations With Academic Outcomes in Middle School and High School Students," J. of School Health, 2010; 80: 31-37.
5. I. Esteban-Cornejo, C. M. Tejero-Gonzalez, J. F. Sallis and O. L. Veiga, "Physical activity and cognition in adolescents: A systematic review," J. of S. and M. in Sport, 2014; 18(5): 534-539.