Comparative study of supervised learning algorithms for student performance prediction

Mehdi Mohammadi [1)], Mursal Dawodi [2)], Wada Tomohisa [2)], Nadira Ahmadi [1)]

1) Kabul university, Faculty of Computer Science, Department of Information Technology Kabul, Afghanistan

2) Dept. of Information Engineering, University of the Ryukyus, Senbaru 1, Nishihara, Okinawa, Japan

***Abstract*— *With huge amount of data in diverse technological areas, and generating such kinds of data rapidly, it needs for proper usage; therefore, Data Mining has emerged. Data Mining can extract prominent knowledge from customary data that can attract attention of people to it which is meaningful information. Regarding this concept that data can be generated rapidly every day or even every moment, data need to take under process for offering better valuable information. Data of educational areas is more that belongs to students, and it's all right a good basis for commence of applying Data Mining. In this paper the focus is on how to use Data Mining techniques to discover information in student`s raw data and different algorithms such as KNN, Naïve Bayes, and Decision Tree are implemented.*

Keywords— *Data Mining(DM),Machine Learning(ML), K-Nearest Neighbor, Naïve Bayes, Decision Tree.*

Introduction

Every year, educational centers, schools, universities, and institutes admit many students in different fields with various grades and performance capability. Recently in educational areas for studying and teaching performance enhancement, DM and other techniques become well-known.

DM techniques and ML algorithms in educational area can simplify the prediction process. In this paper various data mining algorithms comparatively discussed on surveyed data from students at Kabul University. One of the most important problem that has taken from past up to now is vague state of student`s level performance which leads to produce slow and weak learning methods; this problem has arisen up subsequently of misunderstanding between high level and low level. Demonstration of performance level can be time-consuming; it means in short time it would be difficult to know which student is in high level or low level; therefore, DM techniques can solve this problem by predicting the final result according to training dataset. DM algorithms such as Decision Tree, KNN, and naïve Bayes are evaluated here.

LITERATURE REVIEW

DATA MINING: Knowledge Discovery from Data (KDD), it is the process of discovering interesting patterns from data [1]. It has been widely used in recent years due to availability of vast amount of data in electronic forms, and it is useful to convert them into useful information and knowledge that can be used for some applications which are related to different fields of study such as: Artificial Intelligence, Machine Learning, Market Analysis, Statistics, Database Systems, Business Management and Decision Support Systems [2].

CLASSIFICAION**:** It is a data mining technique for dividing data into predefined groups and if groups have label or name that is called supervised learning methods [3]. These kinds of method used to specify all data that have located to one of existence class, so it called supervised. If mining can be executed in educational environment that is educational data mining (EDM).

One other related work in EDM was done by Hijazi and Naqvi that have conducted a dataset of 300 students consisting of 225 males and 75 females from Punjab university of Pakistan, and make prediction of their performances [4]. They used Naïve Bayes classifier on their dataset. Their collected dataset had various features about students such as: hours of study, family income, percentage of recent semesters and so on. By collection of features about students and applying Naïve Bayes for its independent probability, they have selected some features with high probabilities and use them in building predictive model.

Factors	Features higher than 0.5 Probability		
	Variable	*Description*	*Probability*
1	GSS	Students grade in senior secondary education	.8642
2	LLoc	Living Location	.7862
3	Med	Medium of teaching	.7225
4	MQual	Mother`s Qualification	.6788
5	SOH	Students other habit	.6653
6	FAIn	Family annual income status	.5672
7	FStat	Students family status	.5225

Table 1: Features with higher probability than 0.5

Another related work in this area was done by Surjeet Kumar and Saurabh Pal. They applied multiple version of decision tree algorithms such as ID3, CART and C4.5 on student's dataset [5].

The following table shows their classification accuracy.

Table 2: Classifier accuracy

Factors	Classifier accuracy gained by Decision Tree		
	Algorithm	***Correctly classified instance***	***Incorrectly classified instance***
1	ID3	62.2222%	26.6667%
2	C4.5	67.7778%	32.2222%
3	CART	62.2222%	37.7778%

DATA MINING PROCESS

Data mining in education needs student's attributes, and different factors have impact on produced model; therefore, dataset need to be pre-processed in order to improve the model's accuracy. These processes pass through several steps as followings:

A. Data Preparation

In this research the first step is data preparation, so data gathering completed with surveying of 230 students at Faculty of Computer Science in Kabul University. The survey method was distribution of questionnaires among students, and then dataset was created in .csv format.
For data cleaning, preprocessing is essential; in our research we handled the missing values using averaging method,
The following table shows the list of attributes in the dataset.

Table 3: Dataset Attributes

Numbers	Student`s features in dataset		
	Variables	***Description***	***Values***
1	Sex	Gender	Male/Female
2	Cat	Category	IT/SE/IS
3	FLang	First language	Dari/Pashto/Other
4	Stat	Status	Married/single
5	Med	Media of teaching	Dari/Pashto/English
6	Loc	Location	Village/City-Home/City-Dormitory
7	Fsize	Family size	3,4,5,6,More
8	Fstat	Family state	Separated, Together
9	FAIn	Family annual income	Low, Medium, very good, perfect
10	SP	Secondary percentage	Less than 50, 50-60, 65-75, 75-85, more than 85
11	ToC	Type of College	Only girls only boys, coeducation
12	Fqual	Father`s qualification	PHD, Master, bachelor, 14, 12, none
13	Mqual	Mother`s qualification	PHD, Master, bachelor, 14, 12, none
14	Foccp	Father occupation	Governmental, emeritus, free, jobless
15	Moccp	Mother occupation	Governmental, emeritus, free, jobless
16	FQ	Friends quantity	1,2,3,more,none
17	JOB	Occupation	Yes, no
18	HoW	Hours of working	1,2,3,more than 3 hours
19	WSH	Weekly studying Hours	5,6,7, more than 7
20	TrP	Transportation	Walking, Bike, Private car, Ordinary cars
21	GPA	Grade Point Average	High, Medium, Low

B. Feature selection and transformation

In the feature selection, the subset selection method was used to select the most appropriate sub-set features. Here we evaluate the dataset using Weka.
Since we used scikit-learn to implement and compare the supervised algorithms, the selected categorical attributes were transformed to numeric. For this we used "One hot encoding" method and then data became ready to fit and produce the underlying model.

C. Implementation of mining

In this paper we have implemented KNN, naïve Bayes, and decision tree algorithms; each of which had different accuracy.

K-Nearest-Neighbor

In this algorithm; it picks a value for K which is the number of neighbors, search for the K observation in training set which are close to target and use the most popular response value from the K nearest neighbors as the predicted response. For parameter tuning we start from K=1 then KNN would search for one nearest observation until the best value for K is found it increases by one.
KNN implementation is consisted of the following steps:

- Load Data
- Initialize value for K
- Calculate distance between target observation and each records of training data(for instance using Euclidean Distance)
- Sort all calculated distances from lowest to highest (lowest mean nearest distance)
- Find out predicted classes

After applying KNN on dataset, 0.5464% accuracy was achieved, and the following figure shows how parameter tuning can affect the accuracy score.

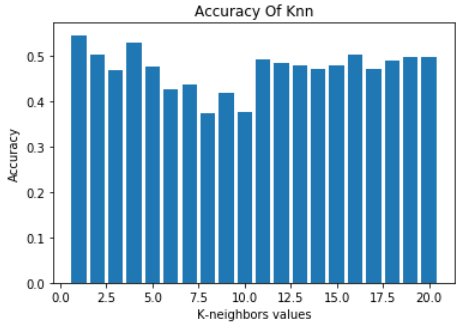

Figure1: Accuracy of KNN

Decision Tree

CART one version of DT is used for making model, CART stands for classification and regression trees introduced by Breiman [6], CART can work with categorical as well as continuous variables. Gini index has been used as attribute selection to construct trees. DT uses pruning to remove unreliable branches for increasing performance and accuracy of model.
ID3 algorithm uses **entropy** to calculate the homogeneity of samples. If samples are completely homogeneous, the entropy is zero. If they are equally divided, it has entropy of one.

Entropy can calculate as following:

$$\mathbf{Info(D)= \sum -p_i.log_2\ p_i}$$

Where p_i is the nonzero probability that an arbitrary tuple in D belongs to class Ci(c_1=high, c_2=medium, c_3=low) and is estimated by $|C_{i,D}|/|D|$ Info(D) is just the average amount of information needed to identify the class label of a tuple in D that is also known as the **entropy** of *D* [7].
Information gain is based on the depletion of entropy after a dataset is split by an attribute. Constructing a DT is all about finding attribute that returns the highest information gain (The most homogenous branch)

$$\mathbf{Gain(T,x) = E(T) – E\ (T,x)}$$

After applying DT on dataset we observed 0.5325% accuracy score, the following plot shows how diversity of Max features parameter can impact on all over result of models.

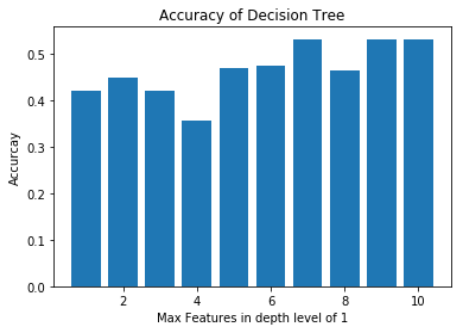

Figure 2: Accuracy of DT

Naïve Bayes

Bayes classification has been proposed with rule of conditional probability. Bayes rule is to estimate likelihood of a given data or input of Bayes rule.
The following formula is Naïve rule that is all about posterior and prior of one attribute, which posterior checks how one feature like x_i has probability that comes under class of any labels (h_i).

$$\mathbf{P(h_i|x_i) = P(x_i|h_i)\ P(h_i)\ /\ P(x_i|h_i)+P(x_i|h_2)\ P(h_2)}$$

This approach is Naïve, for it determines the independence between values of attributes. NB can be both predictive and descriptive algorithm; it means, the probability is descriptive and then use to predict target so called predictive algorithm [8].

With applying Naïve Bayes algorithm we obtain 0.4616% accuracy score.
Because we don't have parameter tuning in NB; therefore, we plot 10 cross folds values and get mean score of accuracy as below:

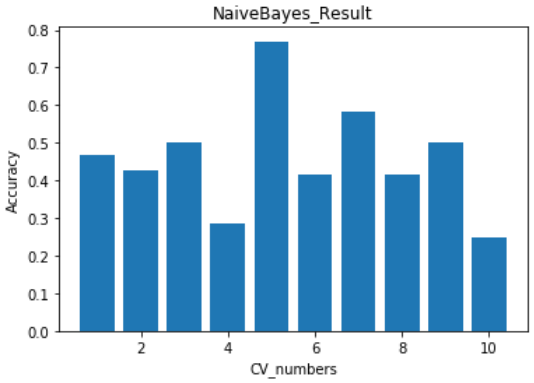

Figure 3: NB result

Experimental Result

Table 4 shows each implemented algorithms with their accuracies.

Table 4: Accuracy of Models

Algorithm	***Accuracy***
KNN	0.5464%
DT	0.5325%
NB	0.4616%

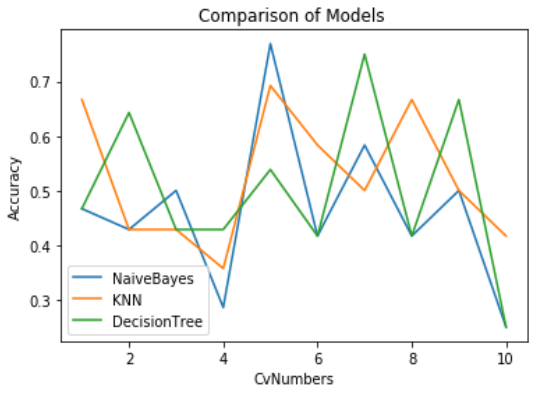

Figure 4: Accuracy Comparison

Figure 6 shows the comparison of each model in term of their accuracies that built by NaiveBayes, KNN and DecisionTree algorithms.

DISCUSSION AND OUTLOOK

As this investigation was done about student`s GPA prediction to increase their performance, and applying supervised learning algorithms, we figured out how to manipulate such kinds of data and class labels with DM techniques. In our work, we used subset feature selection for every algorithm in order to improve the accuracy of produced models. This helps to reduce the computation time and avoid overfitting for some models such as decision trees. It is turn to recommend next trick to enhance performance of models, for this reason; utilization of ensemble methods is recommended, and it is future plan for this research.

Obviously, this investigation goes through with applying some supervised learning algorithms such as: Decision Tree, Naïve Bayes, and KNN, all of these have their own influences and outputs.

In term of NB that it can check all features without consideration of dependencies, so it's usage is better if more than two class labels exist, like: text classification, spam filtering, recommender systems, and so on. If this is not what we want; the solution is to check other algorithms.

Decision tree, for its clarity became more acceptable for classification, and also it can work better with fewer classes. DT will encounter with challenges such as overfitting, but it can be handled with ensemble method like Random Forest that it will be considered in future plane of this investigation.

The next option is KNN that its implementation is based up on distance vector, it has no linear boundary. KNN also has worthy impact on accuracy of model.

For this investigation and data classification, KNN has the highest accuracy, for it uses classified distance matrix for classification and is more accurate; after that, DT has higher accuracy and NB has the lowest accuracy among them, because of its attribute dependency. Overall, NB is not much reliable as two other algorithms in practical problems.

CONCLUSION and Future Work

In this paper, KNN, DT and Naïve Bayes classifiers were used on the dataset of 230 students of Kabul University to predict their GPA as high, medium and low. This process can help instructors to decide easily about performance of students and schedule better methods for their education improvement. In the future, Ensemble Learning approach will be used as well as none grading relationship to produce a more robust model for students' performance prediction.

REFERENCES

[1] Jiawei Han,2012 vol.3 No.5 September international journal of data mining

[2] Kalpesh Adhatrao, Sep, 2013 Predicting students' performance using ID3 and C4.5 classification algorithms

[3] Duhnam, M.H, 2003 Data mining introductory and advanced topics, Pearson Education Inc

[4] Hijazi, S. T., and Naqvi, R.S.M.M., "Factors Affecting Student's Performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006

[5] Surjeet Kumar and Saurabh Pal,2012 B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140,

[6] J.R Quinlan, 1986 Introduction of decision tree", Journal of Machine learning", : pp. 81-106, 1986

[7] Jiawei Han,Micheline Kamber,Jian Pei,(2012).Data Mining Concepts and Techniques.

[8] Pandey, 2011 "A Data Mining View on Class Room Teaching Language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, March -2011, 277-282, ISSN:1694-08

[9] A. Arnold, R. Scheines, J. E. Beck, and B. Jerome, "Time and attention: Students, sessions, and tasks," in *Proc AAAI 2005 Workshop Educ. Data Mining*, Pittsburgh, PA, pp. 62–66.

[10] C.Romero and S.Ventura, "Educational Data Mining.A Survey from 1995 to 2005," Elesvier,Science Direct,Expert Sysytems with Applications, vol. 33, pp. 135-146, 2007