Identifying Key Performance Indicators and Predicting the Result from Student Data

J. Shana
Department of MCA,
Coimbatore Institute of Technology,
Tamilnadu-641014, India

T. Venkatachalam
Department of Physics,
Coimbatore Institute of Technology,
Tamilnadu-641014, India

ABSTRACT
Student information systems hold a lot of information that can be mined for useful patterns. This work aims to build a prediction model to predict the result of students in 'C' Programming course by analyzing the factors that affect the performance of students. We applied feature selection techniques to select the most relevant academic and non-academic factors. The model is implemented using various classification algorithms and it is found that Naïve Bayes classification model gives the highest accuracy of 82.4%.Decision tree based algorithm also showed considerable accuracy of 80.2%.The model was trained using 182 records from student dataset collected from the college with 20 attributes within the year 2008 to 2010. The model was validated using test records. It would predict the class label 'Result' as categorical value, Pass or Fail. Such a prediction model would help the faculty in early identification of 'at risk' students and thereby take timely and proactive measures to improve their performance.

Keywords
Feature Selection, Classification, Prediction, Data mining

1. INTRODUCTION
All data mining projects require working with very large datasets that involve many attributes. Dimensionality reduction is a very daunting task for many applications of data mining especially in predictions. There are many feature selection methods available [1, 4, 5] for all domains. Student dataset involves many attributes which act as predictors. In the Indian scenario the attributes of student dataset that affect performance is different from the students of foreign universities. Students in India do not come in different age groups nor there retention problems in courses. Online courses are very few in India. The attributes that influence a student's performance are entirely different in India. Educational data mining has become a very useful area [3]. In this work we show how we analyzed and applied different feature selection techniques to identify the most probable factors contributing to the success of students in a course. When most influential variables are known researchers can proceed with the prediction model building process with ease [7].Here the final result (pass/fail) in a subject is considered as performance. We analyzed the dependency of the class attribute 'result' on other attributes. For this we tried to study the correlation between the attributes using Pearson's Correlation coefficient and F-Test. Also Chi-Square analysis was performed to identify the degree of dependency. And methods like Information gain and Gain ratio used in tree building were also applied. Later these selected set of attributes were used to build the prediction model and its accuracy was tested. A student dataset of 182 records who have taken up 'C' Programming course has been used for this study. The dataset initially contained 20 attributes.

1.1 Research Objectives
The main goal of this research is to analyze the student data at Computer Technology Department who have taken 'C' Programming course and perform the following:

1. Identify the key performance indicators that affect the result (success or failure) of the student in the course.
2. Analyze various classification models and identify a high accuracy prediction model to predict the result.

This paper is organized as follows. In Section 2 we explain about the various methods employed to identify the most influencing attributes from the student dataset. Section 3 describes the mode of experiment and Section 4 discusses the results. Section 5 concludes the work.

2. FEATURE SELECTION METHODS
2.1 Correlation Analysis
Karl Pearson's coefficient of correlation is the widely used method for measuring the degree of relationship between two variables. This gives an insight into the dependency of each of the attributes on the 'result' attribute [9]. Also F-Test on the attributes specifies the significant difference in the variance of the attributes involved in the study. Both techniques helped in knowing the association between attributes.

2.2 Chi-Square Analysis
Pearson's chi-square test of independence is a statistical method used to identify degree of association between variables [8]. This technique is applied to analyze the dependency of all attributes (factors) on the outcome attribute. So chi-square method proves useful here. For a contingency table that has 'r' rows and 'c' columns the formula for finding the chi-square is as given in equation (1).

$$\chi 2 = \frac{\Sigma(observed-expected)^2}{expected} \quad (1)$$

The predetermined level of significance is taken as 5% and P-values are identified using the chi-square values for each of the attributes.

2.3 Information Gain Analysis
Information can be represented in bits. Given a probability distribution the required information to predict an event is the distribution's entropy. It is calculated using the equation (2) given below.

$$\text{Entropy}(p1..pn) = -p1\log(p1) - p2\log(p2) \ldots\ldots - pnlog(pn) \quad \ldots (2)$$

where p1…pn are number of instances of each class, expressed as a fraction of the total number of instances at that point in the tree and log is base 2.The smaller the entropy greater will be the purity of the subset partitions.

Let A be the set of all attributes and *Ex* the set of all training examples, val (x, a) which defines the value of a specific example *x* for attribute $a \in Attr$, *H* specifies the entropy as in equation (2). The information gain for an attribute $a \in Attr$ is defined as in equation (3).

$$IG(Ex, a) = H(Ex) - \sum_{v \epsilon val(a)} \frac{|\{x \epsilon Ex | val(x,a)=v\}|}{|Ex|} H(\{x \epsilon Ex | val(x, a) = v\}) \quad .. (3)$$

This is used in selecting the most appropriate attributes for building the classifier tree. This set of attributes affect the final outcome attribute.

2.4 Gain Ratio Analysis

Information gain has one limitation that it prefers attributes with many values. To avoid this problem Gain ratio is calculated for each of the attribute using equation (5). This method is also used as a tree splitting criteria in classifiers. The intrinsic value calculation for a test is defined as in equation (4), where n is the number of examples left in the class after the test on the attribute.

$$IV(Ex, a) = -\sum \frac{|\{n \epsilon Ex\}|}{|Ex|} * \log_2 \left(\frac{|\{n \epsilon Ex\}|}{|Ex|} \right) \quad .. (4)$$

Information Gain Ratio is then calculated as follows:

$$IGR(Ex, a) = IG/IV \quad ... (5)$$

3. EXPERIMENTAL METHODOLOGY

3.1 Data Collection and Transformation

The dataset for this study was taken from all the courses which offered 'C' programming as one of its subject at the same under graduate level. Part of the data was extracted from the student information system of the institution and the rest through questionnaires. The data thus collected were later transformed into categorical values as needed by the feature selection techniques. For example, family income attribute was categorized into three ranges High, Medium and Low. Similarly, other attributes like subject difficulty level, medium of instruction, place of stay were also transformed into categorical attributes.

The experiment also took the support of the WEKA 3.x tool which is a widely used open source data mining research tool [10] and OpenStat statistical software. From the knowledge and experience of domain experts, first a set of key dimensions that would most influence the student performance were identified from the initial data set. Table 1 shows the academic and non academic factors that tend to influence the performance of students in a course. It is on this set that we applied the feature selection methods to further reduce the dimensionality.

Table 1: Initial category of attributes selected by domain experts for the study

Family Background
Factor 1: Annual Income
Factor 2: Nativity
Factor 3: Education of parents
Schooling Information
Factor 1: Location of school
Factor 2: Medium of instruction
Factor 3: Marks in high school (HSC)
Factor 4: Marks in Secondary school(SSLC)
Factor 5: Core Subject in school
Academic Information
Factor 1: Faculty approach
Factor 2: Subject difficulty level
Other Personal Info
Factor 1**:** Stay
Factor 2: Social contacts inside the campus
Factor 3: Interest in subject

Using the above knowledge we derived a set of 13 attributes that would influence the student's performance in C programming course. All the four feature selection methods are applied to the dataset consisting of 13 attributes from the initial set of 20 attributes. Feature selection would further filter this attribute set and help to identify the most relevant factors.

3.2 Feature Analysis

Feature selection is the process of removing features from the data set that are irrelevant with respect to the task that is to be performed. Feature selection can be extremely useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy [4].
Feature analysis was performed in two ways. Initially to study how the variables were related two statistical techniques namely Pearson's correlation coefficient and F-Test were applied. It tested the dependency between the 'Result' attribute and other attributes. Table 2 shows the values of both the analysis. Secondly to determine the degree of dependency, Chi-Squared values with 5% level of significance was calculated.

Table 2: Values of Correlation coefficient and F-Test for each of the attributes in descending order

Attributes	Pearson's Coefficient	Attributes	F-Test
Stay	0.203012	Previous Skill	0.909187479
Subject Difficulty	0.183157	HSC Percentage	0.393660611
Senior Secondary marks.	0.153442	Native	0.27235583
Staff Approach	0.141674	SSLC Percentage	0.245409837
HSC Percentage	0.132312	Stay	0.18099089
Previous Skill	0.11304	Staff Approach	0.101155324
Family Income	0.078257	Motivation	0.03482057
Motivation	0.07478	Friends	0.020908705

Medium of Instruction	0.035057	Medium of Instruction	3.19027E-05
Interest in Subject	0.010774	Family Income	3.80093E-07
Native	-0.0357	Interest in Subject	1.6655E-10
Friends	-0.13474	Subject Difficulty	6.16111E-13

Table 2 gives an initial understanding about the probable factors that contribute to the performance of the students. Two commonly used feature selection techniques in tree building namely, Information gain and Gain ratio were also applied. A final set of attributes were selected from the above feature analysis for building the prediction model.

3.3 Model Building

The prediction model can be built using various classification algorithms and one that gives the best prediction accuracy. Decision tree algorithms are more easy to understand because they can be converted to If-Then rules which can be implemented easily[1, 2]. The model was trained using 182 records with the key predictors with cross validations of 10. The trained model was later tested for its accuracy. Algorithms such as decision tree induction, REP tree, Naïve Bayes classifier and CART were applied to the data for analysis.

4. RESULTS AND DISCUSSION

4.1 Feature Selection

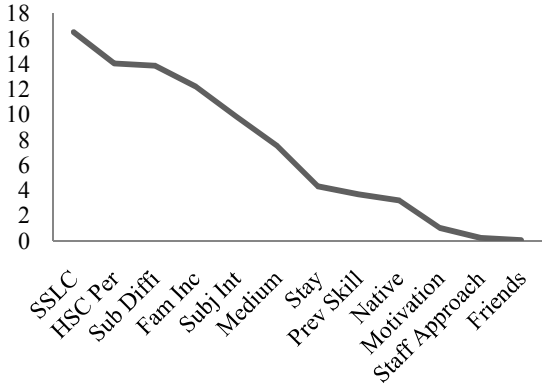

Figure 1. Chi-Square values for the attributes

Attributes with high chi-square values implies highly influential factor.Figure1 shows these values in the ascending order. And Figure 2 shows attributes in the ascending order of the information gain values.

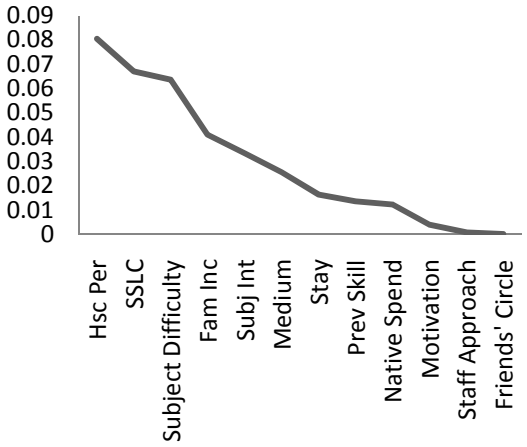

Figure 2. Information Gain values for the attributes

In Figure 3 we can see the attributes that has the highest Gain ratio values. Factors like school marks, difficulty level of subject, family income and interest in subject largely influence the result of students in the course.

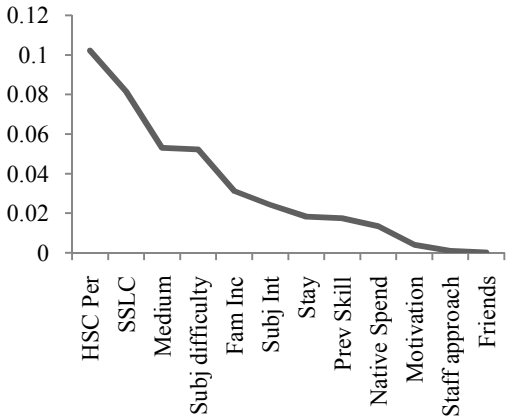

Figure 3. Gain ratio values for the attributes

From all the above feature selection analysis we can conclude upon the factors that tend influence the result of the students. Table 3 shows the list of attributes based on the ascending order of the values for all each of the feature selection methods.

Table 3: High ranked attributes according to the different feature selection methods.

Pearson's coefficient	F-Test	Chi-Sq values	Info gain	Gain Ratio
Stay	Previous Skill	SSLC marks	HSC marks	HSC marks
Subject Difficulty	HSC marks	HSC marks	SSLC marks	Medium
Friends	Native	Subject level	Subject level	Family Income
Staff Approach	SSLC marks	Family Income	Family Income	Stay
HSC	Stay	Subject Interest	Subject Interest	Native
Previous Skill	Staff Approach	Medium	Medium	Staff Approach

From this feature analysis result the attributes with high rank in all the methods is considered in the final list. Table 4 lists the seven attributes selected for model building

Table 4: Final set of attributes

Key Performance Indicators
SSLC percentage
HSC percentage
Subject Difficulty
Family Income
Stay
Medium
Staff Approach

4.2 Prediction Model

The model was trained using four different classification algorithms using the key performance factors. Table 7 shows the accuracy of these classifiers. The Naïve Bayes technique showed an highest accuracy of 82.4% compared to all other methods. Decision tree induction algorithm also showed an acceptable level of accuracy.

Table 5: Percentage of correctly classified instances

Methods	Accuracy (%)
DecisionTree Induction	80.2
RepTree	77.3
Simple CART	74.7
Naïve Bayes	82.4

Figure 4. Classification tree for Decision tree induction algorithm

Classification rules can be derived from the tree in Fig 4. IF-THEN rules as such as the following can be derived.

- IF SSLC percentage=average AND Family Income=low AND staff approach =high AND medium=low THEN Result="PASS"
- IF SSLC=average and Family Income= medium and Medium=high THEN Result="PASS"
- IF SSLC=high THEN Result= "PASS"

These rules can be implemented in any high level language to predict the result of new student who is yet to take the course. The accuracy of such a prediction is found to be 80.2%. For the Naïve Bayes classifier probability needs to be found to predict the result for unseen data.

5. CONCLUSION

In this study we tried to identify from an initial set of 20 attributes a list of 7 factors that influence the performance of Indian students in a 'C' Programming course. Performing correlation analysis and using different feature selection methods a set of factors that influence the result of students have been derived. The prediction model built using Bayes classifier showed highest accuracy of 82.4% and can predict unseen data. The academicians can take measures to improve students if they know that higher secondary marks (HSC), medium of instruction and subject difficulty level contribute to a students' success in their course. This is not the exhaustive list of factors because history of data available when increased would bring out other factors also. Future works can concentrate on working on a larger dataset with other attributes not taken into account here such as assignment marks and attendance and include the type of subject namely theory or practical.

6. REFERENCES

[1] Jiawei Han and Micheline Kamber, 2009. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.

[2] Arun K.Pujari, 2005, Data Mining Techniques, Universities Press (India) Private Limited.

[3] Baker R.S.J.D., and Yacef K, 2009, The state of educational data mining in 2009: A review and future visions, Journal of Educational Data Mining, I, 3-17.

[4] Shyamala Doraisamy, Shahram Golzari, Noris Mohd. Norowi, Md Nasir B Sulaiman, Udiz, 2008, A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Mala Music, ISMIR- Session 3A - Content Based Retrieval, Categorization and Similarity.

[5] M.A Hall, and L.A Smith, 1998, Practical feature subset selection for machine learning, In Proceedings of the 21st Australian Computer Science Conference, pp181-191.

[6] Sk. AlthafHussainBasha, A Govardhan, S.ViswanadhaRaju, Nayeen Sultana, 2010, A Comparative Analysis of Prediction Techniques for Predicting Graduate Rate of University, European Journal of Scientific Research, Vol 46, No 2, pp.186-193.

[7] Huan Liu, Hiroshi Motoda, Rudy Setiono, Zhen Zharo, 2008, Feature Selection: An Ever Evolving Frontier in Data Mining. JMLR: Workshop and Conference Proceedings,10: 4-13, The Fourth Workshop on Feature Selection in Data Mining.

[8] Anne F. Maben, 2005, Chi-square test adapted from Statistics for the Social Sciences.

[9] C.R Kothari, 2006, Research Methodology-Methods and Techniques, New Age International (P) Limited.

[10] Weka available at: http://www.cs.waikato.ac.nz/~ml/weka/ for weka learning software.