ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Determining Key Performance Indicators and Envisioning Results Using Data Mining

[1] Dr. V.B.Narsimha
Asst. Professor, Dept. Of Informatics, University College, Osmania University, Hydetabad

Abstract: **- Information Systems or databases holds enormous amounts of information/data which can be used for mining to extract the effective patterns which can be further used in making better decisions in business, healthcare, banking and other applications. During this research we tend to build a model which predicts the performance of a Student in a Programming Language named 'Python' based on the different factors like the high school marks, parent's education, nature of environment, faculty approach, difficulty level of subject, etc. During the research, we have applied different feature selection techniques using algorithms like Naive Bayes and Decision Tree and found that Naive Bayes have built a prediction model with accuracy of 82.4%. This model helps to understand about the students who may fail which in turn give a sign to faculty to focus on those students to take active measures in improving their performance.**

Keywords **:- Data Mining, Prediction, Accuracy, KPI, Key Performance Indicators, Classification, Decision Tree Induction.**

I. INTRODUCTION

Data Mining is an integrative subject and can be exemplified in different ways. If you consider mining gold from rocks or sand, we say gold mining instead of rock or sand mining. Similarly, data mining should also be named appropriately saying Knowledge mining from data but it makes the name so long. Many people find data mining as a synonym for a popularly used term known as Knowledge Discovery from Data also described as KDD [1]. The Knowledge discovery process consists of a lot of iterative sequence of steps which includes Data Cleansing, Data Integration, Data Selection, Data Transformation, Data mining, Pattern Evaluation and finally Knowledge Presentation [2].

II. FEATURE SELECTION TECHNIQUES:

During this research, we have used some of the feature selection techniques like Correlation Analysis, Chi-Square Analysis, Information Gain Analysis and Gain Ratio Analysis to further transform the dataset in bringing the most important attributes for prediction.

Correlation Analysis:
One of the most widely used method for measuring the degree of relationship between two attributes is Karl Pearson's Method. This analysis gives an observation into the dependency of attributes on the result attribute. Applying F-Test on the attributes indicates the difference in the variance of attributes involved. The two techniques guided in knowing the association between attributes [3].

Chi-Square Analysis:
Chi-Square test is a method to determine degree of association between attributes and it is basically applied to analyze the dependency on the outcome [4]. If you consider a table has 'r' rows and 'c' columns, the formula for finding the chi-square can be written as

$$\chi 2 \quad = \Sigma\,(observed - expected)2/expected$$

Information Gain Analysis:
Data can be characterized in bits. If a probability distribution is provided, all you need is the distribution's entropy to predict an event [5]. Entropy is calculated using the following mathematical equation
Entropy(p1...pn) = -p1 * log(p1) - p2 * log(p2) .... -pn * log(pn)
Where p1,p2...pn are the instances and log is taken with base 2.

Gain Ratio Analysis:
There is one limitation if we use Information Gain Analysis that it gives the attributes with more values and to avoid this, we use Gain Ratio Analysis [6].

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

III. RESEARCH METHODOLOGY

Data Collection and Transformation:
In this research, we have taken a dataset containing all the courses which offers Python Programming at the graduate level. This is actually taken from the Student Information System of the institution. We decided to perform feature selection technique in order to transform the data into different categories. For instance, if you consider family income as an attribute, it is classified into three spaces namely High, Medium and Low. Correspondingly, other attributes like subject complexity, medium of instruction, and place of stay were also reconstructed into distinct categories.

This experiment includes the use of Weka 3.x tool which is one of the widely used open source data mining research tool. Considering the experience and knowledge of domain experts, a set of key dimensions which influence the student performance the most were identified using the dataset. Table 1 depicts the two main factors namely academic and non-academic which tends to influence the performance of students in a course. On this particular dataset, we enforced the feature selection techniques to further shorten the extensity.

Considering the knowledge, we derived a total of 13 attributes which would influence the student's efficiency in Python Programming course. The feature selection techniques namely Correlation Analysis, Chi-Square Analysis, Information Gain Analysis and Gain Ratio Analysis are applied to this dataset containing 13 attributes transformed from the existing dataset of 20 attributes filtering the data and figuring out the most compatible factors.

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

Table 1: Data Collection form Datasets

Feature Analysis

Feature Analysis is a series of actions for removing the trivial features from the dataset according to the prescribed task to be performed. This can be immensely fruitful in contracting the dimensionality of data using classifier, which in turn reduce the execution time and improving predictive accuracy. This technique is performed in two ways. At the beginning, we performed two statistical techniques like Pearson's Correlation and F-Test to study how variables are inter related which in turn gave us the dependency between the Result and other attributes. Later, Chi-Squared values with 5% level of significance were calculated to determine the degree of dependency. Table 2 shows the analysis of Pearson's Coefficient and F-Test.

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

Table 2 shows the analysis of Pearson's Coefficient and F-Test.

This table gives a basic understanding of possible factors which contribute to the performance of the students. Two of the commonly used feature selection techniques used in tree building, Information gain and Gain Ratio were also applied to transform a final set of attributed for building the accurate prediction model.

Model Building:
The Prediction model can be built using various classification algorithms which gives the best prediction with accuracy. Decision tree algorithms are considered to be the most easiest to understand because they can be transformed to If-Then rules which makes the implementation easy. This model was trained using 182 records with key predictors of 10. Further, the model is tested for accuracy. Naive Bayes Classifier, CART and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Decision Tree Induction are the algorithms used for data analysis to name a few.

IV. OBSERVATIONS AND DISCUSSIONS:

Feature Selection:

Attributes with high Chi-Squared values indicates highly influential factor. Figure 1 depicts these values in ascending order and Figure 2 displays attributes in ascending order of information gain values.

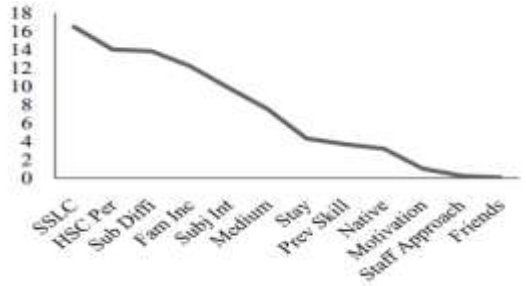

Figure 1 depicts these values in ascending order

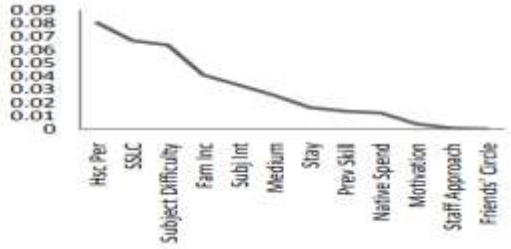

Figure 2 displays attributes in ascending order of information gain values

Figure 3 clearly shows that attributes like school marks, difficulty level of subject, family income and interest in subject highly influence the result having the highest Gain ratio values.

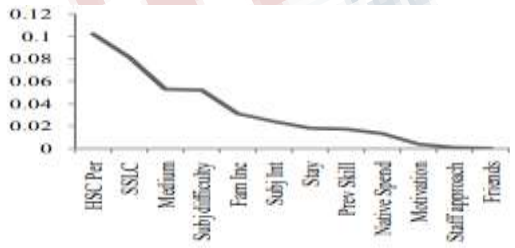

Figure 3 clearly shows that attributes like school marks, difficulty level of subject

Based on the above feature selection analysis, we can conclude upon the factors that tend to influence the result of the students. The above table displays the list of attributes based on the ascending order of the values for all each of the 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

Table 3: list of attributes based on the ascending order of the values for all each of the feature selection methods

Prediction Model:

Based on the performance factors, the model was trained using four distinct classification algorithms. The below table displays the accuracy of these classifiers. Naive Bayes technique predicted the highest accuracy of 82.4% compared to other. Decision tree induction algorithm also showed an acceptable level of accuracy.

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

Table 4: Accuracy of the classifiers

V. CONCLUSION

During the entire research, we demonstrated to analyze from a list of factors which influence the enforcement of Indian Students in a course named Python Programming. We have performed Correlation analysis and distinct feature selection techniques using a set of attributes to find out the Result of the Student's Performance. During the whole process, Naive Bayes Classifier displayed the highest accuracy of 82.4% to predict the unseen data. With this prediction, the academicians in the institutions can make use of this classification technique to take important measures in improving the Students Performance. Based on the different factors like Higher Secondary Marks (HSC) and medium of instruction, the teachers can help the students to bring Success in the respected course.

REFERENCES

[1] NainjaRikh, 2015, Data Mining and Knowledge Discovery in Database, International Journal of Engineering Trends and Technology Volume 23 Number 2

[2] FestimHalili and AvniRustemi , International Journal of Computer Science and Mobile Computing, Vol.5 Issue.8,August - 2016 , pg.207 - 215

[3] C.R Kothari, 2006, Research Methodology Techniques, New Age International (P) Limited.

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

[5] Sebastian Nowozin, 2012, Improved Information Gain Estimates for Decision Tree Induction

[6] Jiang, Zhe, and Shashi Shekhar. "Spatial Information Gain-Based Spatial Decision Tree." Spatial Big Data Science, 2017, pp. 57–76., doi: 10.1007/978-3-319-60195-3_4.