ISSN:2229-6107

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY

E-mail :
editor.ijpast@gmail.com
editor@ijpast.in

www.ijpast.in

INTERNATIONAL JOURNAL OF PURE AND APPLIED SCIENCE & TECHNOLOGY
IJPAST

ISSN 2229-6107 www.ijpast.in
Vol 11,Issuse 3.Sep 2021

Intelligent Data Mining Techniques for Predicting the Academic Performance of Students

Dr. M.M. Bokare, Dr S.B.Thorat, Mr. P.P. Joshi

ABSTRACT:
A big issue in today's society is the pervasive usage of alcohol among students. The effects of drunkenness on students' academic performance are clear. Here we describe a tiny collection of algorithms that were developed to enhance drunken learning. To identify the simplest formula among many alternatives, we employ a common Data Mining technique called "Prediction" in this study. Our research will analyse the pedagogical genius of school employees using the WEKA suite of technologies and R Studio. We conduct this study by utilising the student alcohol consumption datasets available on the kaggle website. There are a total of 395 tuples and 33 attributes. A classification model is constructed using Naive Bayes and ID3. The accuracy comparisons between R and WEKA have been finished. It is possible to predict whether a student will be promoted or demoted in the upcoming school year by looking at their performance in the previous year.

***Keywords**: Data Mining , Prediction ,Naïve Bayes ,ID3 ,WEKA ,R studio.*

1. INTRODUCTION

Researchers and scientists collect massive amounts of data from many different fields every day in order to sift through it for useful insights. The decision-making process makes use of a number of different data processing techniques. Diverse Mining Techniques for Classification, Prognostication, Clustering, and Connection. The goal of employing a classification method to organise information is to reveal commonalities. This technique is an example of supervised learning because it requires labelled training data to build the model. The core tenet of cluster analysis is that similar objects can be grouped together to reduce both their internal and external distances. Making accurate predictions requires analysing a large amount of data and making inferences about possible future events. Prediction is a data processing technique used to estimate a student's future performance. To compare the Naive Theorem and ID3 algorithms, we developed a model in WEKA and R and evaluated its performance metrics.

2. RELATED WORK

They used a lot of different data processing methods to look at student achievement in [1]. Students' performance must be evaluated using some sort of categorization system. Because it was the most efficient classification method, the call tree approach was implemented. The data they need is from the class of 2005 at Yarmouk University who took and passed a C++ course. Using CRISP-DM, a classification model is constructed. Twenty factors were found, but only twelve were shown to have any significant influence on students' performance in school.

Assistant Professor[1,3] , Director[2]
Department of computer science
SSBES ITM College Nanded
bokaremadhav@yahoo.com, suryakant_thorat@yahoo.com, pranavjoshi13@gmail.com

Using a Naive Bayesian approach, the authors of [2] are charged with creating a web-based application. Amrita Vishwa Vidyapeetham in Mysore has 700 students and data about them comprises 19 key characteristics. The simplest and most effective algorithm is the naive Bayesian. In this case, the input and output are the same thing: the student's performance over the course of a semester.

In [3], attention must be paid primarily on data mining techniques that enhance learning, particularly at the university level. This is an example of how knowledge mining can be used to help safeguard a school's standing. Students who are at danger of not completing a bachelor's degree in a satisfactory manner are identified early on, and interventions are provided to help them succeed. In order to gain some understanding, 214 freshmen from 2005-06 and 2006-07 were polled at NEDUET, Pakistan's Department of Applied Sciences. For a call tree to be effective, it must have mechanisms such as the Gini index, data gain, accuracy, Naive Bayesian, neural networks, and a random forest.

For [4], a survey was conducted among 300 engineering students over the course of three years. Some of the many disciplines to which these traits can be traced include English, mathematics, and artificial language. It's important to use Neural Networks, the J48 technique, and the SOM algorithm. Second and third year students at the Amrita School of Engineering in Bangalore are polled for their opinions in [5]. The dataset includes information on twenty different characteristics, including as gender, parental education, marital status, and more. Students whose predicted grades are low are offered guidance on how to improve their study habits using a naïve Bayesian classifier.

Use of alcohol consumption as part of a student's necessary knowledge set (ages 10-14) is mandated by [6].

After comparing the costs of using SVM, call trees, and Naive Bayes algorithms, they decided that SVM was the best option.

In [7], we see some examples of the application of data classification techniques in the medical field. The employment of IF-THEN rules for prediction is a common technique in data processing. The latest findings are presented at the broadest level in this discussion.

The authors of [8] used several data mining techniques on educational datasets to make predictions and assess their findings.

In [9], it is necessary to analyse the academic success of teachers and students in the Villupuram district. We used a bunching strategy based on the k-means bunch formula. The accuracy was enhanced by using a mathematical mixing model created by mathematicians.

In [10], the authors use K-means clustering to analyse student data collected from B. J. School. B.C.A. majors at B. J. College were surveyed to compile the data used in this study.

The author of [11] delves into the relevance of interconnections between a large data set.

The authors provide a method that teachers might use to assess their students' development as learners in the classroom.

Information from the medical field, such as the patient's vitals, diagnosis, and prescriptions, is crucial in [12, 13]. With this data, we're able to train the algorithm and zero in on a predictive pattern.

3. PROPOSED SYSTEM

Dataset:

The sample includes 34 different features and was collected via the internet. The dataset used in the challenge is provided by Kaggle. Name of school attended; gender; age; parents' level of education and occupation; number of hours spent studying per week; access to the internet; alcohol use; health; and academic success are among the most important factors to consider..

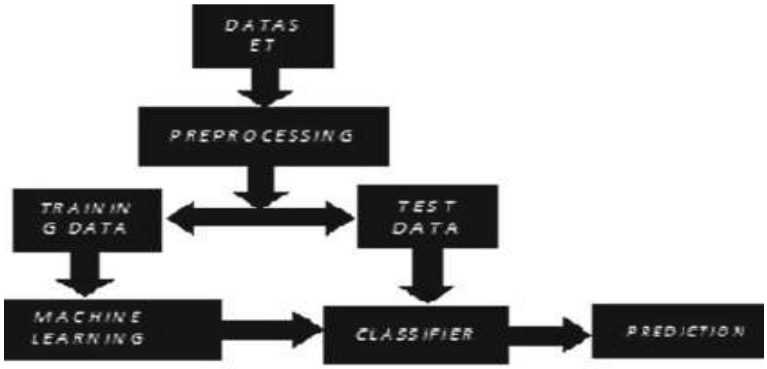

Figure 1: Proposed Model

Preprocessing: We need to preprocess the data before we can organise it. The preliminary processing consists of four stages. As a group, they are

When cleaning data, we remove irrelevant details and supplement them with mandatory information.

Hair Care Data Integration Bringing Together Various Types of Data.

In this phase of data transformation, disparate types of knowledge are brought together.

Data reduction is a technique for making information display more compact.

The first 300 tuples are separated out and used for training. The material utilised for coaching includes labels indicating levels of sophistication. Putting our knowledge of trains to use, we devise a system for categorising data.

Prediction and other machine learning techniques are commonly used to create classifiers. There are a plethora of data analysis algorithms that can be put to use in predicting. We can build a classifier using these techniques.

When it comes to classifiers, we've built one with the help of R and WEKA's mandated Naive Theorem and ID3 algorithmic rule**.**

Naïve Bayesian

Input: dataset

Output: confusion matrix and predicted class labels

Do

For each value of the class label (Ci,Cj) find probability

For each attribute belonging to the class label (either Ci or Cj) find probability

Compare probabilities of each attribute of different class labels

If p(Ci) > p(Cj)

Class label will be Ci else Cj

$P(Ci|X) = P(X|Ci)*P(Ci)/P(X)$

Confusion matrix: It is a tool for finding the accuracy.

ID3

Input: dataset

Output: confusion matrix and predicted class labels

Do

Calculate information gain of all the attribute

The attribute with highest information gain value will be taken as the root node

and according to the outcomes the tree will be further extended till all the leaf

node becomes the class labels.

Test Data

The remaining dataset is used as test data, and the classifier makes a prediction using this test data as input (Fig.2).

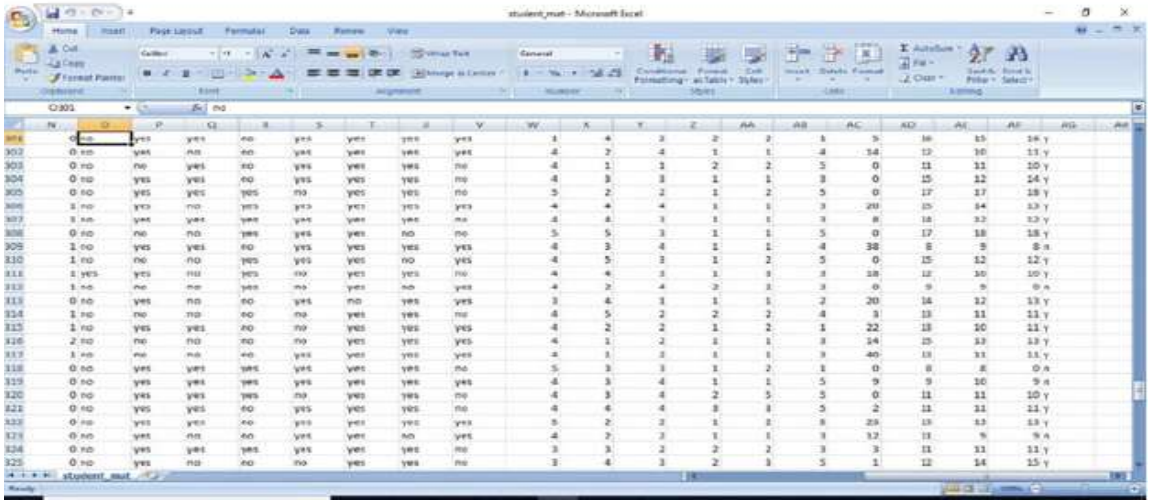

Figure 2 : Test Data

Data mining is a technique used to forecast the value of a tuple of information. During prediction, we provide the model with the check knowledge without the class label and have it estimate the sophistication level based on the trained knowledge consisting of sophistication labels. This label indicates the student's expected academic success or failure for the upcoming school year (yes = pass, no = fail).

4. RESULT ANALYSIS:

Accuracy for each attribute is compared between R and Weka Tool in Table 1.

Table 1: Comparision of Accuracy for ALL Attributes using R and WEKA

Algorithm	*R* (%)	WEKA (%)
Naïve Bayes	96.8	95.95
ID3	94.9	92

Table 2 compares R and WEKA in terms of accuracy for the factors that influence student success.

Table 2: Comparison of Accuracy FEW Attributes using R and WEKA

Algorithm	*R* (%)	WEKA (%)
Naive Bayes	94.7	87
ID3	100	100

Figure 3 depicts the effectiveness and quality of attributes. Above, you can see bar charts depicting the distribution of all attributes, with each colour denoting a different set of categories. Some of the variables examined were time frame, study length, gout severity, uric acid levels (G1, G2, G3, Dalc, Walc), and performance.

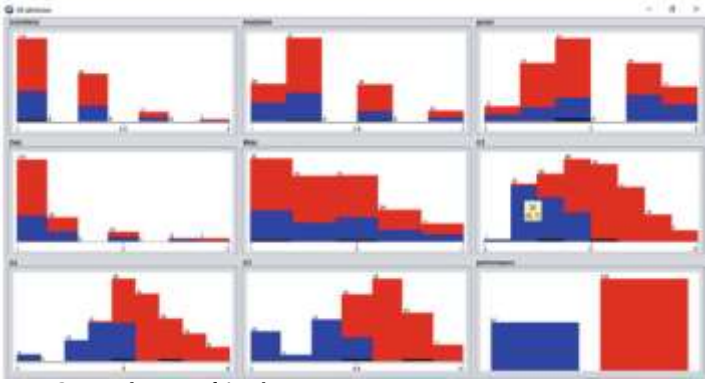

Figure 3: Visualization of Attributes

5. CONCLUSION:

We found the highest accuracy using Naive theorem and ID3 classifier compared to other algorithms. When we utilised the ID3 algorithmic programme, we found that the traits that were previously thought of as having the highest priority really had the lowest accuracy. Naive Bayes in R is more accurate than in rail if all the attributes have been considered. When all qualities for ID3 are included, the accuracy obtained in R and rail is equivalent to the accuracy gained in R and rail when only select attributes are considered. This suggests that once a sufficient number of attributes have been considered, implementing the classifier will take less time than it did for a smaller number. By looking at a student's performance over the course of multiple years, it is possible to predict whether or not they will pass the course.

6.REFERENCES:

1. Al-Radaideh, Q., Al-Shawakfa, E., Al-Najjar, M.I.: Mining student data using decision trees.The Int. Arab J. Inf. Technol.—IAJIT (2006)

2. Devasia, T., Vinushree T.P., Hegde, V.: Prediction of students performance using educational data mining. In: International Conference on Data Mining and Advanced Computing (SAPIENCE), pp. 91–95 (2016)

3. Asif, R., Hina, S., Haque, S.I.: Predicting student academic performance using data mining
methods. Int J Comput. Sci. Netw. Secur. (IJCSNS) 17(5), 187–191 (2017)

4. Ramesh, V., Parkavi, P., Yasodha, P.: Performance analysis of data mining techniques for
placement chance prediction. Int. J. Sci. Eng. Res. 2, 2229–5518 (2011)

5. Krishna, K.S., Sasikala T.: Prognostication of students performance and suggesting suitable learning style for under performing students. In: International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS—2018), December 2018.

6. Fabio, M.P., Roberto, M.O., Ubaldo, M.P., Jorge, D.M., Alexis, D.L.H.M., Harold, C.N.:
Designing A Method for Alcohol Consumption Prediction Based on Clustering and Support
Vector Machines. Res. J. Appl. Sci., Eng. Technol. 14, 146–154

7. Sreevidya B., Rajesh M., Sasikala T.: Performance analysis of various anonymization techniques for privacy preservation of sensitive data. In: Hemanth J., Fernando X., Lafata P., Baig Z. (Eds.) International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. ICICI 2018. Lecture Notes on Data Engineering and Communications Technologies, vol 26. Springer (2019)

8. Krishnaiah,V.,Narsimha, G., Subhash Chandra, N.:Diagnosis of lung cancer prediction system using data mining classification techniques. Int. J. Comput. Sci. Inf. Technol. 4, 39–45 (2013).

9. Shelke, N.: A survey of data mining approaches in performance analysis and evaluation. Int. J. Adv. Res. Comput. Sci. Softw. Eng. (2015)

10. Jyothi, J.K. Venkatalakshmi, K.: Intellectual performance analysis of students by using data mining techniques. Int. J. Innov. Res. Sci. Eng. Technol. 3, (2014)

11. Sreevidya, B.: An enhanced and productive technique for privacy preserving mining of
association rules from horizontal distributed database. Int. J. Appl. Eng. Res. (2015)
12. Bhise, R.: Importance of data mining in higher education system. IOSR J. Hum. Soc. Sci. 18–21 (2013).
13. Sumitha Thankachan, Suchithra, Data mining warehousing algorithms and its application in medical science. IJCSMC, 6 (2010).