International Journal of Pure and Applied Mathematics
Volume 119 No. 12 2018, 221-227
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

PREDICTING STUDENT PERFORMANCE USING DATA MINING TECHNIQUES

Astha Soni[1], Vivek Kumar[2], Rajwant Kaur[3], D. Hemavathi[4]
Dept. of Information Technology, SRM Institute of Science & Technology
Kattankulathur - 603203, Kancheepuram Dt., Tamil Nadu, INDIA
[1]asthasoni912@gmail.com
[2]vivek.it.1108@gmail.com
[3]rajwantkaila43@gmail.com
[4]hemavathi.d@ktr.srmuniv.ac.in

***Abstract*-The paper represents the data mining techniques used for analysing pupil performance. Educational institutions contain an enormous amount of academic database containing student details. These student databases along with other attributes are taken into consideration like family background, family income, etc. It will help us by identifying promising students and by providing us a chance to pay heed and to refine those students who likely get low marks. For answer, we prepare a structure which will analyse the pupil's performance from their last performances using concepts of Data Mining under Classification. Classification Algorithms like Decision Tree, Naïve Bayes and Support Vector Machine can help us for predicting student's performance. This prediction helps parents and teachers to keep track of student's performance and provide required counselling. These Analysis also help in providing scholarship and other required training to the student. We are actually trying to enhance student's acquirement and success more effectively in a way using educational data mining techniques. It can bring the benefits & influence of novice, teachers and educational institutions. Experimental answers show that suggested procedure significantly outperforms prevailing procedure due to the misuse of family incomes and students' personal data component sets. Results of this examination can act as policy improvement technique in higher education.**

Keywords- Data Mining; Student Performance Prediction; Extra-Curricular Activities; Academic Performance; Students Personal Information

I. INTRODUCTION

Higher educational institutions considers Student's performance as one of its most crucial part. This is because of the fact that most of the educational institutions are based on the best record of academic performances. There are lots of discussions and reviews for Student's performance based on the previous researches. Schools, Colleges, and other Educational Institutes are running on high pace to provide scholars in this competitive world. These Educational institutes focus on generating graduates with good academic performances as well as extra-curricular activities. They need to keep track on how the student is performing in particular fields and in what fields they need more training. By using Educational data mining techniques, the educational authorities can have the idea before the starting of the new semester and can have informed decision so it will help them to effectively deal with all problems faced by the students while performing academically or in their personal life as will be already known to them. Large volumes of data are analyzed using educational data mining techniques so as to find different trends and patterns to predict the student performance. Day by day the volumes of data is increasing so to analyze we need to generate algorithms using data mining and then compare them so to get the maximum accuracy rate. More the accuracy rate the more specific the prediction is. There are multiple data classification techniques used for predicting the results each one having its own advantages and disadvantages. This paper comprises the use of decision tree, Naïve Bayes', Support Vector Machine techniques and algorithms based on them. These algorithms were used to study and recognize the space in prevailing examining techniques for analysing of scholar's performance.

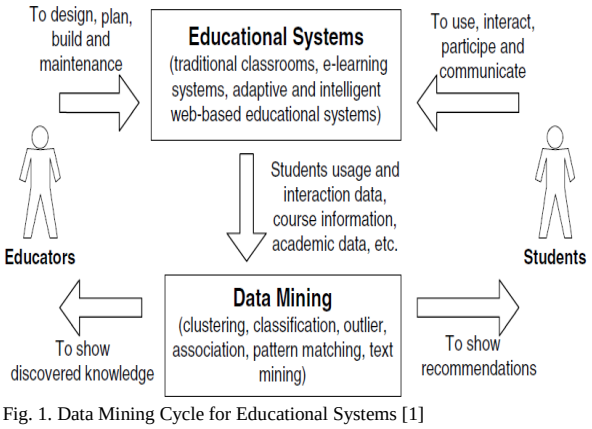

Fig. 1. Data Mining Cycle for Educational Systems [1]

II. RELATED WORK

Baradwaj and Pal [8] proposed a research on about 50 pupils who enrolled in the course for a period of 4 years with

attributes like “students last semester grades”, “weekly marks”, “lab assignments”, “attendance”, “final semester marks”, etc. They brought into use of the ID3 Decision tree for classification of student’s data and then build a decision tree for accurate prediction of student’s performance. Their main aim was to focus on the fields in which the student was not good and decrease the failure ratio. They brought into use of the ID3 Decision tree for analysing the student performance because it is the easiest learning algorithm.

Abeer and Elaraby [9] also did a relevant research to categorize and predict academic performance of a group of people over a time period of 6 years with multiple of features collected from educational institutes. As an output, they were able to predict the grade of students in the certain course and even can improve his performance by taking training in weak fields to avoid failure ratio.

Dorina Kabakchieva [7] did same research in which he used classification models generated by using four data mining algorithms – OneR Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbour. The maximum accuracy is achieved by using neural networks followed by Decision Tree model and K-NN model. The Neural network model works better with the “Strong” class while other three worked well with “Weak” class. The results of each one of the model are compared with others for the same attributes and data set.

Amjad Abu Saa [10] research concluded that student’s performance not only depends upon academics but also depend upon other personal, social and extra-curricular activities. He along with Naïve Bayes algorithm used three decision tree algorithms for classification of data. Firstly he did a survey and collected students data and then pre-processed and explored the data for data mining tasks. Secondly, the data mining algorithms were implemented on the data set to generate classification models for predicting student’s performance.

The research work carried out by Ali Daud and Farhat Abbas [11] introduces the student scholastic prediction process that uses four various types of attributes namely: family expenditure, family income, student personal information and family assets. It modifies the method of attribute subset selection in order to recognize the most important features for student scholastic performance prediction. It is obvious from the comparative analysis that their suggested attributes are effective predictors and achieved F1-score on real-life undergraduate student’s data. They finalized from the results that family expenditure and personal information attributes have a crucial effect on the performance of the student due to instinctive reasons.

III. LITERATURE REVIEW

A. Data Mining

Data mining is also known as the process of Discovery of Knowledge which refers to extracting or mining information from huge bunch of data. It helps in determining fascinating knowledge such as anomalies, changes, associations, patterns and important structures from huge volumes of information stored inside several different kinds of databases as in data warehouses or other information repositories available [12]. It's been popularly used nowadays due to the availability of very huge volumes of data in electronic form and there is a need for converting such data into useful information & knowledge for large applications. Decision Support, Artificial Intelligence, Machine Learning, Statistics and Database Systems and Business Management are some of the fields using its applications [2]. These methods are used to function on very huge amount of data for discovering hidden patterns and relationships helpful in decision making. While data mining and knowledge discovery are mostly treated as same, data mining is actually an integral part of the information discovery process. The step by step process required for extracting information from data are shown in Figure 2.

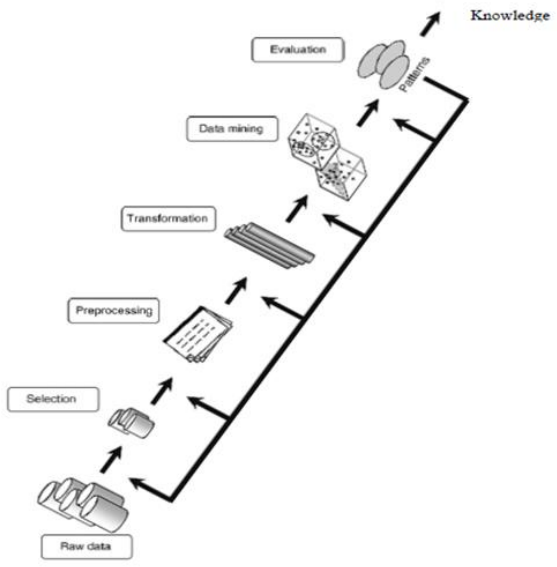

Fig. 2. The steps of extracting knowledge from data

B. Classification

Classification algorithm is a data mining technique that helps us to map data into predefined category. It is a supervised learning technique which needs categorized training data so it can creating rules for categorizing test data into pre-arranged category. [2] Its a 2 phase process. The first phase as the learning phase, where the classification rules are

generated and training data is analysed. The second phase as the classification phase, where test data is classified into predefined groups according to the generated rules. Since classification algorithms requires predefined classes based on values of information component, we had created an component "performance" for all students, for which they may have a value of either "Good" or "Bad".

C. Clustering

Clustering algorithm generally means grouping certain set of components in a way that the components in the same cluster are more similar to each other than to those in other categories [12]. Several fields like pattern recognition, image analysis, machine learning and information retrieval refer to this as a common technique for statistical data analysis. Clustering can be done by several methods that differ between the similar properties required between elements of a cluster and how to efficiently find the elements of the clusters.

D. Classification over Clustering

In process of implementation of clustering different classes can be discovered from the data and are examples of unknown apriori. As our main aim is to analyze students' performance into any of the predefined level - "Good" and "Bad", for which clustering was not appropriate, so we have used classification method instead of clustering method.

E. Prediction of Results

Regression method can be used for analysing as regression analysis can be used to model the relationship between one or more dependent and independent variables. Unfortunately, many real life problems are not simply a prediction. The same model can be used while predicting results for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks technique can be used to create both classification and regression models.

F. Accuracy Measurement

Examining which technique of data mining is good totally depend upon how the users has addressed the issues. Generally each methods performance is inspected by examining the accuracy of the results. Accuracy measurement in Classification technique is done by determining the percentage of placed tuples in the correct class. At the same time, there might be some cost associated with every incorrect assignment to the wrong class which can be neglected.

IV. METHODOLOGY

A. Models

We are using three kinds of classification models so as to learn the predictive function which is required. The models are used for experimental analysis. They are selected on the basis of their frequent usage in the existing literature. The list of methods are as follows:

1) Decision Tree

A decision tree is a tree in which each branch node will represent a choice between several alternatives and each leaf node will represent a decision. A decision tree is commonly used for obtaining information so as to fulfil the purpose of decision making. Decision tree starts from a root node which is there for users to take actions. From root node users split each and every node recursively into different nodes according to decision tree learning algorithm. The final result is a decision tree where each branch represents a possible context of the decision and its outcome.

2) Naive Bayes'

Naive Bayes algorithm is actually based on the probability theory, i.e. the Bayesian theorem [3] and is a simple classification method. It is named as naive because it solves problems based on two critical assumptions: it assumes that there are zero hidden components that will affect the process of analysing and it supposes that the prognostic components are conditionally independent with similar classification. This classifier provides an efficient algorithm for data classification and it represents the promising approach to the discovery of knowledge.

3) Support Vector Machine

Support Vector Machine is used for classification which is also a supervised learning method. There are three research papers that have used Support Vector Machine algorithm as their technique to analyse student's performance to review it thoroughly. Hamalainen et al. (2006) had chosen Support Vector Machine as their analysing method because it suited well in small datasets. [4] Sembiring et al. (2011) demonstrates that Support Vector Machine algorithm has a good ability of performing generalization and is actually found faster than other algorithms. [5] At the same time, the study done by Gray et al (2014) explained that Support Vector Machine algorithm acquires the highest analysing accuracy in identifying student's performance (Failing Risk). [6]

B. Data Preparations

For experimental purpose, the data of graduate and undergraduate students have been collected from different universities during the period (2017 to 2018) through a questionnaire survey. Once we got the details of all the students, we divided the training dataset, considering various feasible dividing components, i.e. the components which will have a major effect on the students' performance. Pre-processing is applied to obtain the most relevant characteristics of students. After removing inconsistencies and duplications in the dataset, we considered student instances for experiments. The student's performance model was created, where performance is measured with the performances in the areas such as "Academic", "Behaviour", "Extra- Curricular", and "Placement". We used 48 variables

as input to the model. The main purpose of our research is to find out and analyze the scholar's overall achievement.

C. Construction of Feature Space

The feature set is constructed by considering four categories of characteristics related to a student and his family. Initially, a pool of features is constructed by combining some existing (baseline) and proposed features and then feature subset selection process is applied to remove/reduce the number of redundant features. Overall, four categories of features are collected. Table 1 presents the description of each feature, its category, and possible values.

CATEGORY	NAME	DESCRIPTION	POSSIBLE VALUES
ACADEMIC	AttendanceA75	Attendance above 75%	A – YES; B – NO
	MarkA80	Marks above 80%	A – YES; B – NO
	MarkA40	Marks above 40%	A – YES; B – NO
	MarkA0	Marks above 0%	A – YES; B – NO
	InterestinStudy	Student interested in study	A – YES; B – NO
	Understanding	Student understanding of things	A – YES; B – NO
	MemorizeLession	Memorizes answers	A – YES; B – NO
	SentenceOwnWords	Write their own words	A – YES; B – NO
	Extracourses	Has taken extra courses	A – YES; B – NO
	BecomeScientist	Does research	A – YES; B – NO
BEHAVIOUR	Parents	Parents Alive	A – YES; B – NO
	ParentStudy	Literate Parents	A – YES; B – NO
	ParentReadWrite	Parents can read or write	A – YES; B – NO
	ParentEmployed	Parents are employed	A – YES; B – NO
	FamilyIncomeL30000	Nett income less than 30000 per month	A – YES; B – NO
	FamilyIncome30Kto60K	Nett income more than 30000 but less than 60000 per month	A – YES; B – NO
	ParentStrict	Parents are strict	A – YES; B – NO
	ParentCare	Parents are caring	A – YES; B – NO
	Emotional	Student is Emotional	A – YES; B – NO
	FailToleranceCapacity	Students tolerance for failures	A – YES; B – NO
	Anger	Student gets anger	A – YES; B – NO
	Patience	Student is patient	A – YES; B – NO
	RespectElder	Student respects elder	A – YES; B – NO
	FightwithFriends	Get in fight with friends	A – YES; B – NO
	BadHabit	Has bad habit	A – YES; B – NO
	BadHabitfromAdolescent	Has bad habit from adolescence age	A – YES; B – NO
	BadHabitfromFriend	Got bad habit from friend	A – YES; B – NO
	BadHabitfromFamily	Got bad habit from family	A – YES; B – NO
	BadHabitAddiction	Heavily addicted to bad habits	A – YES; B – NO
	BadHabitinEveryday	Everyday routine of bad habits	A – YES; B – NO
	BadHabitinweek	Sometime or weekly does bad habit	A – YES; B – NO
	StartedCuriosity	Started realising about bad habits	A – YES; B – NO
	WantRidofBadHabbit	Want to get rid of bad habits	A – YES; B – NO
	Pocketmoney	Gets pocket money	A – YES; B – NO
	SpendMoneyUseful	Spends money to buy needed items	A – YES; B – NO
	SpendMoenyBuyCigarAlcohol	Spends money to buy cigarettes and alcohol	A – YES; B – NO
	StealMoney	Steals money	A – YES; B – NO
	PoliceCompliant	Has police complaints	A – YES; B – NO
	RoamwithFriend	Roam around with friends all day	A – YES; B – NO
EXTRA-CURRICULAR	SocialService	Does social service	A – YES; B – NO
	SportsInterest	Interested in sports	A – YES; B – NO
	ExtraCurricular	Participates in extra curricular	A – YES; B – NO
	PoliticalInterest	Has interest in political science	A – YES; B – NO
PLACEMENT	PartTimeJob	Does part time job or internship	A – YES; B – NO
	Aptitude	Aptitude practice	A – YES; B – NO
	Coding	Coding practice	A – YES; B – NO
	GroupDiscussion	Group discussion practice	A – YES; B – NO
	PersonalInterview	Personal interview practice	A – YES; B – NO

Table 1: Features Distribution

V. EXPERIMENTAL RESULTS

In this part, we have discussed the results we have obtained by doing analysis on our student's dataset. Firstly we have prepared 45 questions and asked students from various department to fill the questionnaire. We prepared a dataset of about 2000 students with about 50 attributes. Then we used a decision tree, Naïve Bayes and Support Vector Machine techniques to do an analysis of 1000x50 matrix.

A. Individual Feature Analysis

This section evaluates the impact of each feature for the prediction of student's performance. Twenty features were selected by the feature extraction process out of 48 which are further selected for experiments. In experiments, three classifiers are used (NB, SVM and Decision Tree) to analyze the influence of each feature for predicting the performance of students. We find the "InterstedInStudy" is the best predictor of the desired student's performance using Decision Tree classification method. SVM and NB methods show second and third highest F1- scores using same features. Other Extra-Curricular features also play important roles. The "RespectElder" feature has the lowest performance for prediction and all classifiers present same F1-score (0.333). By analysing the performance of best and worst features, that concludes the proposed proposition based on the Academic Performance and Placement Performance feature seem to be helpful in improving the classification accuracy while performing analysis. "AttendanceAbove75" is found to be the second-best feature that also belongs to the proposed feature set of student's academic information and all classifiers show 0.77 F1-score which represents a better performance by using the proposed feature. The third best feature is "PoliceComplaint" which belongs to the student's personal behavior feature set. If we critically analyze the impacts of other proposed features in comparison with old features, better accuracy is obtained by using our proposed feature space as compared to the previously obtained features space. Hence, it can be concluded by saying that students' "InterstedInStudy", "AttendanceAbove75", "PoliceComplaint" and "ParentsReadWrite" characteristics are most influential for prediction of student's performance.

B. Comparisons

In this section, we have compared all the three algorithms of educational data mining which we have used for students analysis. Among all three algorithms, the Support Vector Machine algorithm has good accuracy as compared to other two. The accuracy using support vector machine is 83.33% which is far better than decision tree and Naïve Bayes. We have also done Clustering which is the method of arranging important data into subclasses or groups. This algorithm clusters the given data into defined clusters and we can obtain the mean of it using Euclid's distance method.

C. Discussions

In this part, we will discuss upon which feature contributed mostly to the desired outcome of students performance. The analysis showed us that the students' academic performance not totally depend upon students marks but also depend upon extracurricular as well as personal habits. We have distributed the data set into 5 other datasets to predict the result for scholarship, students' performance, students' behavior, aptitude skills and overall performance. From the analysis, we have observed that students' academic performance is contributed by family income, extracurricular activities, any bad habits, parent's strictness, parents' literacy, and student's emotional stability. Aptitude skills required for placement includes group discussion marks, mock interviews response as well as coding and quantitative and qualitative aptitude marks. The analysis predicted shows that parent's income and expenditure plays a major role in how their child's future will be. Indulging in bad habits like smoking or gambling also effects the students' academics as well as behavior. Teachers can keep a track of student's performance by analyzing the performance in various fields and share the same with parents. During placement, they can predict which student lack in which field and according to that they can provide training to students as well as scholarship.

VI. CONCLUSIONS

In this research, an effort is made to find the impact of our proposed features on student performance prediction with the help of classification models. A feature space is constructed by considering characteristics of family expenditure, family income, personal information and family assets of students. The potential/dominant features selection is unavoidable as it provides us with a subset of features. By using SVM classification algorithms we found our analysis very effective for our proposed features of family expenditure and student personal information categories. It can be easily derived from the results we got that academic information, family details and personal information have very strong impact on the students' performance due to instinctive reasons provided in discussions. The meta-analysis on analysing student's performance has encouraged us to carry out further examination to be applied in our educational institutes. Hence, Educational system can take the help of this model to review the student's performance in a suitable manner.

VII. FUTURE WORK

This experiment can be done with more components to get more accurate outputs which will be useful for improving the results of students learning process. Also the experiments can be done by using some other technologies for getting a broader approach and more accurate results. Some different tools can be used while at the same time different factors will be used. Most of the educational institutes mostly find it difficult to provide skilful employee to the society. Many universities/institutes are not in the state that they can provide proper learning environment because of lack of information and lack of proper guidance. To better administer and serve

student population, the universities/institutions need better assessment, analysis, and prediction tools. A considerable huge volume of examination is done in field of analysing student performance but all these are detached. So, it is easily understood that a combined approach is needed. Other than academic attributes, there are other components also which are responsible for students overall performance like personal and emotional stability. So proper data mining techniques are used to analysing the existing components and then classifying them in order to provide relevant results or outcomes. Hence if all factors and components are considered for the analysis, it can effectively increase the prediction model accuracy.

REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert systems with applications, vol. 33, no. 1, pp. 135-146, 2007.

[2] Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.

[3] Witten, I.H. & Frank E. (2000), Data Mining – Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann, San Francisco.

[4] W. H¨am¨al¨ainen, M. Vinni, Comparison of machine learning methods for intelligent tutoring systems, in: Intelligent Tutoring Systems, Springer, 2006, pp. 525–534.

[5] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, E. Wani, Prediction of student academic performance by an application of data mining techniques, in: International Conference on Management and Artificial Intelligence IPEDR, Vol. 6, 2011, pp. 110–114.

[6] G. Gray, C. McGuinness, P. Owende, An application of classification models to predict learner progression in tertiary education, in: Advance Computing Conference (IACC), 2014 IEEE International, IEEE, 2014, pp. 549–554.

[7] Kabakchieva, D., Stefanova, K., Kisimov, V. (2011). Analyzing University Data for Determining Student Profiles and Predicting Performance. Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 6-8 July 2011, Eindhoven, The Netherlands, pp.347-348.

[8] Baradwaj, B.K. and Pal, S., 2011. Mining Educational Data to Analyze Students' Performance. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[9] Ahmed, A.B.E.D. and Elaraby, I.S., 2014. Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology, 2(2), pp.43-47.

[10] Amjad Abu Saa, 2016. Educational Data Mining & Student's Performance Prediction. International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.

[11] Ali Daud, Farhat Abbas, 2017. Predicting Student Performance using Advanced Learning Analytics. International World Wide Web Conference Committee (IW3C2), published under Creative Commons, Pages 415-421.

[12] Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques, Elsevier.