International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 12, December-2019

Exploring Prediction Modeling of Students Alcohol and Drug Addiction Affecting Performance using Data Mining Approach

Tanvi Trivedi
Department of Computer Engineering,
V.V.P. Engineering College, Rajkot
Gujarat, India

Devangi Kotak
Department of Computer Engineering,
V.V.P. Engineering College, Rajkot
Gujarat, India

***Abstract*:- In our society, there are number of problems arises due to the students consuming alcohol and drugs during its teen age. Retrieving exact and accurate students which consumes alcohol and drugs is the main task. It is a real-world problem in our society. The major challenge for finding alcohol addicted students with respect to given data is to find accurate and efficient method which takes less time to generate results. There is large amount of data available, but getting the right information accessible when needed is very important. The availability of educational data has been growing rapidly, and there is a need to analyze hedge amount of data generated from this educational ecosystem. Educational data mining (EDM) has been emerged as a process of applying data mining tools and techniques to analyze the data at educational institutions. This area of research is gaining popularity due to potential benefits to the educational field. Educational institutions use educational data mining (EDM) to gain deep and through knowledge to enhance its assessment, evaluation, planning, and decision making in its educational programs. EDM helps academic programs to identify and discover hidden patterns in the data. These extracted patterns can be used for finding students who are consuming alcohol and drugs and its affect on their academic performance. In our proposed system we will use some educational institute's student's data and generate prediction weather student is alcohol addicted or not, we will do this by using clustering, classification, and filtering methods of data mining.**

Keywords:-Performance, Student performance, alcohol addiction, drugs, hybrid approach

I. INTRODUCTION:

Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets.Data mining is all about discovering unexpected/previously unknown relationships among the data. It is a multi-disciplinary skill that uses machine learning, statistic, and AI and data base technology. Data mining is also called as knowledge discovery, knowledge extraction, data/pattern analysis, information harvesting, etc.[1]

The practice of examining large pre-existing databases in order to generate new information is known as data mining. Prediction in data mining is to identify data points purely on the description of another related data value.It is not necessarily related to future events but the used variables are unknown. Prediction derives the relationship between a thing you know and a thing you need to predict for future reference.

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in.

The availability of educational data has been growing rapidly, and there is need to analyze hedge amount of data generated from this educational ecosystem. Educational data mining (EDM) has been emerged as a process of applying data mining tools and techniques to analyze the data at educational institutions. Recent educational data mining is evolving and helping the educational sector to adapt new teaching techniques for the learning process and the learner. This area of research is gaining popularity due potential benefits to the educational field. Educational institutions use educational data mining (EDM) to gain deep and through knowledge to enhance its assessment, evaluation, planning, and decision making in its educational programs.EDM helps academic programs identify and discover hidden patterns in the data. These extracted patterns can be used for gain data about which students are consumes alcohol and drugs and how it affect on their academic performance.

One of the frequently occurred problems in educational institutions is youth are drinking. Alcohol drinking has many short term and long term health effects. Taking alcohol at teenager age reduces child's mental and physical abilities as well as affect judgment and coordination which can lead to trouble.[2]

Different types of methods are used for predicting student's performance or predict which student consumes alcohol based on their performance. They all use either classification or clustering for prediction. In our proposed system we are using both classification and clustering hybrid approach for provides more accuracy.

II. LITERATURE REVIEW

This section summarizes the existing data mining techniques employed for prediction of student academic performance and prediction of student consuming alcohol and how it affect their academic performance. We studied different recent papers form IEEE, Springer, Google scholar, etc. that employed data mining for prediction of student performance or prediction of students alcohol consumption.

Few recent research works in the domain of interest are summarized in Table 1 and are detailed later in this section.

TABLE 1. SUMMARY OF RELATED WORK

Work	Research Question	Methodology
[3]	Prediction of alcohol consumption among Portuguese secondary school students	decision tree, KNN, Random forest and naïve bayes
[4]	Relevant factors and classification of student alcohol consumption	decision tree and random forest
[5]	Using data mining techniques for predicting alcohol consumption in Portuguese secondary schools	Five level classification method
[6]	Using data mining to predict secondary school student alcohol consumption	Business intelligence and data mining
[7]	Performance analysis of students consuming alcohol using data mining technique	SMO, Bagging, REP Tree, Decision tree
[8]	Is alcohol affect higher education student performance searching and prediction pattern using data mining algorithms	Decision tree algorithms
[9]	Educational data mining and learning analysis	Clustering
[10]	Predicting GPA and academic dismissal in LMS using educational data mining: a case mining	CRISP
[11]	Prediction of student performance using educational data mining	Naïve bayesian
[12]	Implementation of data mining to analyze drug case	Decision tree

In 2018, Shuhaida Ismail, Nik Intan Areena Nik Azlan and Aida Mustapha perform a comparative experiment on prediction of alcohol consumption by using four classification algorithms which include the decision tree, k-NN, random forest and naïve bayes. The result shows that the decision tree produced highest values for accuracy. [3]

In 2018, Auth pisutaporn, Burit Chonvirachkul and Daricha Sutivong carries out educational data mining to study the student alcohol consumption by using decision tree and random forest algorithm. Result shows that random forest algorithm perform better than the decision tree algorithm. [4]

In 2016, Mahsa Afsharizadeh and Hossein Ebrahimpour-Komleh predict amount of alcohol consumption using five level classification methods by using naïve bayes, decision tree and KNN.Result shows that decision tree classifier had the highest prediction accuracy. [5]

In 2016, Fabio Pagnotta and Mohmmad amran hossain intends to approach student addiction on alcohol in secondary level using business intelligence and data mining techniques. The result shows that a good predictive accuracy can be achieved, provided that addiction of alcohol can impact to the student performance. [6]

In 2017, Saurabh pal and Vikas chaurasia provide performance analysis of students consuming alcohol using data mining algorithms such as SMO, bagging, REP tree and decision tree. The result shows that bagging classification is better than other. [7]

In 2017, again Saurabh pal and Vikas chaurasia intends to predict Is alcohol affect higher education students performance using four decision tree algorithms (BFTree, J48, Rep Tree and simple cart).the result shows that BFTree algorithm mostly proper to classify and predict. [8]

In 2017, Akansha Mishra, Rhashi bansal and Dr. Shailendra Narayan predict educational data mining and learning analysis by using clustering. [9]

In 2012, Mahdi Nasiri, Fereydoon Vafaei and Behrouz Minaei describe an educational data mining case study based on the data collected from learning management system using CRISP methodology. Result shows that there can be confident models for predicting educational attributes. [10]

In 2016, Ms. Tismy Devasia, Ms. Vinushree T P and Mr. Vinayak Hegde proposed web based application for prediction of student's performance using educational data mining by using naïve Bayesian mining technique. Result shows that naïve Bayesian algorithm provides more accuracy over other methods. [11]

In 2018, Sri Wahyuni analyze drug cases using C4.5 Decision tree data mining technique. Result shows that Decision tree of C4.5 algorithm is effectively used in data processing. [12]

III. THE MAIN OBJECTIVES ARE:

1) Identify results from educational data mining by using data mining tools and methods for study the student alcohol and drug consumption.
2) Retrieve optimum and accurate data rather than using faulty data.
3) Finding best and accurate method
4) Finding time saving method and which are use for all data like large and small amount of data.
5) This prediction helps both institute and family to identify student's academic performance and their addiction.

IV. METHODOLOGY:

Methodology used for prediction of student consuming alcohol is dividing into following phases.

4.1) Phase 1: Understanding of Business:

The main aim of this study is to develop a model of prediction for student's alcohol consumption affect academic performance of students using data mining.

4.2) Phase 2: Data collection:

The dataset used for predicting alcohol consumption was originally collected and analyzed by [13] who performed prediction on secondary student school performance. To build the model of prediction the students features and their description which was gathered by [14] who performed academic performance of students was shown in the below table (Table 2)

The data has been gathered based on demographic features, academic features, behavior features and extra features.

4.3) Phase 3: Data Pre-processing:

After the collection of data set pre-processing methods are applied to develop the data set quality. This includes cleaning of data, feature selection, data transformation and data reduction. In transformation technique the data will be transformed using several transformation method which are z-transformation, range transformation, preposition transformation and interquatile transformation.

4.4) Phase 4: Applying Data Mining Classification Algorithms:

This study carried out the experiments using Decision tree and Naïve bayes classifier.

Decision trees are probably the most commonly used technique for data mining. It is a structure of flow chart where every internal node indicates an attribute test and every branch indicates a result of the test and class label is indicated by every internal node. Decision tree uses a decision tree as a predictive model which represents observations about an item to inference about the target value of item. [15]

Naïve bayes is among the entire easiest probabilistic classifier. It always performs well in real world applications despite the powerful assumption that entire features are independent conditionally. In the classifiers learning process with the known structure, conditional probabilities and class probabilities are estimated using training information and then these probabilities values are used to classify new observations. [16]

Table 2: Student features and their description

Category of features	Sub-features	Description
Demographic features	Gender Nationality Place of birth, etc.	Students gender Students nationality Students place of birth
Academic features	Stage ID Grade ID Section ID Semester	Educational stage of student. Grade level of student Classroom section Semester of student
Behavioral features	Raised hand Announcement view Discussion	Behavior of students during learning of education
Extra features	Parent answering survey Parent school satisfaction Student absence days	Survey ans by parents provided from school Parents satisfaction regarding school Absence days of students in school

4.5) Phase 5: Apply K-means clustering plus majority voting:

The data mining classification algorithms are applied to k-means clustering plus majority voting which is proposed in this study. K-means clustering is the most vastly used algorithm. This allots n points of data into k number of clusters so that same points of data can be grouped together. It is an iterative process which allots every point to cluster whose centroid is the closest. Then it again evaluates this groups centroid by taking is average. In this research a new algorithm is proposed by integrating k-means clustering plus majority voting which predicts the best accuracy.

4.6) Phase 6: Find the result:

After applying the k-means clustering plus majority voting the two classifier are compared and the best accuracy is found. The algorithm used for decision tree is ID3. According to Bedi (2015) ID3 algorithm uses information gain as criterion of splitting. Topmost decision node is the good predictor and it is also known as root node. The attribute with greater gain of information is chosen as split attribute. Information gain is employed to create tree from instances of training. This tree is employed to categorize test information. When an information gain method to zero or entire instances belongs to individual target then growth of tree terminal.

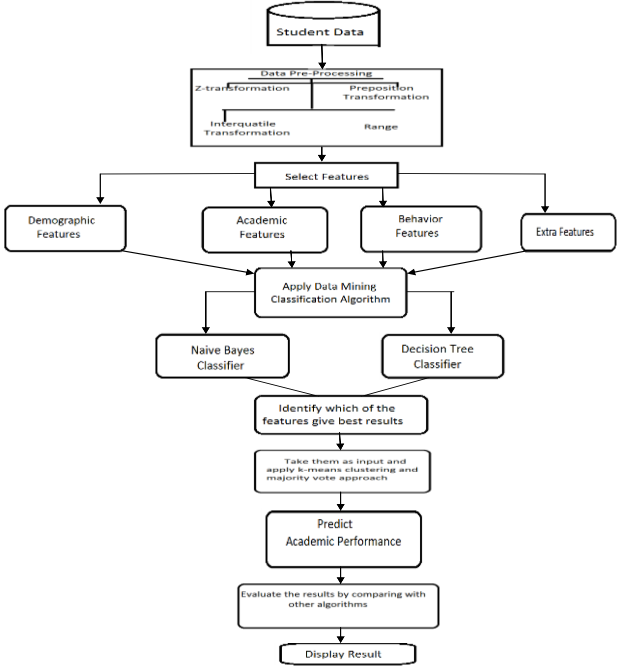
4.7) Phase 7: Evaluation of the algorithm proposed:

In this research four measures have been used for the evaluation of the quality of classification. The four measures are precision, recall, fscore and accuracy. Precision is the ratio of the properly classified cases to the total number of misclassified cases and properly classified cases. Recall is the proportion of correctly classified cases to total number of correctly classified cases and unclassified ones. F-score integrates the precision and recall measure which is regarded as a good indicator of relationship between them. Accuracy is the ratios of the total number of predictions were calculated properly.

4.8) Phase 8: Display the result:

The last phase is the result display which provides the details of the features and algorithm that gives best accuracy.

Fig. 1 Flow diagram of the proposed methodology:


Student Data
Data Pre-Processing
Z-transformation
Preposition Transformation
Interquatile Transformation
Range
Select Features
Demographic Features
Academic Features
Behavior Features
Extra Features
Apply Data Mining Classification Algorithm
Naive Bayes Classifier
Decision Tree Classifier
Identify which of the features give best results
Take them as input and apply k-means clustering and majority vote approach
Predict Academic Performance
Evaluate the results by comparing with other algorithms
Display Result

The equation of precision, recall, fscore and accuracy is stated below:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\text{Fscore} = 2\frac{\text{Precision c*Recall c}}{\text{Precision c + Recall c}}$$

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{True Positive + False Negative + False Positive + True Negative}}$$

V. RESULT AND DISSCUSSION:

Layer 1: Identify essential transformation technique based on the accuracy from algorithms.

The percentage rate for accuracy, recall and precision of the classification algorithms were evaluated for decision tree and naïve bayes algorithms. Results are as follows. First is the comparison results when using different transformation methods, which are z-transformation, range transformation, preposition transformation and interquartile transformation. Table II until V shows the difference in performance metrics using different normalization methods.

Table 3. Classification results using Z-transformation

Metrics	Decision Tree	Naïve bayes
Accuracy	**98.77%**	98.00%
Recall	9.57%	**57.39%**
Precision	**58.99%**	56.46%

Table 4. Classification results using Rang Transformation

Metrics	Decision Tree	Naïve bayes
Accuracy	**98.77%**	98.00%
Recall	**59.57%**	57.39%
Precision	**58.99%**	56.24%

Table 5. Classification results using Preposition Transformation

Metrics	Decision Tree	Naïve bayes
Accuracy	**98.77%**	39.30%
Recall	**59.57%**	11.31%
Precision	**58.99%**	11.43%

Table 6. Classification results using Interquatile Transformation

Metrics	Decision Tree	Naïve bayes
Accuracy	**98.77%**	9815%
Recall	**59.57%**	57.83%
Precision	**58.99%**	56.87%

Based on the presented results in Table II to V, DT algorithm produces a consistent percentage of accuracy and precision for all types of data normalization methods.

Layer 2: Clustering Students based on these features into three clusters:

The students are clustered based on these features into three clusters. The obtained features are then passed into K-means clustering algorithm to acquire clusters indicating high, medium and low performing students. New students are represented to these clusters and allotted labels based on majority voting of students in these clusters. Then the accuracy is computed on this test set.

Results of the features

The classification results by using all features are represented by a table.

Demographic features (Table 7)

Classifier	Precision	Recall	Fscore	Accuracy
Naïve Bayes	0.083333	0.194444	0.116667	0.333333
Decision Tree	0.267442	0.638889	0.377049	0.522013

Academic features (Table 8)

Classifier	Precision	Recall	Fscore	Accuracy
Naïve Bayes	0.098901	0.25	0.141732	0.314465
Decision Tree	0.119048	0.277778	0.166667	0.522013

Behavior features (Table 9)

Classifier	Precision	Recall	Fscore	Accuracy
Naïve Bayes	0.277108	0.638889	0.386555	0.540881
Decision Tree	0.242857	0.472222	0.320755	0.54717

Extra features (Table 10)

Classifier	Precision	Recall	Fscore	Accuracy
Naïve Bayes	0.351064	0.916667	0.507692	0.597484
Decision Tree	0.366667	0.916667	0.52381	0.622642

Academic + Behavior + Extra features (Table 11)

Classifier	Precision	Recall	Fscore	Accuracy
Naïve Bayes	0.270588	0.638889	0.380165	0.528302
Decision Tree	0.472727	0.861111	0.571429	**0.754717**

Features type vs. best accuracy (Table 12)

Features type	Best accuracy
Demographic features	0.528302
Academic features	0.415094
Behavior features	0.54717
Extra features	0.622642
Academic+Behavior+Extra features	**0.754717**

Result of new approach (Table 13)

Algorithm	Precision	Recall	Fscore	Accuracy
Clustering	0.641509	0.641509	0.641509	0.641509
Decision tree	0.622642	0.622642	0.622642	0.622642

From the previous analysis this research found that behavioral and extra features work the best for the accuracy of the system. After applying the new algorithm proposed in this study the result of the new approach is (Table 13).

VI. CONCLUSION:

This paper investigates two classification algorithms, which are Decision tree and Naïve bayes for predicting alcohol consumptions among students. Four normalization methods were also investigated, which are z-transformation, range, preposition and interquartile transformation. The classification results were then compared against each other. The result showed that the DT algorithm produced highest value for accuracy as compared to other classification algorithms. Other than that, the results of the application of the proposed hybrid algorithm show that there is a strong relation between behavior of student and their academic performance. The accuracy of the proposed hybrid model combining clustering and classification is 0.7547 when applied to features of the student data set and is found to be superior to that of the other existing algorithms. This model can help to reduce the failure rate of the academies. In future, more efficient prediction tools can be developed in order to pay more attention to the students and control alcohol influence in their life.

VIII. REFERENCES:

[1] https://www.guru99.com/data-mining-tutorial.html

[2] Auth pisutaporn,Burit chonvirachkul,Daricha sutivong; “Relevant factors and classification of student alcohol consumption”;IEEE International conference on innovative research and development(ICIRD);2018

[3] Shuhaida Ismail, Nik Intan Areena Nik Azlan, Aida Mistapha; “Prediction of alcohol consuming among Portuguese secondary school students”; IEEE Symposium on computer applications and industrial electronics(ISCAIE);2018

[4] Auth pisutaporn,Burit chonvirachkul,Daricha sutivong; “Relevant factors and classification of student alcohol consumption”;IEEE International conference on innovative research and development(ICIRD);2018

[5] Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh; “International Conference on new research achievement in electrical and computer engineering”; IEEE; 2018

[6] Fabio pognotta, Mohammad Amran hossain “Using data mining to predict secondary school student alcohol consumption”; Google scholar; 2018

[7] Saurabh pal and vikas chaurasia;“Performance analysis of student consuming alcohol using data mining technique”; International journal of advance research in science and engineering; 2017

[8] Saurabh pal and vikas chaurasia; “Is alcohol affect higher education students performance: searching and predicting pattern using data mining algorithms”; International journal of innovations and advancement in computer science (IJIACS); 2017

[9] Akansha mishra, rashi bansal and dr. shailendra narayan singh; “Educational data mining and learning analysis”; International conference on cloud computing, data science and engineering confluence; IEEE; 2017

[10] Mahdi nasiri, fereydoon vafaei and behrouz minaei; “Predicting GPA and academic dismissal in LMS using educational data mining: A case mining”; International conference of e-learning and e-teaching; IEEE; 2012

[11] Ms. Tismy devasia, Ms. Vinushree T P, Mr. Vinayak hedge; “Prediction of students performance using educational data mining”; International conference on data mining and advanced computing (SAPIENCE);IEEE; 2016

[12] Sri wahyuni; “Implementation of data mining to analyze drug cases using C4.5 Decision tree”; Journal of physics conference; 2018

[13] P. cortez and A. M. G Silva; “Using data mining to predict secondary school student performance”; 2008

[14] Bindhia k. francis and suvanam sasidhar babu; “Predicting academic performance of students using a hybrid data mining approach”;Journal of medical system; 2019

[15] Sharma H. and Kumar S. “A survey on decision tree algorithms of classification in data mining”; International journal of science and research (IJSR);2016

[16] Taheri. S. and Mammodov, M.; Learning the naïve bayes classifier with optimization models; International journal of application maths computer science; 2014