



The Survey – Predicting an Education Performance of Students Based on Data Mining Techniques

M. Mayilvaganan¹, D. Kalpanadevi*²

Associate Professor, Department of Computer Science, PSG College of Arts and Science College, Coimbatore, India¹

P.hD Scholar, Department of Computer Science, PSG College of Arts and Science College, Coimbatore, India²

ABSTRACT: First review concern about the relevant literature survey in order to evaluate the academic performance of students from the learning ability and based on CGPA score, family background, extracurricular activities which is analyzed by various data mining techniques. Predicting the student's achievements involves in the analysis on the assessment of the result of data gathered from questionnaires distributed to the respondents. In this paper focus the reviews of performance of various data mining techniques which can be used to give new idea for future work of a scope of research analysis of deep approach to predict performance where an assessment requires analysis, synthesis and critical thinking which can intended for use with students at an individual level, to identify strengths and weaknesses in individual study behaviour. In this research gives a new idea for further work, application of data mining techniques in educational field can be used to develop performance monitoring and evaluation tools system, implies that the skills will be stimulated over time through intentional support and also helps for various resources based on different categorize.

KEYWORDS: Data mining techniques; academic performance; K-nearest neighbor; decision tree; Naïve Bayes; Neural network; regression analysis; ANOVA; Multiple regression; Correlation

I. INTRODUCTION

Data mining is the process of digging through data and looking for meaningful trends and patterns that connect variables in database [1], actually part of the knowledge discovery process. Data mining techniques such as K-nearest neighbor, decision tree; Naïve Bayes, Neural network, Fuzzy, Genetic and other techniques are applied in various environments [3]. This paper describes literature survey on to analyze the academic performance which can give an idea of the scope of research; it will help the educational system to monitor the students' performance in a systematic way.

II. REVIEW - DATA MINING TECHNIQUES INVOLVES EDUCATIONAL ENVIRONMENTS

An application of Data mining [20-22] is a rich focus of Classification algorithm, Association algorithm, Clustering algorithm which can be applied to the field of some resources it concerns with developing methods that discover the knowledge from data originating from any other resource environment. Regression analysis can predict a value of a given continuously valued variable based on the values of other variables, assuming either a linear or nonlinear model of dependency. These tasks are studied in statistics and neural network fields. Deviation Detection – to discover the most significant changes in data from previously measured or normative values.

A. Association Techniques

A frequent item-set typically refers to a set of items that frequently appear together in a transactional data set [27] produce dependency rules which will predict the occurrence of an item based on patterns found in the data. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 6, June 2017

For eg: if (lower_class=Poor, Higher_class=Good) then Performance=Average.

B. Classification Techniques

In Classification process [1, 22] the derive model is to predict the class of objects based on class label which is unknown. In educational data mining, the work of data was predicted by logical rule of the Classification algorithms with the represent of common domain values for analyzing the qualitative performance of required details.

Data mining technique classified as two tasks such as discovery and verification. From Fig.1 represents the taxonomy of data mining techniques, the verification task considers as user hypothesis, while the discovery task searches for unknown knowledge hidden in the data. In general the discovery task can further classified into prediction and description. To predict a value of a given continuously valued variable individually for behaviour of other variables, assuming either a linear or nonlinear. Description can represents to identify a set of categories that describe the data.

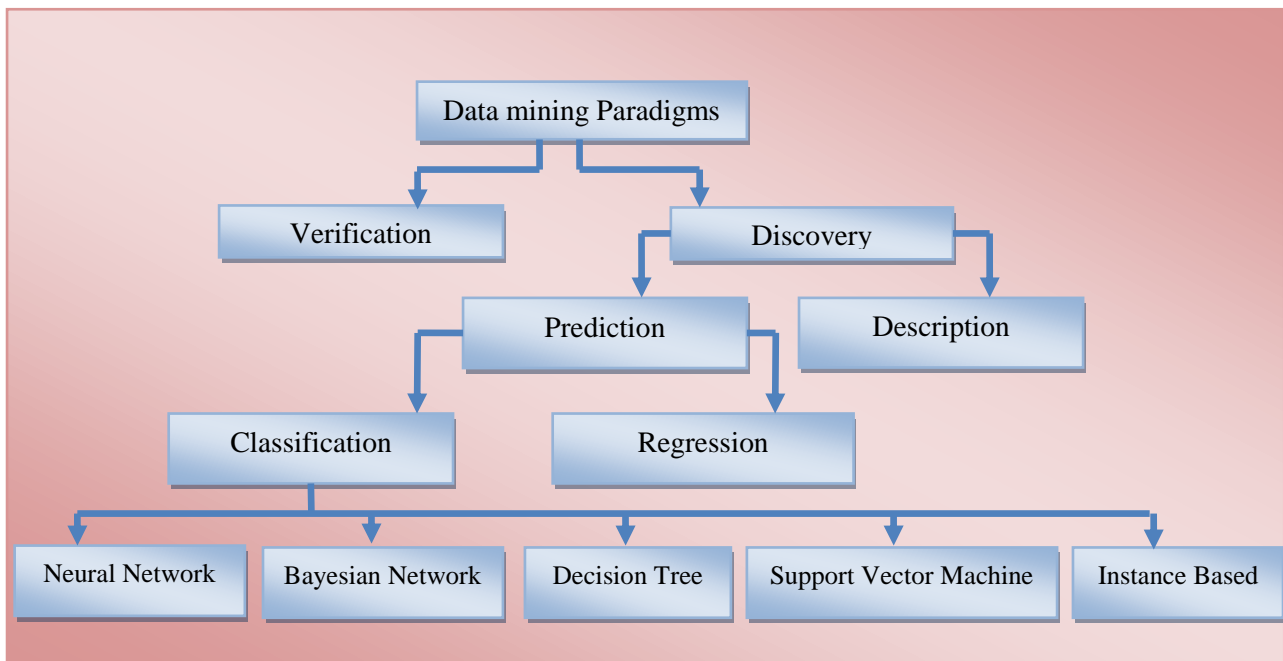


Fig.1 Data Mining Taxonomy

Classification is a data mining technique used to classify each item in a set of data into one of predefined set of classes or groups. Data classification is a two-step process.

- In the first step, a model is built by analyzing the data tuples from training data having a set of attributes. For each tuple in the training data, the value of class label attribute is known. Classification algorithm is applied on training data to create the model.
- In the second step of classification, test data is used to check the accuracy of the model. If the accuracy of the model is acceptable then the model can be used to classify the unknown data tuples.

C. Clustering Techniques

The task is to learn a classification from the data [1, 2]. The representation can then be investigated to see if the data group according to preconceived ideas or to suggest new experiments. The merging or division of clusters is performed according to some similarity measure, chosen so as to optimize some criterion.

Clustering can be used for:

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirce.com

Vol. 5, Issue 6, June 2017

- Exploratory data analysis: visualize the data at hand and get a feel of what the data looks like, what its properties are.
- Generalization: discover instances that are similar to each other, and hence can be handled in the same way.

The representation can then be investigated to see if the data group according to preconceived ideas or to suggest new experiments. The characteristics of clustering techniques can be shown in the fig.2, it is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

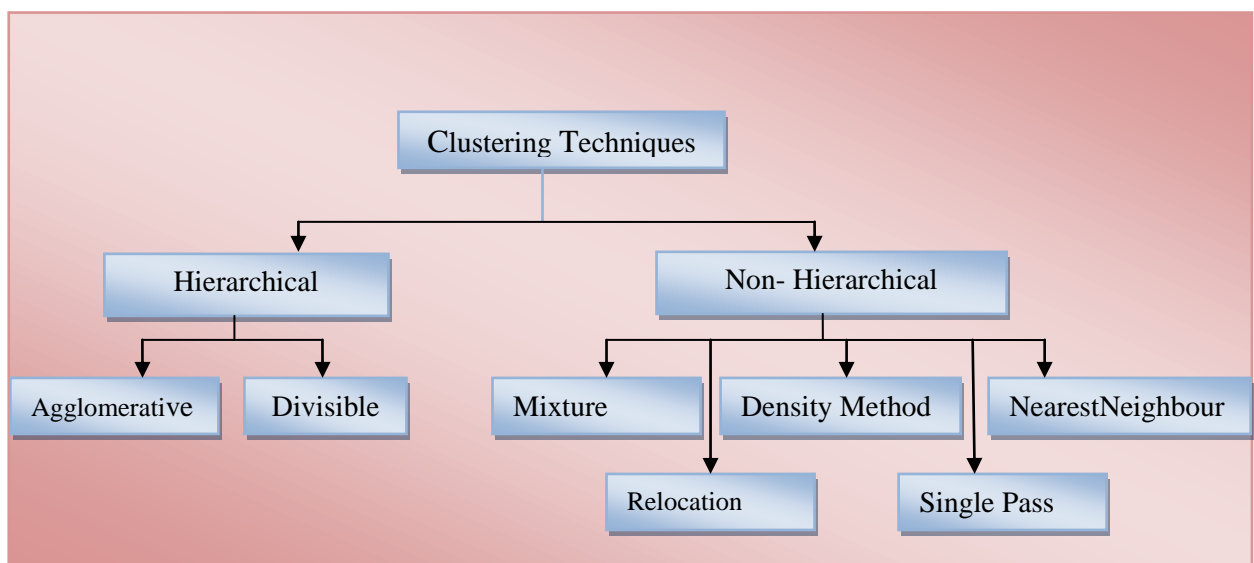


Fig.2 Summary of Clustering Techniques

Correlation depends upon two variables; change in one variable effect a change in second variable. Its value lies in the range of -1 and +1. Whereas in covariance two variables vary together, this can be negative or positive. Correlation is the measure of the strength of the relation between two variables.

The actual value and predicted value agree very closely. In most cases a linear regression line captures this close relationship, in which case slope of the regression line is nearly equal to one. The coefficient of determination R^2 indicates estimation capability of the model used for estimation.

Soraya hakimi. et.al., discussed an academic performance of the students for improving education quality and investments in effective factor regarded country's pervasive development [9]. Based on regression analysis revealed personality traits accounted for 48 percent of the variance in academic achievement. In this research, they confirmed the hypothesis of study and were consistent with much other kind of researches considering conscientiousness as the most reliable predictor of academic performance and conscientiousness is related to academic achievement.

A work had done Jedsarid Sangkapan, et. al., discussed logistic regression analysis can be used for findings, male students' academic probation was higher than their female counterparts investigate the factors affecting academic achievement of students who were put in probation status at Prince of Songkla University [31]. A statistical analysis can be employed to analyze data from questionnaires and interviews. The research supports three main categories of student sides such as Bio-social factor, Affective factor and environmental factor. In Bio-social factor represents education of students, hometown, career, income including personal differences which also affect students' academic probation. An affective factor like Achievement motivation, Study habits, anxiety, adjustment, responsibility, an interaction between instructors and students. An environmental factor consists of parents' support, teaching quality and classroom atmosphere. The final hypothesis can be framed under the three categories and implemented in logistic regression analysis. From the result the Male students were 4.42 times more probation than female students was due to male students spent more time involved in playing games, hanging out with friends, or chatting on the Internet.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirce.com

Vol. 5, Issue 6, June 2017

Educational data of students characteristics can be discussed by Moyosola Jude Akomolafe, in this research correlation analysis can be evaluated for investigate the relationship of personality characteristics to the academic performance of secondary school students [10]. The training data of 398 student details can be constituted relevant of independent and dependent variables were measured. The big five personality dimension such as neuroticism, extraversion, agreeableness, openness to experience and conscientiousness. The facets of neuroticism like angry hostility, Anxiety, Depression, Impulsiveness, Self-consciousness and Vulnerability. The facets of extraversion are Excitement – seeking, Positive emotions Gregariousness, Warmth, Assertiveness and Activity. The facets of openness to experience are Fantasy, Aesthetics, Actions, Feelings, Ideas and Values. The facets of agreeableness are Straightforwardness, Trust, Altruism, Compliance, Modesty and Tender-mindedness. The facets of conscientiousness are Order, Dutifulness, Achievement striving, Competence, Self-discipline and Deliberation. Pearson Product Moment Correlation and Multiple Regression analyses [20-22] can be used to evaluate the factor of the relational ship of the personality characteristics and academics performance of the students in an effective manner to encourage the students to be well organized, proactive about learning and disciplined. Individual students should be encouraged to develop study habits and time management of good academic performance in the education environment. In further they recommend arranging the counselling for identify the students experiencing through educational psychologists and assists an appropriate therapy such as Cognitive Behavioural Therapy, Rational Emotive Behaviour Therapy, Emotional Intelligence Training, Self- Management Technique, Problem Solving Technique and so on to deal with their concerns [5].

Baris Cetin et. al., discussed multiple regression analysis [12] can be used to predicting the college students of academic achievements by identifying the relationship between learning and age are significantly correlated to grade point average (GPA) in early childhood education students.

Fadhilah Ahmad et. al., discussed the prediction of students academic performance based on classification techniques, the 497 training data can be taken from eight of first-year bachelor students in Computer Science course [14]. The domain values can be followed by demographics, family background information and previous academic records. To compare the required techniques such as Decision Tree, Naïve Bayes, and Rule-Based classification [17][21] are applied to the students' data and produced the best students' academic performance prediction model from the highest accuracy. The classified rule base of GPA values of the first year categories into three different classes such as poor, average, and good. An Experiment can be analysed by WEKA tool and result of Rule-Based classification given best model among the other techniques by receiving the highest accuracy value of 71.3%. The students' GPA is selected as a dependent parameter for determining the students' level of success in an efficient way.

J.K. Jothi Kalpana, et. al., discussed education on data mining of new system of predicting the intellectual performance analysis can be evaluated in the primary, secondary and higher level from the year of 2008 to 2013 [16]. This paper improved the investigation of an educational domain of data can be collected from the graduate students from the college of Engineering and Technology Villupuram. In this paper data mining methods of centroid-based, distribution based and density based clustering can be applied for performance analysis using MATLAB [19]. Each one of these tasks can be used to improve the performance of the graduate student. The performance of student's multi-level of optimization formulated by using clustering Experiments attempts to improve the accuracy by using the method of the Gaussian mixture model. The dataset is modelled with a fixed number of Gaussian distribution that is initialized randomly and the parameters are iteratively optimized to fit better to the data set. The density-based clustering method is a linkage based clustering. The range parameter ϵ produces a hierarchical result related to that of linkage clustering. The future work can be recommended on applying data mining techniques on an expanded data set with more distinctive attributes to get more accurate results.

Amirah Mohamed Shahiria, et.al., discussed the higher learning institution in Malaysia challenging the prediction of student academic performance can be investigate due to the lack of affecting factor of students achievement in particular courses with in Malaysian context [18]. First the research work can study the existing method and second the work can be focused to suitable method for predicting the performance in high accuracy for large volume of data in educational database. The criteria of research question can be followed by Kitchenhams steps in structure. The research attributes such as Internal assessments, Psychometric factors, External assessment, CGPA, Student Demographic, High school background, Scholarship, Social network interaction, Student Demographic, High school background, External assessment, Student Demographic, High school Background, Internal assessments, External assessment are taken for analysis and prediction techniques are Naive Bayes, K-Nearest Neighbour, Decision tree, Artificial Neural Networks



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 6, June 2017

and Support Vector Machine can be implemented [21][19]. The result on classification technique of neural network and decision tree method can predicted highest accuracy from the main attribute of CGPA.

Sajadin Sembiring, et. al., discussed the higher learning education system for analysing the academic performance of the students based on data mining techniques to analyze the relationships between students behavioural and their success [32]. The criteria of question can be framed for finding the relationship between behavioural of students and academic performance based on the prediction of SSVM algorithm and clustered using kernel k-means clustering algorithm [21]. In this research, 300 sample data can be carried out for analysing the strong correlation between behaviour and final academic performance. Based on the binary classification process 90% of data can be classified in training dataset and 10% data can be classified in testing dataset. The predicted data can be classified under the rule base of Excellent, very good, good, average and poor for analysing the status of the student performance. Finally they recommend developing the educational field by monitoring and evaluating tool system offering for historical perspectives of students' performances.

d. Outlier Detections

Outlier analysis [28] is to detect an error and remove the problem of several applications. It is a right step to improve the quality of such application in future. In this survey this analysis was approached based on distance based and density based. It will be analyzed by clustering algorithm

- Distance Based Outlier: An approach of general idea is to judge a point based on distance to its neighbors.
- Density Based Outlier: Compare the density around a point with the density around its local neighbors.

III. REVIEW OF COGNITIVE PROCESS AND HUMAN COMPUTER INTERFACE SYSTEM

Human computer interaction is concerned with how people use computer system to perform tasks, usually in a real life work setting [30]. Cognitive processes is the process that involve knowledge, attention, memory, problem solving, producing and understanding the language and decision making. All these are very important for human behaviour. The working process of each task can be analysed by finding the human behaviour [2], attitude and attention performance in respect way.

A. Cognitive Models

In review of cognitive model, scope of knowledge is accumulated information, problem solving schemas, performance skills, expertise, memory capacity, problem representation ability, abstraction and categorization abilities, synthesis skills, long-term concentration ability, motivation, efficiency and accuracy [5, 6].

Data which is collected by using variety of techniques like Video and audio recording, software logging, Scan converters, think- aloud protocols or pencil and paper field notes [29]. These techniques can be followed by several models of cognitive such as GOM model, KLM model, Cognitive complexity which has to be evaluating by language based model such as Command Language grammar, Task Action Language, Task Action Grammar, and Knowledge Analysis of tasks.

In cognitive complexity, the tasks can be analysed by the group of domain variables that have to be related in a single representation [23-26]. For e.g.: The hypothesis ideas such as collecting the academic performance, family background, , extracurricular activities etc., are the basic attributes for analysing the performance skill for required person.

The techniques are outlined for analysis of cognitive complexity in general cognition, cognitive development, mathematics education, reasoning tasks, psychometric test items, and industrial decision making, problem solving etc. The role of questionnaire format on the basis of problem solving, reasoning task, behaviour methods to analyse in effective way.

B. GOM model

GOM model stands for Goals, Operators, Methods, and Selection rules needed to perform a task. Tasks are broken down into their components to predict performance times. Fig.3 represents the process of GOM model describe as

- Goals - are objectives
- Operators - are the actions that change the system state or the cognitive state.
- Methods - are description of procedures for achieving goals stored in the users Knowledge structure of the task built-up for the problem solving.
- Selection Rules - are If ... Then statements to enable the user to choose between the methods under the time complexity.

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 6, June 2017

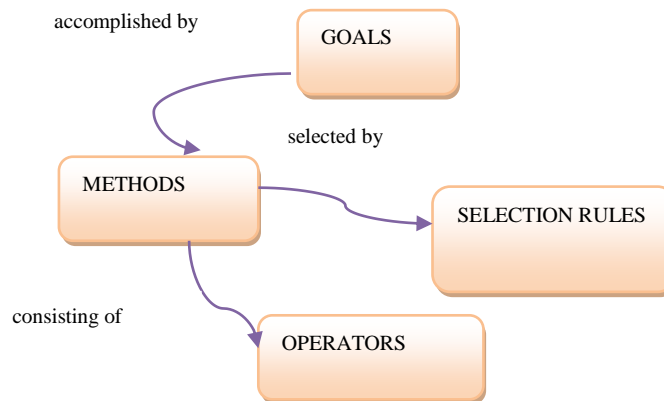


Fig.3 GOM model

C. Keystroke Level Model (KLM)

KLM is derived from-GOMS [4] [7]and describes the time taken to execute sub-task using the system facilities. Total time taken for an action is arrived at by simply adding together the times for each component task. To obtain the predicted time for a task and add the times for individual operators based on Fitt law, Steerling Law.

D. Cognitive Complexity Theory

Cognitive complexity theory is an extension of GOMS [8]. It attempts to predict how difficult learn and use a system will be base on a GOMS model of the task and its required knowledge, a model of the user current knowledge and a list of the items of knowledge to be learned in order for the user to be able to make error- free use of the system [23-26].

E. Knowledge Analysis of Tasks (KAT)

KAT is an evaluate model to identify the task gathered from variety of techniques including interviews and questionnaires, observation, rating scales, repertory grids and conduct online test for problem solving. The completed tasks will analysis for the performer by producing the result. KAT involves several stages such as

- Identify the person goals, sub goal and subtasks
- Work out order in which sub goal are to be carried out.
- Identify task strategies.
- Identify procedures.
- Identify Task Objects and Actions.

F. Observation methods of evaluation

Observation evaluation may be either laboratory based or carried out in the field form to evaluate performance with end users and either prototypes or complete system. Quantitative data is collected and analysed relating to task performance [3]. For e.g., Time to complete a task, number of task completed number of times help consulted etc. Qualitative data is also collected about how users go about performing tasks their attitudes to the system etc.

IV.SCOPE OF RESEARCH- AN IDEA OF PREDICTING THE PERFORMANCE OF STUDENTS IN EDUCATIONAL ENVIRONMENT

From the survey of the research work, an idea of predicting the performance of students be recommend the suggestion of future work by self observance, behaviour, problem solving, education task, reasoning task and other resources will be handle for arranging the counselling for identify the students experiencing. A Scope of research Analysis from review of cognitive process in the Human computer Interface system contributes better performance of the human behaviour which will be analyse by data mining techniques to evaluate the knowledge of person in efficient manner.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirccce.com

Vol. 5, Issue 6, June 2017

The variety of domain values which are related with performance based on their required result carried out by cognitive process model followed by the GOMS model, KLM model, and Cognitive complexity theory for end-user testing through knowledge task analysis integrates with data mining techniques design at the stage of effectiveness, learning ability, flexibility, attitude which is evaluated by problem solving analysis, an area of human thinking, behavior, analyse the performance from the knowledge criteria are gathered by the techniques of data mining [1]. An ability of cognitive performance is essential in various environment, which is influenced by many qualitative attributes are included for forming the data set.

V. CONCLUSION AND FUTURE WORK

In this review, it can be concluded that the research work of the survey gave a new idea of predicting the process of students skill analysis through human computer Interface integrates with Data mining techniques with respect to cognitive models of GOMS model, KLM model, cognitive complexity theory for analyzing human behavior, performance, skills, attention, memory level etc are gathered by using several parameters of self observance, behavior, problem solving, education task, reasoning task and other resources.

REFERENCES

1. Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao. "Data mining techniques and applications – A decade review from 2000 to 2011", Expert Systems with Applications, Elsevier (39):11303–11311, 2012.
2. Bainbridge, L. "The change in concepts needed to account for human behaviour in complex dynamic tasks", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 27, 351–359.
3. Arbi Ghazarian, "Pauses in man-machine interactions: a clue to users' skill levels and their user interface requirements", Int. J. Cognitive Performance Support, Vol. 1, No. 1, 2013.
4. G. D. Abowd. Classroom 2000: an experiment with the instrumentation of a living educational environment. IBM Systems Journal, 38(4):508–30, Special issue on Human–Computer Interaction: a focus on pervasive computing, 1999.
5. P. Barnard, M. Wilson and A. MacLean. Approximate modelling of cognitive activity with an expert system: a theory-based strategy for developing an interactive design tool. The Computer Journal, 31(5):445–56, 1988.
6. M. Blattner, D. Sumikawa and R. Greenberg. Earcons and icons: their structure and common design principles. Human–Computer Interaction, 4(1):11–44, 1989.
7. S. K. Card, T. P. Moran and A. Newell. The keystroke-level model for user performance with interactive systems. Communications of the ACM, 23:396–410, 1980.
8. S. K. Card, T. P. Moran and A. Newell. The Psychology of Human–Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, NJ, 1983.
9. Soraya hakimi, Elaheh Hejazi, Masoud Gholamali Lavasani, "The Relationships Between Personality Traits and Students' Academic Achievement", Procedia - Social and Behavioral Sciences, 2011, 29 : 836 – 845 doi:10.1016/j.sbspro.2011.11.312.
10. Moyosola Jude Akomolafe, "Personality Characteristics as Predictors of Academic Performance of Secondary School Students", Mediterranean Journal of Social Sciences, 2013 4(3), 657-664, Doi:10.5901/mjss.2013.v4n2p657.
11. Age diseth, "Personality and Approaches to Learning as Predictors of Academic Achievement", European Journal of Personality, 2003, 17: 143–155 DOI: 10.1002/per.469.
12. Baris Cetin, "Approaches To Learning And Age In Predicting College Students' Academic Achievement", Journal of College Teaching & Learning, 2016, 13(1), 21-28.
13. Angus Duff *, Elizabeth Boyle, Karen Dunleavy, John Ferguson, "The relationship between personality, approach to learning and academic performance", 2004, 36: 1907–1920, doi:10.1016/j.paid.2003.08.020.
14. Fadhilah Ahmad*, Nur Hafieza Ismail and Azwa Abdul Aziz, " The Prediction of Students' Academic Performance Using Classification Data Mining Techniques", Applied mathematical Sciences, 2015, 9(129), 6415 – 6426, <http://dx.doi.org/10.12988/ams.2015.53289>.
15. Ruby-Ann B. Dela Cruz., Ryan Manuel D. Guido, "Factors Affecting Academic Performance of BS Astronomy Technology Students", International Journal of Engineering Research & Technology (IJERT), 2013, 2(12), 84-94, ISSN: 2278-0181.
16. J.K. Jothi Kalpana, K. Venkatalakshmi, "Intellectual Performance Analysis of Students by Using Data Mining Techniques", International Journal of Innovative Research in Science, Engineering and Technology, 2014, 3(3), 1922-1929 ISSN (Online) : 2319 – 8753.
17. Tomas Chamorro-Premuzic, Adrian Furnham, Martin Lewis, "Personality and approaches to learning predict preference for different teaching methods", Science direct, 2007, 17: 241–250, doi:10.1016/j.lindif.2006.12.001.
18. Amirah Mohamed Shahiria, Wahidah Husaina, Nur'aini Abdul Rashida," A Review on Predicting Student's Performance using Data Mining Techniques", Procedia Computer Science, 2015, 72 : 414 – 422, doi: 10.1016/j.procs.2015.12.157.
19. Rajn Jindal and Malaya Dutta Borah, "A Survey on Educational Data Mining And Research Trends", International Journal of Database Management Systems, 2013, 5(3): 53-73, DOI : 10.5121/ijdms.2013.5304.
20. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nded., Morgan Kaufmann Publishers, 2006.
21. Arun K Pujari, "Data mining techniques", University Press (India) Private Limited.
22. David Hand, Heikki Mannila & adhraic Smyth, "Principles of Data Mining", MIT Press, 2001.



International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirce.com

Vol. 5, Issue 6, June 2017

23. Richard E.Clark et.al., "Cognitive Task Analysis", October 14, 2006.
24. David H. Jonassen, "Analysis of Task Procedures", Copyright _@ 1986.
25. Anderson, J.R., 'The Architecture of Cognition', Harvard University Press, Cambridge (1983).
26. A. Monk, editor, Fundamentals of Human-Computer Interaction. Academic Press, London, 1985.
27. Sheikh,L Tanveer B. and Hamdani,S., "Interesting Measures for Mining Association Rules". IEEE-INMIC Conference December. 2004.
28. M. O. Mansur, M.Sap and M. Noor, "Outlier Detection Technique in Data mining: Research Perseptive", In Postgraduate Annual Research Seminar, 2005.
29. P. Barnard and J. May. Interactions with advanced graphical interfaces and the deployment of latent human knowledge. In F. Paternó, editor, Proc. of Design, Specification and Verification of Interactive Systems, DSVIS'94. Springer-Verlag, 1995.
30. A. Dix. Managing the ecology of interaction. Proceedings of Tamodia 2002 – First International Workshop on Task Models and User Interface Design, Bucharest, Romania, 18–19 July 2002
31. Jedsarid Sangkapan, Kasetchai Laeheim, "Factors Affecting Students Academic Achievement into Probation Status at Prince of Songkla University", In proceeding of The 3rd International Conference on Humanities and Social Sciences, April 2, 2011, Community Empowerment.
32. Sajadin Sembiring, et.al., "Prediction of Student Academic Performance By An Application of Data Mining Techniques", in proceeding of International Conference on Management and Artificial Intelligence , 2011, IPEDR vol.6, 110-114, IACSIT Press, Bali, Indonesia.

BIOGRAPHY



M. Mayilvaganan, Associate Professor, Department of Computer Science, PSG College of Arts and Science, Coimbatore. He has more than 21 years teaching and research experience in the field of Computer Science. His area of research includes data mining, big-data analysis, and networking. He has published more than 25 international publications and 13 in IEEE journals.



D. Kalpanadevi is a research scholar of Computer Science. She has completed MCA, M. Phil and pursuing her Ph.D. in Computer Science at PSG College of Arts and Science from Bharathiar University, Coimbatore, India. Her research area of interest includes data mining. She has attended six international conferences and presented her paper and published three indexed Journal.