International Journal of Pure and Applied Mathematics
Volume 119 No. 16 2018, 2261-2271
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue
http://www.acadpubl.eu/hub/

Forecasting Reasons for Students' Health Failure in Tertiary Institutions

J. K. Alhassan[1], S. Adebanjo[1], M. Sanjay [2], Robertas Damaševičius[3],
Rytis Maskeliūnas[3]

[1] Federal University of Technology, Minna,, Nigeria
[2], Covenant University, Ota,Ogun State, Nigeria
[3]Kaunas University of Technology, Kaunas, Lithuania
jkalhassan@futminna.edu.ng,ssopam@gmail.com, {robertas.damasevicius,Rytis.maskeliunas}@ktu.lt

Abstract. One of the prime requirement of healthcare administration is to bring excellence service to patients, by making comprehensive conclusions. Such conclusions can only be made with the existence of satisfactory knowledge derivative from healthcare data that cannot be acquired by simple observation. The usage of data mining is very supportive in health care administration for forecast and conclusion making, since it is a convergence of numerous disciplines, which comprises database management systems (DBMS), Statistics, Artificial Intelligence, and Machine Learning. In this study, Decision tree technique was used to forecast the periodic causes of students' health failure. Four dissimilar decision tree models were articulated, which comprise the J4.8, Random Forest, Random Tree and Decision Stump. This was attained by performing a 10-fold cross validation on a dataset containing of seven nominal variables and ninety instances. It was detected that the J48, which is the Weka's implementation of the C4.5 decision tree model performed better with 76.6% accuracy, 0.829 precision and 0.708 recall than the remaining three models.

Keywords: Forecasting, Reasons, Students, Health Failure, Data Mining, Decision Tree.

1 Introduction

The dawn of computers, in what way they have advanced over the years has gone a lengthy way in refining how work is completed. Mainly, persons in office, secretaries, office assistant and managers similar have found computers valuable particularly in the handling of data. This principally is what data processing is all around. The usage of computers in processing data. The necessity to form a warehouse for data has been required by the degree at which data has developed over the years. This requirement has required organization to have databases and data storeroom for a varied array of their data. The information industry is rich with data. Unless these data are converted to appropriate information, they are unusable. To gain information that is beneficial, it is vital to mine these data. Information extracting is not the only process needed; the mining of data or rather knowledge discovery includes processes like cleaning the data, integrating the data, transforming the data, mining the data, evaluating patterns and presentation of data, [1] and [2].

A method of receiving information out of data by mining it is demarcated as data mining. In principle, it can be regard as the procedure of gaining knowledge out of data, [3], The understanding or information acquired can be relevant in: fraud detection, market analysis, production control and science exploration, [3]. Data mining is very valuable in the following arenas: Market analysis, fraud detection in a management, cooperate analysis and risk management, [4]. Apart these, it is also usable in manufacturing control, retention of customers, scientific exploration, sport, and web surf-aid. According to [5], Data mining (an analysis step in the process " Discovering of Knowledge and Mining of Data") which is an interdisciplinary sub-field in computer science is a computational procedure of unearthing patterns out of huge sets of data consisting methods that harmonizes machine learning, database system, artificial intelligence, and statistics. The general notion of the data mining process is to gain information from a dataset and change it into an understandable form for further use, [5]. The term "Data mining" was obtainable in the 1990s, but it is the evolution of a area having a lengthy past, [6].

Separately from afore stated ranges of applying data mining, it might likewise be used in the healthcare area, [7]. The healthcare area is one that produces enormous dimensions of data connecting to patients, hospital properties, diagnosis of diseases, electronic records of patients, and medical gadgets. The use of data mining in health care has been predominant over the ages, mainly in kidney dialysis, and in forecasting the survival chance in heart disease in patients. Recently, the cases of students visiting the school clinic at a specific period of the academic session for health-related challenges have been on the rise. Following this drift to reach at concrete deductions in other to offer suitable answers is not one that can be done by careless or less thorough observation. Therefore, the necessity to have a system that can forecast students' health failure in tertiary institutions.

Study has it that great fraction of circumstances of poor academic performance in student is mostly owing to poor health situations amid numerous other factors, [8]. The usage of data mining is found to be very supportive in health care administration for forecast and conclusion making, [9]. Therefore, this study applied data mining in forecasting reasons of students' health failure in tertiary organizations, using Federal University of Technology, Minna as example.

2 Literature Review

[10] focused on the use of various algorithms used in predicting a combination of different target attributes. They offered an efficient and intelligent approach for pre-

dicting heart attack using data mining. The approach was to extract relevant patterns from the heart disease data warehouses for efficiently predicting heart attack. Using a calculated significant weightage, frequent patterns that had values exceeding a predefined threshold were used for the relevant prediction of heart attack.

Home healthcare is a way of providing care to patients at the comfort of their home. It is an important area of study. [11] evaluating performance for home healthcare practices in the US, applied the classification and regression tree (a non-parametric approach) on a dataset comprising of hip replacement, heart failure and chronic obstructive pulmonary disease in determining determinants of outcomes in home healthcare service, that is, length of stay and discharge destination. From this study it was observed that patients with age 85years and above were the driving force in discharge destination and length of stay for all three health situations considered. The CART procedure was adequate in correctly classifying patients in all three conditions which suggests continuing utility in home health care.

[12] in their study gave an overview of different decision tree algorithms ranging from ID3, C4.5, k-NN and SVM. The study captured the application of these four decision tree algorithms in diagnosing treatment effectiveness, hospital infection control and clinical data. Their study only captured the advantages of these methods in various domains.

Data mining comprises of different approaches. These approaches serve varying purposes, which have both pros and cons. Data mining task are either predictive or descriptive. [13] in this study made an outline of some data mining approaches and their applicability to healthcare data for prediction and decision making. They highlighted a difference between traditional data mining and healthcare data mining to be that typical data mining is aimed at description rather than explanation of patterns and trends. According to [13] one difficulty of data mining in medical practice is the voluminousity and heterogeneity of uncooked medical data. Before data mining will take place collection and integration of data must be done, and it is quite expensive to build a data warehouse before beginning data mining. Data classification is a two-way procedure, comprising: a learning step (where we build a model for classification). This is done by analyzing the training data using a classification algorithm and classification step (where a model for predicting the class labels in a particular data is applied). A test data is used to estimate the accuracy of the classification rule.

[14] applied data mining technique called association rule mining in attractive the quality of student’s performances. They mined rules that aids to forecast the performance of the students and it recognize poor, good and excellent students. The performance account of the student likewise aids to progress the result of the student. In their study, they approved that their system also aids to find those students which require special care to decrease fail ratio and taking suitable action for the next semester examination.

[15], used genetic programming and different data mining methods to foresee student failure at school. They projected a genetic programming model to get precise and understandable classification rules. They maintained that forecasting student failure at school is a tough duty not just because it is a multifactor problem but likewise because the accessible data are normally imbalanced. They used real data from high school

students in Mexico. Though, how they joint different data mining methods to attain their result was not clear from their report.

[16] applied data mining approach to discover useful data and legitimate pattern. They established a way to forecast the final status of students based on their continuous valuation test and attendance status. Using data mining methods, they resultant rules that allow the classification of students in their foretold classes. They proclaimed in their study that the use of k-means and Apriori they were able to find patterns based on their academic records, which normally is not possible by going through the huge database manually.

[17] applied decision tree and genetic algorithms to investigate academic performance of students using homework assignments, and effort to derive short rules that explain and forecast success or failure in the final exams. They argued in their investigation that genetic algorithm-based induction of decision trees has a latent for developing into an alert tool for early prediction.

[18] presented a generic, multi-criteria model based on fuzzy logic ideas to build decision support system for admitting student into the university. He used Evolutionary computation techniques, fuzzy Logic, and genetic Algorithms. He firstly worked on a first order aggregation model and performed its learning stage by using Steady State Genetic Algorithms. The main focus of his paper is to develop intelligent decision support system which spots the lights on the problems of multi-criteria decision that dealing with subjective and indefinite data.

3 Materials and Methods

3.1 Data Collection and Variables (Input and Target)

The research requires dataset of health-related cases of students to be collected for use in this research work. The data were gotten from a random sample of students using questionnaire. It consists of 7 (6 nominal attribute and 1 nominal class) and 90 instances.

There are two classes of variables used to carry out the research: the dependent variables and the independent variables. The dependent variables are those variables whose values depend on that of another. On the other hand, the independent variables are those variables that stand alone and are not changed by the other variables being tried to measure.

Independent Variables: Academic period, stress, dirty environment, poor dieting, poor medication, and weather season. These are independent variables as shown in Table 1, because they are measured and are the ones who determine the status of the dependent variables.

Dependent Variable(s): Class label. This is the predicted variable. The class label has two sub-dependent variables: yes, or no. The variable with the highest prediction rate becomes the class label.

Table 1. Variables and their descriptions

Attributes	Description	Label
Academic period	Academic period	Early semester, mid semester, exam
Stress	Stress from academic work and other commitments	True, False
Dirty environment	Dirty toilets in the hostel and bath rooms, bushes and poor drainage.	True, false
Poor dieting	Poor feeding habit, attending lectures and exams on empty stomachs	True, false
Poor medication	Self-medication, not taking medications as prescribed	True, false
Weather season	Seasonal weather condition	Rainy, Dry
Class	The probability that a student fell ill as a result of one or more factors	Yes, No

Data collected were cleansed by pruning in order to rid it of inconsistencies or errors. This is a fundamental data mining process. In this case, the data were cleansed manually through repeated careful observation of the data. This pruning operation helps to tune the data.

.

3.2 Data Mining Tool Used

Weka, which stands for Waikato Environment for Knowledge Analysis, is a popular machine learning software developed by Waikato University, in New Zealand. It is an open source software, presented under the GNU General Public License. Its workbench includes a collection of tools for visualization and several algorithms for analyzing data and predictive modelling, and also a graphical user interfaces to easily access this functionality. It is a software built using Java that is able to run on all platforms. The algorithms can either be run from Weka interface or an application programmer interface.

Weka is a collection of machine learning algorithms used in solving real-world data mining problems. It runs on almost all major platform including Microsoft Windows, Apple Mac and Linux. There are various techniques for running the data mining process, and many of these are taken care of in the software package. Data can be fed into the software using a direct way which encompasses loading in a file or interfacing with the software using the Java programming platform by coding. The latter is time-consuming and gives the same result as the former. The techniques include the Classify, Cluster, Associate, and so on. Each technique requires different parameters to run.

The software is a two-in-one combo comprising of both command line interface and a graphical user interface. First, the software is clicked on in the Installed Programs windows or by clicking on the weka.jar file located at the installation folder. Thereafter, a widow pops up showing four buttons: Explorer, Simple CLI, Experimenter and Knowledge Flow**.** The Explorer is used to carry out most of the data mining tasks. Clicking on it opens up a window with a couple of menu tabs such as Preprocess, Classify, Classify, Associate, Select Attributes, and Visualize. There are as many as 10 other buttons scattered around the graphical user interface of the application.

3.3 Data Processing

The computer used for the processing is a Dell Latitude E6400 Note book PC. Summary of configuration are as follows:

Processor: Intel Core 2 Duo with Nvidea™ HD Graphics 2.53 GHz

Installed Memory: 1.50GB

System Type: 64-bit Operating System

Though the research is using classification method one of the techniques in data mining, there are subclasses of methods under it. Among them is decision tree, the one used in this research. Others are Support Vector Machine and Naïve Bayes Classifier. Decision tree classification is a machine learning model which decides the target value (a variable that depends on some other variables to get an output) from a new sample based on different attribute values of the existing data.

How Decision Tree Works: The J48 decision tree follows the following algorithm. In order to order a new item, it first needs to produce a decision tree based on the attribute values of the existing training data. So, when it meets a set of items it identifies the attribute that discriminates the various instances most visibly. This characteristic that is able to identify the data instances so that it can classify them the best is said to have the highest information gain.

Decision trees are formed by algorithms that find several ways of rending a dataset into branch-like parts. These parts form an inverted decision tree that starts with a root node at the top of the tree. The object of analysis is reflected in this root node as a simple, one-dimensional representation in the decision tree interface. Decision trees attempt to point out solid relationship between input values and target values in a set of observations that makeup a data set. When a set of input values is recognized to having a solid relationship to a target value, then all of these values are assembled in a bin that becomes a branch on the decision tree. Decision trees provide exclusive capabilities to complement and substitute for the following: traditional statistical methods of analysis (like multiple linear regression); varying data mining tools and techniques (like neural networks) and multidimensional forms of report and analysis used in business intelligence. Decision trees usually consist of three different types: decision nodes – represented by square; chance nodes – represented by circle; and end nodes – represented by triangle. Commonly, decision trees are drawn using flowchart symbols, since they are very easy to understand.

3.4 Experimental Design

To predict the cause of student health breakdown, various classifications model was built using J4.8, Random Forest, Random Tree, and Decision Stump. Decision trees can handle continuous and categorical variables and can be used for both classification and regression. They automatically handle interactions between variables and identify significant variables which it uses to make “intelligent” predictions. Extensive data pre-processing resulted in a clean dataset comprising 90 instances void of missing values. After evaluating the data and selecting the predictive models to be used, a series of experiments were performed.

Three evaluation criteria were used to evaluate the results; they are accuracy, precision and F-measures.

Accuracy -the accuracy measures the percentage of correctly classified samples.

$$\text{Accuracy} = \frac{TP+FP}{TP+FP+TN+FN} \tag{1}$$

Precision - precision measures the proportion of positive patterns that are correctly classified as positive.

$$\text{Precision} = \frac{TP}{TP+FN} \tag{2}$$

F-measures -it combines both positive prediction value and precision

$$\text{F-measures} = \frac{2xFP}{2xTP+FP+FN} \tag{3}$$

Where:

TP is true positive of the correctly classified class positive
TN is true negative of the correctly classified class negative
FP is the false positive of the correctly classified class positive
FN is the false negative the correctly classified class negative

4 Results and Discussion

WEKA’s explorer generally chooses reasonable defaults, the J4.8, Random Forest, Random Tree, and Decision Stump algorithm were performed using their default parameters. The classification task for each model were performed using a 10-fold cross validation. Table 2 shows the combined results of the models.

Table 2. Combined results of the classification models

Classification Technique	Accuracy (%)	Class	Precision	Recall
Decision Stump	75.56	Yes	0.861	0.646
		No	0.685	0.881
Random Tree	70.00	Yes	0.723	0.708
		No	0.674	0.69
J48	76.60	Yes	0.829	0.708
		No	0.714	0.833
Logistic Model Tree (LMT)	74.40	Yes	0.821	0.667
		No	0.686	0.833

As seen in Table 2, Random Tree is found to perform least next to LMT followed by Decision Stump, and J48 decision tree having the best performance for the classification. Fig 1 shows the decision tree classification results.

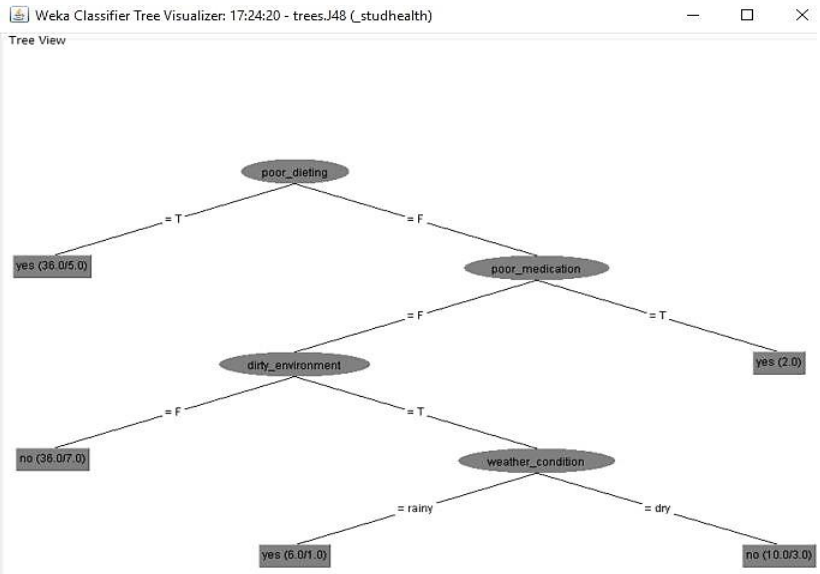

Fig. 1. Decision tree classification result

Four decision tree algorithms were used using a 10-fold cross validation in WEKA tool and the results were compared. From the result presented in table 4.1, the Decision stump model has 75.60%, 86.1% and 73.8% of accuracy, precision and accuracy respectively, Random tree has 70%, 72.3%, 71.6% of accuracy, precision and accuracy respectively, J48 has 76.6%, 82.9%, 76.4% respectively while Logistic model tree has 74.4%, 82.1%, 73.6% respectively. The results present the J48 model to be of better performance in accuracy and f-measure.

5 Conclusion

Based on the success of the J48 decision tree model in predicting causes why student fall ill in tertiary institution, it is recommended to health care organizations in tertiary institutions in determining type of drugs to be purchased and also in determining period where more manpower might be needed. The healthcare sector is one rich in terms of data, making the application of data mining to mine data and gain useful knowledge realistic. Therefore, it is recommended that all records be automated, for easy access and use. The scope of this study can be extended by making use of more data with more set of attributes to classify causes of student health breakdown.

References

1. Giraud-Carrier, C., Povel, O. : Characterising Data Mining Software. Intelligent Data Analysis, *7*(3), (2003) 181–192
2. Ngai, E. W. T., Xiu, L., Chau, D. C. K. : Expert Systems with Applications Application of Data Mining Techniques in Customer Relationship Management : A Literature Review and Classification. Expert Systems With Applications, *36*(2), (2009) 2592–2602. http://doi.org/10.1016/j.eswa.2008.02.021
3. Liao, S.-H., Chu, P.-H., Hsiao, P. Y. : Data Mining Techniques and Applications – A Decade Review from 2000 to 2011. Expert Systems with Applications, *39*(12), (2012) 11303–11311. http://doi.org/10.1016/j.eswa.2012.02.063
4. Wells, J. T. : Occupational Fraud: The Audit as Deterrent. Journal of Accountancy, 193(4), (2002) 24–28.
5. Kapoor, A.: Data Mining : Past , Present and Future Scenario, *3*(1), (2014) 95–99
6. Sharma, M.: Data Mining : A Literature Survey, 9359(2), (2014) 1–4.
7. Harris, J. B. F, Ryan, E. T, Calderwood, S. B. : Cholera. Lancet, 379, (2012) 2466–2476
8. Daniels, D. Y. : Examining Attendance, Academic Performance, and Behavior in Obese Adolescents. The Journal of School Nursing : The Official Publication of the National Association of School Nurses, *24*(6), (2008) 379–387. http://doi.org/10.1177/1059840508324246
9. Kirkos, E., Spathis, C., and Manolopoulos, Y. (2007). Data Mining techniques for the detection of fraudulent financial statements, *32*, 995–1003. http://doi.org/10.1016/j.eswa.2006.02.016
10. Rani, B. K., Govrdhan, A., Srinivas, K., Rani, B. K., Govrdhan, A.: Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering, 02(2010), 250–255. http://doi.org/10.1.1.163.4924
11. Madigan, E. A., Curet, O. L. : A Data Mining Approach in Home Healthcare: Outcomes and Service Use. BMC Health Services Research, *6*(1), (2006) 1–10. http://doi.org/10.1186/1472-6963-6-18
12. Dave, M., Dadhich, P. : Applications of Data Mining Techniques : Empowering Quality

Healthcare Services, 1(1), (2013) 13–16

13. Milovic, B., Milovic, M. : Prediction and Decision Making in Health Care using Data Mining Corresponding Author :, 1(2), (2012) 69–76.
14. Charanjit B., Minakshi B., Nirali M. Vandana M.: Mining Association Rules in Student Assessment Data. International Journal of Advanced Research in Computer and Communication Engineering, 3(3),(2014)
15. Carlos M., Alberto C., Cristóbal R. Sebastián V: Predicting Student Failure at School using Genetic Programming and Different Data Mining Approaches with High Dimensional and Imbalanced Data, Appl. Intel. (2013) DOI 10.1007/s10489-012-0374-8
16. Khyati M., Madhuri R.: Predicting Students Performance in Higher Education: A Data Mining Approach. International Journal of Scientific & Engineering Research, 5(2), (2014)
17. Kalles D. Pierrakeas C. : Analyzing Student Performance in Distance Learning with Genetic Algorithms and Decision Trees. Imitational Federation for Information Processing, Volume 204, Artificial Intelligence Applications and Innovations, eds. (2006) 9-18
18. Wafa S. : Steady State Genetic Algorithm in University Admission Decision. Technology Journal, 4(4), (2013) 32–36