Prediction of Students' Academic Performance based on their lifestyle through Machine Learning Methods

Thendral Puyalnithi[1], Madhu Viswanatham V[2] and Mithilesh Kumar Singh[3]

[1] School of Computing Science and EngineeringVIT University,Vellore, Tamilnadu, India
thendral.p@vit.ac.in, thendral_psg@yahoo.com
[2] School of Computing Science and EngineeringVIT University,Vellore, Tamilnadu, India
vmadhuviswanatham@vit.ac.in
[3] School of Computing Science and EngineeringVIT University,Vellore, Tamilnadu, India
,savanmorya@gmail.com

ABSTRACT

In this paper we are going to analyse the addiction of student towards alcohol in early life and how that can affect their life in other words Children who start to drink by age 13 are more likely to go on to have worse grades, to skip school and, in the worst case scenario, to be excluded from school. We are going to use data set of students on some courses which was collected and analysed by University of Minho, Portugal. Our work intends to approach student addiction on alcohol in secondary level using Data Mining (DM) techniques. The result shows that a good predictive accuracy can be achieved, provided that addiction of alcohol can impact to the student performance. In addition, the result also provides the correlation between alcohol usage and the social, gender and study time attributes for each student. As a direct outcome of our paper, more efficient prediction tools can be developed in order to pay more attention to the student and share how the alcohol impact so badly in his life. We have used the Classification and clustering algorithm of the data mining to analysis the result.

Keywords: Classification algorithms, Machine Learning, Classification accuracy comparison, Data Mining, Disease diagnosis, Orange

1. INTRODUCTION

Alcohol had lots of bad impact in our life. Drinking Too much on a single occasion or over time can Take a serious toll on our health. If who drinks alcohol it's likely he is experienced first-hand at least some of its short-term health effects, be it a hangover or a bad night's sleep. Alcohol have many short-term and long term health effects. Taking alcohol as teenager age, reduce a child's mental and physical abilities, a effecting judgment and co-ordination which can lead to trouble. The level of alcohol gets so high that the brain's vital functions, which include breathing control, are blocked. Alcoholics were more likely to get injured or have accidents than non-drinkers. More worrying still, they're more likely to be a passenger in a drink-driving incident. When children drink, their decision making skills are a effected and they're more likely to take big risks like having unprotected sex. That can lead to sexually transmitted diseases and unwanted pregnancy. Now a days the student who are under 18 are falling under different bad activity like alcohol consumption as we mentioned and many peer behaviour there are various factor to affect any student performance like their background, where they living, parent status, previous pears activity. We are taking in consideration the all the factors and predicting the performance of the student based on whatever we have the datasetby use of this dataset we are training the models and by using the trained data we are testing the new data.

The researchers used pattern recognition and data mining methods in predicting models. The experiments were carried out using classification algorithms Naïve Bayes, Decision Tree, K-NN and Neural Network and results proves that Naïve Bayes technique outperformed other used techniques. The researchers uses K-means clustering algorithm on a student performance warehouse to extract data relevant to data prediction, and applies MAFIA (Maximal Frequent Item set Algorithm) algorithm to calculate weightage of the frequent patterns significant to student performance predictions.

The researchers proposed a layered neuro-fuzzy approach to predict occurrences of coronary education system, simulated in MATLAB tool. The implementation of the neuro-fuzzy integrated approach produced an error rate very low and a high work efficiency in performing analysis for coronary student failure occurrences. The researchers also proposed a new approach for association rule mining based on sequence number and clustering transactional data set for grade predictions. The implementation of the proposed approach was implemented in C programming language and reduced main memory requirement by considering a small cluster at a time in order to be considered scalable and efficient.

The researchers used the data mining algorithms decision trees, naïve Bayes, neural networks, association classification and genetic algorithm for predicting and analysing student report from the dataset. An experiment performed by the researchers on a dataset produced a model using neural networks and hybrid intelligent algorithm, and the results shows that the hybrid intelligent technique improved accuracy of the prediction.

The research paper describes the prototype using naïve Bayes and weighted associative classifier (WAC) to predict the probability of student receiving less marks. The researchers developed a web based intelligent system using naïve Bayes algorithm to answer complex queries for prediction of report and help the parent of the student to take the responsibility of the student.

The researcher uses association rules representing a technique in data mining to improve prediction with great potentials. An algorithm with search constraints was also introduced to reduce the number of association rules and validated using train and test approach. Three popular data mining algorithms (support vector machine, artificial neural network and decision tree) were employed by the researchers to develop a prediction model using 502 cases. SVM became the best prediction model followed by artificial neural networks.

The researcher's uses decision trees, naïve Bayes, and neural network to predict final report with 15 popular attributes as risk factors listed in the dataset.

Two kinds of data mining algorithms named evolutionary termed GA-KM and MPSO-KM cluster the cardiac disease data set and predict model accuracy. This is a hybrid method that combines momentum-type particle swarm optimization (MPSO) and K- means technique. The comparison was made in the research conducted using C5, Naïve Bayes, K-means, Ga-KM and MPSO-KM for evaluating the accuracy of the techniques. The experimental results showed that accuracy improved when using GA-KM and MPSO-KM.

2. PROPOSED METHOD

2.1 Dataset

We use a data set about Portuguese student

which was composed by Paulo Cortez and AliceSilva, University of Minho, Portugal. In Portugal,The secondary education consists of 3 years of schooling, preceding 9 years of basic education and followed by higher education. Most of the students join the public and free education system. This study will consider data collected during the 2005-2006 school year from two public schools, from the Alentejo region of Portugal. Hence, the database was built from two sources: school reports, based on paper sheets and including few attributes; and questionnaires, used to complement the previous information.

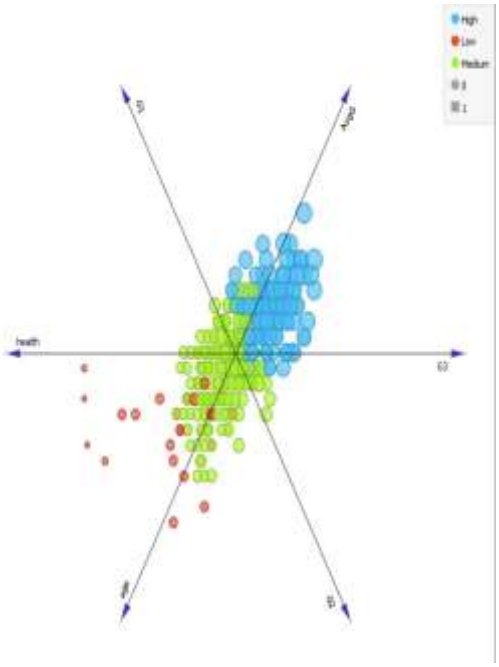
Data Set Characteristics	Multivariable
Attribute Characteristics:	Integer
Associated Tasks:	Classification, Regression
Number of Instances	649
Number of Attributes	36
Missing Values	NA
Area	Social
Date Donated	2014-11-27

2.2 Components of the Model

These proposed methods can be said to be a combination of 4 stages as the order is mentioned in the following context.

The first stage is to pre-process the given data for any outlier detection. This stage is called the pre-processing stage, where the model gets rid of faulty data and outliers.

2.3 Outlier detection is a primary step in many data mining applications. We present several methods for outlier detection, while distinguishing between univariate vs. multivariate techniques and parametric vs. nonparametric procedures. In presence of outliers, special attention should be taken to assure the robustness of the used estimators. Outlier detection for data mining is often based on distance measures, clustering and spatial methods.



Once the outliers are detected and removed we move to the next stage of the model. This stage is to form decision models or trees based on the input data received from the last stage. For our purpose of study we are going to choose four different models, namely, SVM, Random Forest, Naïve Bayes Classifier and Classification Tree.

2.4 SVM Classifier: support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

2.5 Naïve Bayes: These classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

2.5 Random Forest: These are a notion of the general technique of random decision forests that are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

2.6 Decision Tree (Classification Tree): Here learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called **regression trees**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. This page deals with decision trees in data mining.

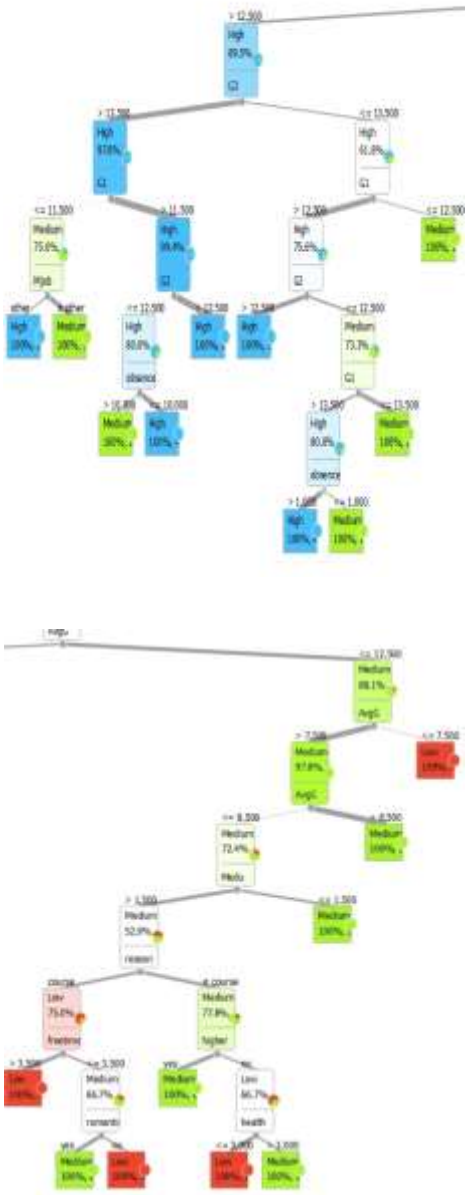


Fig 2: Decision Tree for given Data Set

Now we have four different classifiers in hand. We have to choose the best among them. This is where the training data comes in handy. We now use the training data to measure the accuracy of the various methods obtained. The best one among them is to be chosen and further to be forwarded to next stage for prediction analysis.

We have three methods to test the accuracy of the data. The classifiers are tested on all possible tests and the one with better results is taken forward for prediction.

These methods of validation have different parameters based on which they are judged.

AUC: It is an abbreviation for *area under the curve*. It is used in classification analysis in order to determine which of the used models predicts the classes best. An example of its application is ROC curves.

CA: Classification Accuracy is a method to know the accuracy of the classifier on valid testing data. It is the ratio of correct predictions made to total predictions made.

Recall: The training data used to model the classifier is further used again to test the recalling abilities of the modelled classifier.

F1: In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.

Precision: Precision is the number of correct positive results divided by the number of all positive results.

These are the various parameters of the validation methods based on which the classifier is judged for its dependability.

Most commonly used methods to evaluate the classification methods accuracy are **Leave-One Out** and**Cross Validation.**

2.7 Leave One Out: the simplest and a commonly used method of cross validation in chemo metrics is the "leave-one-out" method. The idea behind this method is to predict the property value for a compound from the data set, which is in turn predicted from the regression equation calculated from the data for all other compounds. For evaluation, predicted values can be used for *PRESS*, *RMSPE*, and squared correlation coefficient criteria.

Method	AUC	CA	F1	Precision	Recall
NB	0.889	0.898	0.566	0.417	0.878
CT	0.911	0.974	0.828	0.820	0.837
RF	0.561	0.934	0.218	1.000	0.122

Fig 3: Leave-One Out analysis results

Cross Validation: This technique sometimes called rotation estimation is a model validation technique for assessing how

the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of *known data* on which training is run (*training dataset*), and a dataset of *unknown data* (or *first seen* data) against which the model is tested (*testing dataset*). The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the *validation dataset*), in order to limit problems like overfitting, give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem), etc.

Method	AUC	CA	F1	Precision	Recall
NB	0.868	0.894	0.543	0.402	0.837
CT	0.934	0.982	0.878	0.878	0.878
RF	0.571	0.935	0.250	1.000	0.143

Fig 4: Cross Validation Analysis Results

The next stage is to predict the results of the various classes. The best method for the training data set is definitely Classification Tree or Decision Tree as it has the best results for Classification Accuracy and Recall for both validation cases.

Now we forward this classifier to next stage to predict the disease that may further lead to a cardiac arrest. Some where we may compare these model but we only focus on the Classification Tree.

2.8 Tool

We have used the orange tool throughout the completion of the paper, Orange is a free software machine learning and data mining software (written in Python). It has a visual programming front-end for explorative data analysis and visualization, and can also be used as a Python library. We can use the already defined features or can use our script in python to do the same.

3. RESULTS ANALYSIS

The results are compared using a confusion matrix. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. Or in other word I can say that confusion matrix is a useful tool for analysing how well your classifier can recognise tuples of different classes.

We have analysis the performance of the data using all four model let us consider the naïve Bayes classifier the corresponding confusion matrix of the dataset.

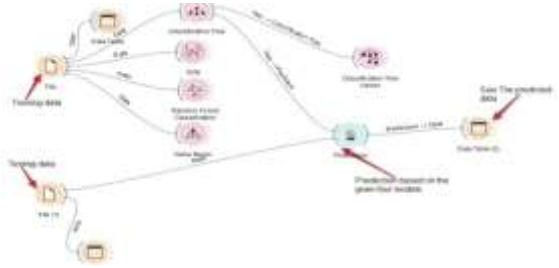
		Predicted			
		High	Low	Medium	∑
Actual	High	204	0	9	213
	Low	0	46	3	49
	Medium	24	60	303	387
	∑	228	106	315	649

Fig 5: Naïve Bayes Confusion Matrix

We are now considering only the classification Tree model because of its performance.

		Predicted			
		High	Low	Medium	∑
Actual	High	212	0	1	213
	Low	0	41	8	49
	Medium	7	0	380	387
	∑	219	41	389	649

The prediction of the student performance are done on the basis of the classification tree classifier and with help of the orange tool



4. CONCLUSION

In this paper a by using the classification method to predict the performance of the student are proposed. The student has less family support are consistently poor in their grade or if they are started use of alcohol also getting less marks in their class. The most effective way to keep student growing, we need to consider that they should have all the basic requirements, good parent support, good peer to peer connection all these factor leads towards the good performance of the student. Many student are failing because of the bad peer connection, if they have not having any bad habit they will take it from their bad peer groups. This prediction system may provide easy and a cost effective way to predict the performance of the student based on the training data and their background support. This system can also use the new data for learning purpose and will be used for the next prediction of the data.

REFERENCES

[1] P. Cortez and A. Silva. Using data mining to predict secondary school student performance.[in a.brito and j. teixeira eds. proceedings of 5th futurebusiness technology conference (f ubutec2008)pp:5-12 portoportugal]. 2008.

[2] V.Krishnaiah "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" International Journal of Computer Science and Information Technologies, Vol. 4 (1) 2013, 39 – 45 www.ijcsit.Com ISSN: 0975-9646.

[3] Jiawei Han MichelineKamber, Jian Pei.DataMining: Concepts and Techniques. Morgan Kaufmann,third edition.

[4] A. Sahar "Predicting the Serverity of Breast Masses with Data Mining Methods" International Journal of Computer Science Issues, Vol. 10, Issues 2, No 2, March 2013 ISSN (Print):1694-0814| ISSN (Online):1694-0784 www.IJCSI.org

[5] Rajashree Dash "A hybridized K-means clustering approach for high dimensional dataset" International Journal of Engineering, Science and Technology Vol. 2, No. 2, 2010, pp. 59-66

[6] RituChauhan "Data clustering method for Discovering clusters in spatial cancer databases" International Journal of Computer Applications (0975-8887) Volume 10-No.6, November 2010.

[7] Dechang Chen "Developing Prognostic Systems of Cancer Patients by Ensemble Clustering" Hindawi publishing corporation, Journal of Biomedicine and Biotechnology Volume 2009, Article Id 632786.

[8] S M Halawani "A study of digital mammograms by using clustering algorithms" Journal of Scientific & Industrial Research Vol. 71, September 2012, pp. 594-600.

[9] ZakariaSulimanzubi "Improves Treatment Programs of Lung Cancer using Data Mining Techniques" Journal of Software Engineering and Applications, February 2014, 7, 69-77

[10] Labeed K Abdulgafoor "Detection of Brain Tumor using Modified K-Means Algorithm and SVM" International Journal of Computer Applications (0975 – 8887) National Conference on Recent Trends in Computer Applications NCRTCA 2013

[11] Alaa. M. Elsayad "Diagnosis of Breast Cancer using Decision Tree Models and SVM" International Journal of Computer Applications (0975 – 8887) Volume 83 – No 5, December 2013

[12] NeelamadhabPadhy "The Survey of Data Mining Applications and Feature Scope" Asian Journal of Computer Science and Information Technology 2:4(2012) 68-77 ISSN 2249-5126

[13] S. Santhosh Kumar "Development of an Efficient Clustering Technique for Colon Dataset" International Journal of Engineering and Innovative Technology" Volume 1, Issue 5, May 2012 ISSN: 2277-3754

[14] RafaqatAlam Khan "Classification and Regression Analysis of the Prognostic Breast Cancer using Generation Optimizing Algorithms" International Journal of Computer Applications (0975-8887) Volume 68- No.25, April 2013

[15] K.Kalaivani "Childhood Cancer-a Hospital based study using Decision Tree Techniques" Journal of Computer Science 7(12): 1819-1823, 2011 ISSN: 1549-3636

[16] Boris Milovic "Prediction and Decision Making in Health Care using Data Mining" International Journal of Public Health Science Vol. 1, No. 2, December 2012, pp. 69-78 ISSN: 2252-8806

[17] T.Revathi "A Survey on Data Mining Using Clustering Techniques" International Journal of Scientific & Engineering Research Http://Www.Ijser.Org, Volume 4, Issue 1, January-2013, Issn 2229-5518

[18] ShomonaGracia Jacob "Data Mining in Clinical Data Sets: A. Review" International Journals of Applied Information System (IJAIS) - ISSN: 2249-0868 Foundation of Computer Science FCS, New York, USA, Volume 4-No.6, December 2012-www.ijais.org

[19] G. Rajkumar " Intelligent Pattern Mining and Data Clustering for Pattern Cluster Analysis using Cancer Data" International journal of Engineering Science and Technology Vol. 2(12), 2010, ISSN: 7459-7469.

[20] M. Durairaj "Data Mining Applications in Healthcare Sector: A Study" International journal of Scientific & Technology Research, Volume 2, Issue 10, October 2013, ISSN: 2277-8616

[21] VikasChaurasia "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability" International journal of Computer Science and Mobile Computing (IJCSMC), Vol.3, Issue. 1, January 2014, pg.10-22, ISSN: 2320-088X

[22] T.Sridevi "An Intelligent Classifier for Breast Cancer Diagnosis based on K-Means Clustering and Rough Set" International Journal of Computer Applications (0975 – 8887) Volume 85 – No 11, January 2014

[23] Com (S rl). Villars sous Yens.Global status report on alcohol and health.Technical report, World Health Organization, 2014.

