International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 01 | Jan 2019 www.irjet.net p-ISSN: 2395-0072

Association Rule Mining to Identify the Student Dropout in MOOCs

BOBBY K SIMON[1], ANJANA P NAIR[2]

[1]BOBBY K SIMON, M.Tech Computer Science & Engineering. Sree Buddha College of Engineering, Ayathil, Elavumthitta Pathanamthitta , Kerala, India.
[2]Ms. ANJANA P NAIR, Assistant Professor Computer Science & Engineering. Sree Buddha College of Engineering, Ayathil, Elavumthitta, Pathanamthitta, Kerala, India

---

Abstract - *Data mining are quite popular for especially on mining system (MOOCs, FUZZY, APRIORI, and so forth.), identification &mitigation procedures are only functioning after the dropout was initiated Prevention, however, attempts of student dropout can be monitored before it is executed. This survey gives us knowledge about how students are been analyzed with this strategies can be acknowledged either by the aggregate data's of student system. It also shows the method for minimizing the student dropout and how to reduce the count of number of students been dropped. Our survey gives an answer for no student collaboration, with every student utilizing just internal knowledge picked up by online information. This also shows the benefits of the different techniques dropout models for better understanding of the online courses and its prevention. We recognize their respective motivations and distinguish their advantages and drawbacks in this paper.*

***Key Words*: MOOCs, FUZZY, APRIORI, ASSOCIATION RULE MINING, EDM**

1. INTRODUCTION

Data mining is the computational procedure of finding designs in vast informational indexes including strategies at the convergence of computerized reasoning machine learning, bits of data, and database structures. It is not a disciplinary sublevel of programming arranging. The honest to goodness data mining undertaking is simply the customized or loader examination of tremendous measures of data to isolate effectively dark, intriguing examples, for example, gatherings of information records, surprising records, and conditions (affiliation administer mining, consecutive example mining). This as a rule includes utilizing database methods, for example, spatial lists. These examples would then be able to be viewed as a sort of rundown of the information, and might be utilized as a part of further examination. For instance, the information mining step may recognize numerous gatherings in the information, which would then be able to be utilized to acquire more exact expectation results by a choice emotionally supportive network.

Rule mining is a strategy which is intended to discover visit designs, connections, affiliations, or causal structures from informational indexes found in different sorts of databases, for example, social databases, value-based databases, and different types of information vaults. Given an arrangement of exchanges, affiliation run mining expects to discover the principles which empower us to foresee the event of a particular thing in light of the events of alternate things in the exchange. Affiliation lead mining is the information mining procedure of finding the standards that may oversee affiliations and causal questions between sets of items. The fundamental utilizations of association rule mining:

1.1 Illustrative Applications

These MOOCs [1], one of the most recent improvements in the advancement of web based learning have seen an expansion in its fame in the course of the most recent couple of years. It is by and large generally acknowledged because of its usability, open access and ease. The vision is of boundless online courses, accessible to for all intents and purposes anybody with an Internet association that would drastically reshape the standard classroom while additionally changing the existence ways of understudies in creating nations, at practically zero cost [19]. The different MOOCs stages at present being used are edX, Udacity, Future Learn offered by various colleges like MIT, Stanford, Harvard and so on.

An ever increasing number of colleges are putting forth courses through MOOCs. Feelings are differed with regards to the viability of MOOCs. Likewise, the wearing down rates are a reason for concern [20]. A gander at the quantity of confirmed and uncertified understudies in any course gives us disturbing signs. The check of uncertified understudies is pretty much equivalent to the aggregate number of understudies enlisted. This paper centres around the initial move towards this - how to recognize understudies who are in danger of not getting affirmed in a course.

1.2. Formal Problem Statement

The word 'at-risk' generally implies presented to some peril or damage. In this specific circumstance, the term alludes to understudies who may drop out at any phase of the course or the individuals who does not meet the essential passing criteria. Additionally, in danger infers a probability which can be corrected with help which is the thing that this exploration work recognizes.

1.3. Objective

Past the hard errand on distinguishing the understudies who can have conceivable danger of dropping out, a similar dropout additionally conveys an enormous harm to current money related and social assets. In this manner, the general public likewise loses when they are ineffectively overseen, once the understudy fills the opportunity however he surrenders the course before the end. Online instruction frequently manages the issue identified with the high understudies' dropout rate amid a course in numerous territories.

There is immense measure of recorded information about understudies in online courses. Thus, a pertinent issue on this setting is to look at those information, going for finding powerful instruments to comprehend understudy profiles, distinguishing those understudies with attributes to drop out at beginning time in the course. In this paper, we address this issue by proposing prescient models to give instructive chiefs the obligation of distinguishing understudies who are in the dropout bound. This prescient model took in thought scholarly components related with their execution at the underlying orders of the course therefore we go for affiliation lead digging for finding the help, certainty, lift can be ascertained and for additionally picking up exactness in understudies drop out we utilize Fuzzy calculation for limiting the understudies from dropout list.

1.4. Overview of the Project

Online instruction regularly manages the issue identified with the high understudies' dropout rate amid a course in numerous regions. There is a colossal proportion of valid data about understudies in online courses. Henceforth, an applicable issue in this setting is to analyze those information, going for finding viable instruments to comprehend understudy profiles, distinguishing those understudies with ascribes to drop out toward the starting time in the course. An instrument to help the pre-dealing with arrange was used as a piece of a demand to design data for use of Data Mining counts.

In this paper, the issue by proposing prescient models to give instructive chiefs or the staff individuals with the obligation to distinguish understudies who are in the dropout bound. For computing this affiliation control mining strategies to figure the help, certainty, lift. What's more, fluffy calculation is utilized for a productive outcome from the dataset of understudies from the specific organization, which were utilized amid the assessment, keeping in mind the end goal to locate the model with the most astounding precision in foreseeing the profile of dropouts understudies. This makes simpler the file age in the good organization with the information digging for online courses. Thusly, for what was uncovered above, it legitimizes the requiring of a speculation to create proficient expectation techniques, appraisal and line up of the understudies with dropout hazard, permitting a future booking and appropriation of proactive measures pointing the abatement of the expressed condition or to diminish the quantity of understudies in chance level.

2. RELATED WORK

In this segment [2], audit of some past takes a shot at amass irregularity MOOCs. This study is an endeavor to give an organized and expansive outline of broad research on MOOCs strategies traversing various research zones and application spaces. The greater part of the current overviews on MOOCs either centre around a specific application space or on a solitary research zone. There are various related works for online courses [18], with numerous classifications and talk about strategies under every class. This study expands upon this work by altogether extending the talk in a few bearings.

These include two more classifications of MOOCs systems [4], data theoretic and ghastly strategies, to every one of the classifications talked about on this overview. For every one of the classes, it examines the strategies as well as distinguishes novel suppositions with respect to the idea of understudies made by the systems in that classification. These presumptions are basic for deciding when the systems in that classification would have the capacity to recognize understudies who are in danger, and when it would come up short. In outline [5], a couple of examinations investigating the utilization of Online courses for MOOCs frameworks to anticipate and recognize understudies who are in threat dropout. In any case, those works share similitudes:

(i) Identify and compare algorithm performance in order to find the most relevant EDM techniques to solve the problem. (ii) Identify the relevant attributes associated with the problem. Some works use past time-invariant student records (demographic and pre-university student data).

In this study, commitment to those introduced in this segment, gives us a different criticism with various frameworks, assembling a bigger number of qualities, factors and time-invariant. Other than being worried about the ID and correlation of calculations, distinguish the properties of incredible significance and take care of the issue to foresee in more precedence the prone to dropout understudies [6]. While a portion of the current overviews specify the diverse uses of MOOCs, it gives a definite dialog of the application spaces where forecast methods have been utilized. For every space it talks about the idea of forecast for MOOCs, the distinctive parts of the understudy issue, and the difficulties looked by the online courses and its strategies. It additionally gives a rundown of systems that have been connected in every application area [7].

3. SYSTEM ANALYSIS

System analysis is a critical thinking strategy which disintegrates the framework into its pieces for the concentrate how well this part will function and achieve their motivation. It alludes to an efficient, organized process for recognizing and tackling issues. The framework investigation process lifecycle system comprises of four stages. They are:

i. Study phase
ii. Design phase
iii. Development phase
iv. Implementation phase

3.1 Existing System

Beyond the hard task undertaking of recognizing the understudies who can have a conceivable danger of dropping out, a similar dropout additionally conveys an enormous harm to current money related and social assets. Accordingly, the general public likewise loses when they are ineffectively overseen, once the understudy fills the opening however he surrenders the course before the end.

Online training frequently manages the issue identified with the high understudies' dropout rate amid a course in numerous regions. There is an immense measure of chronicled information about understudies in online courses. Consequently, an applicable issue in this setting is to inspect that information, going for finding compelling components to comprehend understudy profiles, distinguishing those understudies with attributes to drop out at a beginning period in the course.

Disadvantages

i. Nowadays many students face this dropout during their academics.

Dropout students are also not able to analysis, where their lack in their academics or which subject.

3.2. Proposed System

The word 'at-risk' as a rule implies presented to some peril or mischief. In this unique situation, the term alludes to understudies who may drop out at any phase of the course or the individuals who don't meet the essential passing criteria. Likewise, in danger suggests a plausibility which can be redressed with help which is the thing that this exploration work distinguishes yield.

This paper mainly deals with, “How to enhance student execution in MOOCs?" and furthermore "how to distinguish understudies who are in danger of not getting affirmed in a course?". For this some datasets of few foundations alongside the understudy detail and their scholarly courses. Pre-handling is done on the information gathered in light of the properties chose for distinguishing a student as in danger. An expectation show utilizing affiliation rules is defined from pre-handled information. Likewise, the same pre-handled information is standardized and afterward affiliation rules are connected to shape a forecast demonstrate.

The two yields are thought about. Note that these means are finished with a base number of characteristics. Also, upon the examination, the chose characteristics are re-characterized or included till the two outcomes focalize. It can be seen as an iterative procedure of detailing the model, correlation, and redefinition. Later again this dataset is utilized as a part of the Fuzzy calculation for increasing better productivity in our outcome examination. Since the fluffy framework yield is an agreement of the majority of the data sources and the majority of the tenets, fluffy rationale frameworks can be very much carried on when input esteems are not accessible or are not dependable. Weightings can be alternatively added to each run in the administer base and weightings can be utilized to control how much a manage influences the yield esteems. These manage weightings can be founded on the need, dependability or consistency of each run the show. These govern weightings might be static or can be changed progressively, even in view of the yield from different guidelines.

Advantages

i. Can identify students at risk level or border.
ii. Proposed method also help faculty or staff member to know which student need to be focused for the improving.
iii. User or students will get awareness about their academic level and if they are at risk. Then awareness also will be given to them along with the suggestion to improve them.
iv. With this help uncertified students level will get decreased.

4. SYSTEM ARCHITECTURE

i. Input can be taken from define attributes
ii. Set the data set to Input dataset- student details of a college including their courses.
iii. Then, the output to two methods
    a) Prediction model to get the output
    b) Another one is normalization and then prediction to get the output.
iv. Next we can determine the attributes such as;
    a) Dataset.
    b) Attribute list.
    c) Selected attributes.

v. Then it generates the student data according to their selected attributes.
vi. On this pre-processed data topic processing is done which involves support and confidence calculation.
vii. Mining rule
viii. To shortlist the student data
ix. According to their percentage analysis we can define those students who are in border.

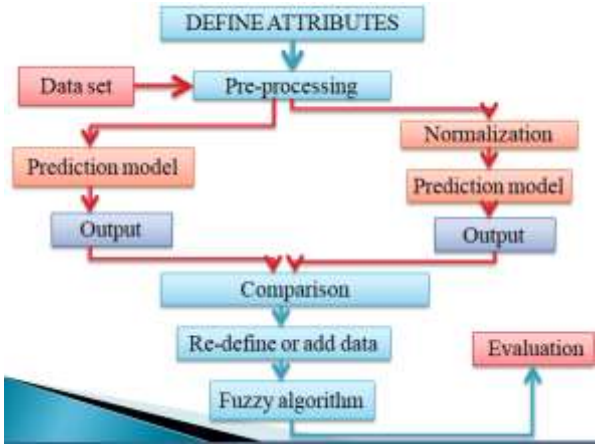


Fig -1: Proposed System Architecture

4.1 Algorithm

4.1.1. Association Rule Mining

Association rule learning is a manage based machine learning strategy for finding fascinating relations between factors in substantial databases. It is proposed to recognize solid standards found in databases utilizing a few measures of intriguing quality. Data mining is the route toward finding associations or cases among numerous fields in tremendous social databases [12]. Different information mining advances are available like affiliation guidelines, grouping, and characterization investigation and so on. Affiliation Analysis [13] helps in finding concealed examples from extensive accessible informational indexes. It expects to find intriguing relationships and incessant examples among information of intrigue [14].

The well-known Market Basket Analysis [15] is done to see which all blends of things are obtained together, for the most part done by retailers to comprehend buy conduct of clients. Such an investigation is more valuable to comprehend designs that are less seen generally. It works by first producing successive item sets in view of the help edge and after that creating rules from these incessant item sets in view of the certainty limit. In the present work, affiliation examination is done to recognize those students who are in danger. The example for portraying an understudy as ensured or not is found as far as different characteristics. On the off chance that the student is confirmed; else, he is in danger.

There are different affiliation investigation calculations like Eclat calculation, Apriori calculation, Relim calculation and the FP development calculation [16], [17], [18], [19]. Among these, the Apriori calculation lessens the computational many-sided quality of successive item set age. The rule expresses that if an item set is a visit then the majority of its subsets will likewise be visited [19]. An affiliation govern is a suggestion articulation of the form X→Y; where X∩Y=∅. The three factors that describe affiliation lead our support, confidence and lift.

Remembering the true objective to pick charming principles from the game plan of each possible control, restrictions on various proportions of significance and intrigue are utilized. The best-known imperatives are least limits on support and confidence

i. **Support**

$$P(X \cap y) = (X \cap y)/N$$

Support means that how much of the time the item set shows up in the dataset. Support decides how frequently a material to a given informational collection. It is the proportion of various exchanges in which the things X and Y have jumped out at a sum of N transactions [19]. A low help demonstrates that manage happened by possibility.

ii. **Confidence**

$$P(Y|X) = P(X \cap Y)/P(X)$$

Confidence means that how regularly the administer has been observed to be valid. Confidence decides how every now and again things in Y show up in exchanges that contain X. It is the proportion of various exchanges in which things X and Y have happened together to the quantity of exchanges where X has happened [16]. It means that how regularly the control has been valid.

iii. **Lift**

$$P(X \cap Y)/P(X) * P(Y)$$

Lift is a measure to foresee the activity of affiliation standards to enhance the reaction of the run the show. On the off chance that the esteem is certain; it speaks to a positive connection between the things X and Y [16]. It gives a measure of the execution of the model at accurately foreseeing the cases. The lift of a lead is characterized as: The two fundamental systems connected in this affiliation lead mining is to discover visit thing sets which discover all item sets that fulfill the base help limit and afterward locate the high certainty rules from the incessant item sets.

4.1.2 Fuzzy Algorithm

Data mining utilizes different systems and hypotheses from an extensive variety of regions for the learning extraction

from expansive volumes of information. In any case, vulnerability is a far-reaching marvel in information mining issues. As needs be, fuzzy rationale is connected to adapt to the vulnerability in reality. The fuzzy technique for thinking is a kind of many-respected support in which reality estimations of segments might be any authentic number somewhere in the extent of 0 and 1. It is used to manage the possibility of midway truth, where reality regard may stretch out between absolutely obvious and absolutely false. The three procedure steps incorporated into the fuzzy calculation are as per the following:

i. Fuzzy if all input values into fuzzy membership functions.
ii. Execute all applicable rules in the rule base to compute the fuzzy output functions.
iii. De-fuzzy if the fuzzy output functions to get "crisp" output values

Each value will have an incline where the esteem is expanding, a principle where the esteem is equivalent to 1 (which can have a length of 0 or more noteworthy) and a slant where the esteem is diminishing. They can likewise be characterized utilizing a sign id capacity. One regular case is the standard calculated capacity characterized as:

$$S(x) = \frac{1}{1 + e^{-x}}$$

This has the following symmetry property,

$$S(x) + S(x) = 1$$

From this it follows that

$$(S(x) + S(x)).(S(y) + s(-y)).(S(z) + s(-z)) = 1$$

With this, it will assist us with defining us know the precision of the outcome accommodated every understudy and the hazard level execution will be examined by that specific foundation.

5. RESULT AND ANALYSIS

In this part of the experiment fig 2, the student details of a college were used as the dataset. The experimental results were measured in terms of accuracy, and with chances of their prediction, our project was analyzed and a better form of student feedback could pick up from that. The project was divided into three phases as admin, user and data analysis. On the field of admin we add the course and the subject with respect to their academic details. Next is the User section, Users is referred as students of a particular institute. Students can register respective with their email id. Video can be uploaded for their online course. They can view their performance. Result of each students. Awareness will be given to those Students who are at in risk. On this each user's have to register their details and wait for the approval and apart from this in this online they can also take the count of student's uploaded video and audio. And next is data analysis, at first we inserted some details of the student and with their respective subjects and the verified subject codes. Input dataset- student details of a college including their courses. Next the attributes can be determined as;

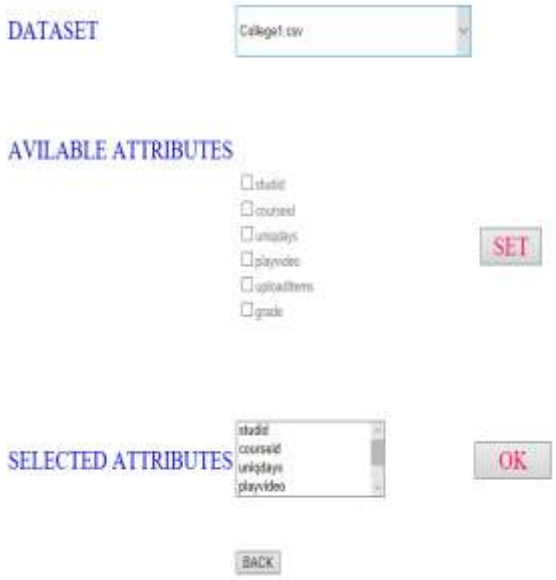


Fig -2: Student dataset and its attributes

Then it generates the student data according to their selected attributes.

5.1 Association Mining

In this some specified college with their subjects of some students. And enter the values for minimum support, which

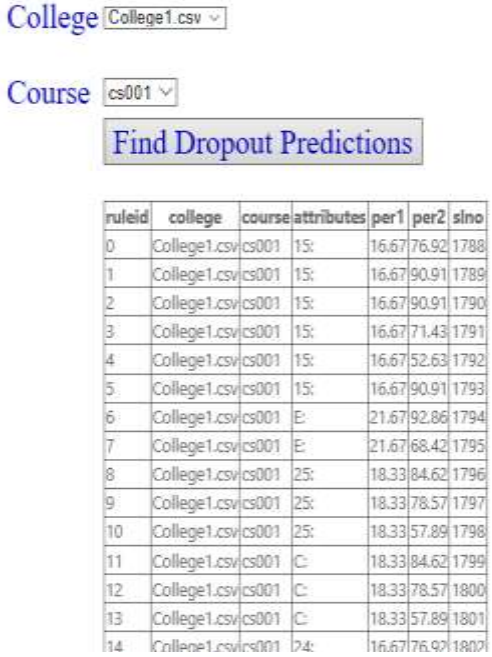
studid	courseid	uniqdays	playvideo	uploaditems	grade
1	cs001	15	DQ	14	C
2	cs002	25	DR	15	D
3	cs003	24	DS	12	B
4	cs004	23	DT	17	A
5	cs005	14	DU	18	O
6	cs001	21	DV	15	E
7	cs002	18	DW	19	C
8	cs003	24	DX	10	D
9	cs004	29	DY	14	C
10	cs005	21	DZ	12	B
11	cs001	14	EA	13	A

Fig -3: Generate the Student dataset and its functions

is Support- individual value. And then enter the values for minimum confidence which is Confidence- two strings. After this step generate item-set fig 3 that is to calculate the support and confidence using this association rule mining. By the help of Mining rule the student data can be shortlisted.

5.2 Fuzzy Dropout Predictions

This algorithm is used to predict students who are at risk level. Which is more easy to identify or to sort-out the students easily. Here, two different prediction ways that can be sorted. At first is According to their percentage analysis we can define those students who are in border. And second way is fuzzy algorithm.



ruleid	college	course	attributes	per1	per2	slno
0	College1.csv	cs001	15:	16.67	76.92	1788
1	College1.csv	cs001	15:	16.67	90.91	1789
2	College1.csv	cs001	15:	16.67	90.91	1790
3	College1.csv	cs001	15:	16.67	71.43	1791
4	College1.csv	cs001	15:	16.67	52.63	1792
5	College1.csv	cs001	15:	16.67	90.91	1793
6	College1.csv	cs001	E:	21.67	92.86	1794
7	College1.csv	cs001	E:	21.67	68.42	1795
8	College1.csv	cs001	25:	18.33	84.62	1796
9	College1.csv	cs001	25:	18.33	78.57	1797
10	College1.csv	cs001	25:	18.33	57.89	1798
11	College1.csv	cs001	C:	18.33	84.62	1799
12	College1.csv	cs001	C:	18.33	78.57	1800
13	College1.csv	cs001	C:	18.33	57.89	1801
14	College1.csv	cs001	24:	16.67	76.92	1802

Fig -4: Dropout prediction using fuzzy algorithm

Fuzzy based dropout prediction was done with the calculation of lift, confidence and support. Fig 4 Thus, the Membership value was calculated and with that we gained an accuracy value which was plotted in a graph. Comparison and Evaluation of Fuzzy algorithm v/s Association Rule algorithm to know which algorithm was able to give a better dropout of student's prediction.

6. CONCLUSION

This project was to identify why the students dropout was high at each level during their courses and also to analysis how to reduce this drawback. A new way to Presents a fuzzy algorithm for reducing dropout with association rule mining. With this it is possible to calculate and analysis the Ability of students inter-knowledge and were able to let them know where they were lacking, that is feedback to student about their performance, thus this method was done as prevention. Shows how they fail to learn. This program has expanded the quantity of understudies has passed the course. What's more, along these lines of the technique can be utilized in online courses as well as for the normal regular courses.

A fuzzy model has been implemented with success, and was able to perform named entity extraction and student identification on documents of a dropout nature. This work has achieved promising results, and, in conclusion, is predicted to open a new path for future research related to information extraction in the student domain proceeding for management growth and for educational growth.

REFERENCES

[1] Srilekshmi M, Sindhumol S, Shiffon Chatterjee, Kamal Bijlani, "Learning Analytics to Identify Students at-risk in MOOCs" , 2016 IEEE 8th International Conference on Technology for Education.

[2] Kannan Govindarajan, Vivekanandan Suresh Kumar, David Boulanger and Kinshuk, "Learning analytics solution for reducing learner's course failure rate", IEEE Seventh International Conference on Technology for Education (T4E), Warangal, pp. 83-90,2015

[3] Balakrishnan, Girish, and Derrick Coetzee. "Predicting student retention in massive open online courses using hidden markov models." Electrical Engineering and Computer Sciences University of California at Berkeley (2013).

[4] Halawa, Sherif, Daniel Greene, and John Mitchell. "Dropout prediction in MOOCs using learner activity features." Experiences and best practices in and around MOOCs 7 (2014).

[5] L.Athira, Aswini Kumar and Kamal Bijlani, Discovering learning models in MOOCs using empirical data, Springer India 2015,N.R.Shetty et al. (eds.), Emerging Research in Computing, Information, Communication and Applications, 2015.

[6] Matthew D. Pistilli and Kimberly E. Arnold, Course signals at Purdue: using learning analytics to increase student success, www.interscience.wiley.com, 2010.

[7] StanfordNews,http://news.stanford.edu/2015/10/15/ moocs-no-panacea- 101515/

[8] Massive Open Online Courses, https://www.massiveopenonlinecourses.com.

[9] Harvard Dataverse Network, http: //thedata:harvard:edu/dvn/.

[10] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Vol. 1. Boston: Pearson Addison Wesley, 2006.

[11] Agrawal, Rakesh, Tomasz Imieli_ski, and Arun Swami. "Mining association rules between sets of items in large databases." ACM SIGMOD Record 22.2, pp. 207-216, 1993.

[12] Kotsiantis, Sotiris, and Dimitris Kanellopoulos. "Association rules mining: A recent overview." GESTS International Transactions on Computer Science and Engineering 32.1, pp. 71-82, 2006

[13] Berry, Michael J., and Gordon Linoff, "Data mining techniques: for marketing, sales, and customer support", 3rd ed., John Wiley & Sons, Inc., 1997.

[14] Borgelt, Christian. "Efficient implementations of apriori and eclat." FIMI'03: Proceedings of the IEEE ICDM workshop on frequent itemset mining implementations. 2003.

[15] Borgelt, Christian. "Simple algorithms for frequent item set mining." Advances in machine learning II. Springer Berlin Heidelberg, 2010. pp. 351-369.

[16] Borgelt, Christian. "An Implementation of the FP-growth Algorithm." Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations. ACM, 2005.

[17] Introduction to Data Mining, http://www.users.cs.umn.edu/kumar/dmbook /ch6. Pdf.

[18] Marcelo A. Santana, Evandro B. Costa, Baldoino F. S. Neto, Italo C. L. Silva, Joilson B. A. Rego, "A predictive model for identifying students with dropout profiles in online courses" Institute of Computing, Federal University of Alagoas, marcelo.almeida@nti.ufal.br

[19] Liu, B. Hsu, W., Ma, Y., "Mining Association Rules with Multiple Minimum Supports," Proc. Knowledge Discovery and Data Mining Conf., pp. 337-341, Aug. 1999.

BIOGRAPHIES

BOBBY K SIMON, received the Bachelor's Degree in Computer Science and Engineering from Karpagam University, Tamil Nadu, India in 2017. He is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Kerala, India. His area of research includes internet security, data mining and technologies in Department of Computer Science and Engineering.

Anjana P Nair received the Bachelor's degree from LBS Institute of Technology for Women, Kerala, India and master's degree in Computer Science and Engineering from Sree Buddha College of Engineering, Kerala, India in 2013. She is a lecturer in the Department of Computer Science and Engineering, Sree Buddha College of Engineering. Her main area of interest is Core Computers and has published more than 10 referred papers.