ISSN (e): 2250 – 3005 || Volume, 08 || Issue, 9|| Sepetember – 2018 ||
International Journal of Computational Engineering Research (IJCER)

Comparison Of Student Academic Performance On Different Educational Datasets Using Different Data Mining Techniques

Mrs. K. Deepika [1], Dr. N Sathyanarayana [2]
Department of Computer Science and Engineering [1] Tallapadmavathi College of Engineering
Department of Computer Science and Engineering [2]
Nagole Institute of Technology and Sciences, Ranga Reddy Dist.Telangana – India.
Corresponding Author: Mrs. K. Deepika

ABSTRACT: Educational data mining focus on developing different methods for solving educational problems which are hidden in an education field. The major problem which is faced in an education field is student dropouts or failure. There are many factors which are influencing the student dropouts. Many Data mining methods are used for identifying and predicting student's failures. In this paper comparison of different educational datasets like UCI, Kaggle is used to analyze the attributes which are causing an impact for student academic failures. How many data mining techniques are applied to these datasets and the results analysis among these two datasets are made. From the comparison, it is observed that parent responsibility attributes' has more impact on student academic performance. From the result analysis of both datasets, Decision Tree classifier performs high prediction on student performance.

KEYWORDS: Business Intelligence in Education, Educational Data Mining, E-learning, Student Performance Prediction, Classification, Behavioral Factors.

Date of Submission: 06-09-2018 Date of acceptance: 22-09-2018

I. INTRODUCTION

Educational data mining is used for developing methods and solving the problem in education data and used to discover the hidden patterns form different environments on education [1].EDM is used to find the patterns and to characterize the behavior and achievement of learners by making predictions. A student failure is a major social problem where educational professionals need to understand the causes, why many students fail in completing their education. It is a difficult task as there are many factors that cause for student failure. Therefore data mining task like classification was applied for predicting student dropouts'."One thousand factor problem" [26] is considered as student failure.

There are different sources through which Educational data can be collected are educational institute databases, e-learning systems, and traditional surveys. Therefore the hidden information can be extracted EDM using Decision Tree, Naïve Bayes and others [2, 3]. The knowledge that is discovered helps the decision makers of an educational institute to enhance their education system and for improving the education quality.

In this paper, comparison of two datasets is made. First, from UCI, the work was related to achievement of student in secondary education. The data is analyzed form two Portuguese secondary schools. The data consist of social features, student grades, school features and demographic features, collected by using some questionnaires and some reports. Mathematics and Portuguese are two core classes that are modeled by binary/five level classifier and by the regression. RandomForest (RF), Support vector machine (SVM) and Decision Trees (DT) are four DM techniques which are tested by three input feature selector which considered with previous grades or without previous grades [27].

Second dataset from Kaggle which is collected from e-learning system that called Kalboard 360 [4]. Here the experience API (XAPI) dataset is categorized as demographical features, academic background features, and behavioral features, to predict the performance of a student and concentrated on a new feature called behavioral feature to improve student performance. These features presented the learner and parent participation in learning process.

The data mining techniques applied to the student performance model are Artificial Neural Networks [5], Decision Tree [6] and Naïve Bayes [7] further ensemble methods like Boosting, Bagging, and Random Forest are also applied to improve these classifier performances. Then the nature of this feature was understood by expanding the data collected and by preprocessing steps.

This paper includes the following sections: Section 2 included with related works on datasets. Section 3

includes comparison on data collection and preprocessing is performed. Section 4 presents methodology applied on datasets. In Section 5 experiments and results are compared. Finally, the paper is concluded with advantages, disadvantages, comparisons and future work in Section 6.

II. RELATED WORKS

Many works are related to this work are as follows. The author Ma et al. [22] identified school students that belong to weak tertiary of Singapore and conducted some remedial classes using Association Rule of DM technique. They have considered demographic attributes as input attributes such as region, sex etc. and also considered the performance in school from previous years. Therefore solution proposed by them was outperforming traditionally. The author Minaei-Bidgoli et al. [23] worked on student grades on online for University of Michigan state. Three classification approaches have been modeled for these student grades like binary which includes a pass or fail, 3-level that considers low level, middle level, high level, and the 9-level includes from 1 – 9 that is from lowest grade to highest score [27]. The data was considered with 227 samples of online features like numbers of answers were corrected or trying for homework. The classifier ensemble methods like DT and NN showed the best results with an accuracy rate of a 94% with binary classes, 72% with 3-classes and 9-classes with 62%.

Kotsiantis et al. [21] worked on the University of Distance Learning Program for predicting computer science student's performance. For binary pass/fail classifiers many demographic attributes like sex, age, marital stages, and the attributes of performance like marks in a given assignment were considered as input variables and NB method showed the best result with 74% of accuracy.

Pardos et.al [24] worked on the online tutorial system at USA considering 8^{th} math test grade,to predict individual skills. Bayesian networks was used and obtained the best results with 15% of predictive error.

Many researchers worked on kaggel dataset for improving E-Learning systems by applying DM techniques. The author explored on some factors that show the impact on achievement of student using some DM techniques at Istanbul University[8].The features that effect the student achievement are extracted by path analysis.

The Students success is relating to the management of school and environment of school [9].The other teacher plays the major role in student success was proposed by authors in [10].The author in [1], worked on a case study using EDM to analyze the student learning methods.

The another author worked on categorizing the student performance into five groups using Expectation Maximization Algorithm [11].The classification method proposed by Shannaq et al in[12] shows Predicting the number of students that are enrolled.

K-mean clustering was applied by Ayesha et al in [13], where students learning activities are predicted. Number of researchers has applied many Data Mining tasks for solving the problems of educational institute. On UCI dataset, and kaggle dataset, many author applied various techniques of Data Mining to solve the problem of educational data [25, 27].

III. COMPARISION ON DATA COLLECTION AND PREPROCESSING

The data set collected from UCI [28] Table 1 consists of achievement of student in the secondary school of education which includes two Portuguese schools. The attributes considered in the dataset are student grade attribute, demographic features, social features and also features related to schools, which were collected by school reports and by some of the questionnaires. It is provided with two datasets in order to consider the performance within two subjects such as Portuguese language and mathematics.Cortez and Silva [18] worked on two different datasets by considering classifier like binary/five level and regression.

The educational data set of Kaggle [29] in Table 3 is collected from Learning Management System (LMS) called Kalboard 360[25]. Kalboard 360 is a multi-agent LMS, which was designed for facilitating the learning through the use of leading-edge technology. Data is collected through API (xAPI) which is a tool for tracking learner activities. The xAPI is the training and learning architecture (TLA) component that enables to monitor learning progress and learner's actions like reading an article or watching a training video [25]. The experience API helps the learning activity providers to determine the learner, activity and objects that describe a learning experience. The dataset includes 480 records of a student with sixteen features. These features were categorized into three groups such as (1) Features of Demographic which includes gender, nationality of student (2) Features of Academic background that includes stage of education, Level of grade and section of student (3) Behavioral features includes raising the hands in class, opening/visualizing the resources, answering survey by parents, and satisfaction of school.

The dataset includes 305 male students and 175 female students. Students coming from various regions are recorded such as students coming from Kuwait are179 students, 172 students recorded form Jordan, students recorded form Palestine are 28, 22 students are recorded from Iraq, Lebanon are recorded as 17, students coming from Tunis are 12, students coming from Saudi Arabia are 11, students coming from Egypt are 9,

students coming from Syria are 7, students recorded from USA, Iran, and Libya are 6, students coming from Morocco are 4 and student coming from Venezuela is 1.

The dataset was collected based on two semesters of education. For the first semester 245 records of students were collected, and for second semester 235 records of students were collected. Attendance feature is also included in dataset .this feature has been divided into two categories based on days of student absent. It was recorded that 191 students were absent more than 7 days, and 289 students are absent less than 7 days.

This dataset includes also a new category of features; this feature is participation of parent in the educational environment. Participation of parent feature has two sub-features one is Survey of parent answering and other is Satisfaction of school with parents. Therefore a number of parents answered the survey are 270 and not answered are 210. It was recorded the number of parents that are satisfied with the school are 292and 188 of parents are not satisfied.

After Data collection process some preprocessing techniques are applied on datasets in order to remove the noisy data, Then Feature selection process is considered as reducing number of attributes[23,24]. A filter-based approach can be applied using some selection algorithm like information gain, Gain ratio, Gini index, for evaluating the features ranks and checks which among the features are more important for building model of student performance. The information gain based selection is considered to evaluate which feature shows the impact on student performance [14, 15]. Student performance architecture [25] is shown in Fig 1.

IV. METHODOLOGY

The methodologies applied on UCI dataset [27] are classification and regression which are data mining goals. The difference between classification and regression is classification represents the discrete values where as regression represents continuous Values. Classification is evaluated using the percentage of correct classification (PCC), and regression using (RMSE) Root Mean Squared. A good classifier suggests high PCC i.e. near 100% where as regression should suggest low global errors i.e. close to zero.

The dataset is compared with the grades of mathematic and Portuguese. Therefore G3 of (Table 1) which is final grade is modeled based on three supervised methods [27].

1. **The Bin**ary classifier includes **pass** if G3 is greater than or equal to 10, otherwise, it includes **fail**;
2. A **5-L**evel classifier includes a Erasmus[1] system for conversation of grade, which is considered from Table 2.
3. A **Reg**ression considers the value of G3 which has numeric value between 0 and 20, which is considered as output.

The algorithms of the data mining are used for classifying and performing regression task on UCI datasets [27] are Decision Tree [17], Random Forest (RF) [16], Neural Networks and Support Vector Machines [20].

The methodologies applied on kaggle dataset [29] are classification methods. Classification is a technique which is applied on kaggle dataset to evaluate the features which have an impact on student performance. The classification technique which has been used are Naive Bayesian [7].classifier, Decision Tree [6], and Artificial Neural Networks [5].For further extension ensemble methods are applied in order to improve these classifier performances.

The common methodologies applied on these two data sets are DT, which has shown the good result for predicting student performance and the model is easily understood by the human.

V. EXPERIMENTS AND RESULTS

RMiner was conducted on UCI dataset [27] through which the data mining techniques can be facilitated. R environment is an open source library with a set of coherent functionalities for classification and regression task. Therefore rpart (DT), nnet (NN), random forest (RF) and kern lab (SVM) packages are used in this library.

The kaggle dataset [28, 29] was used in order to evaluate the classification methods and there comparisons. They applied cross-validation with 10 folds in order to divide data set as training and testing partitions.

Evaluation on UCI Dataset:

Before applying the models on UCI dataset some preprocessing was applied on NN and SVM methods. The nominal attributes (eg.Mjob) have been transformed and encoded as 1-of-C and therefore zero mean and standard deviation [20] are standardized to all attributes. After applying DM model. The DT reduced the sum of squares, and some of the parameters that are default were considered for the RF. Therefore the example considers the value T=500 for the NN, The value of E=100 epoch regarding BFCS Algorithm and SVM with eg. Sequential Minimal Optimization Algorithm is considered.

Therefore G_1 and G_2 are considered as having a great impact for each DM model. The three input configurations were tested as follows.

- **A** indicates that all variable are considered form Table 1 and except G3 is considered as output
- **B** indicates that it is same as **A,** without considering G2 which includes the second-period grade.
- **C** indicates that it is same as **B,** without considering G1 which includes the first-period grade.

To access predictive performance, ten cross-validations were applied for each configuration from 20 runs [20].The data is divided randomly into ten equal subsets.10% of data is tested in one subset and Data Mining techniques were applied on remaining data. The test set which is evaluated contains whole data set and predictions are made based on 10 variations of same DM model.

Fig. 1: Architectural diagram of the student performance

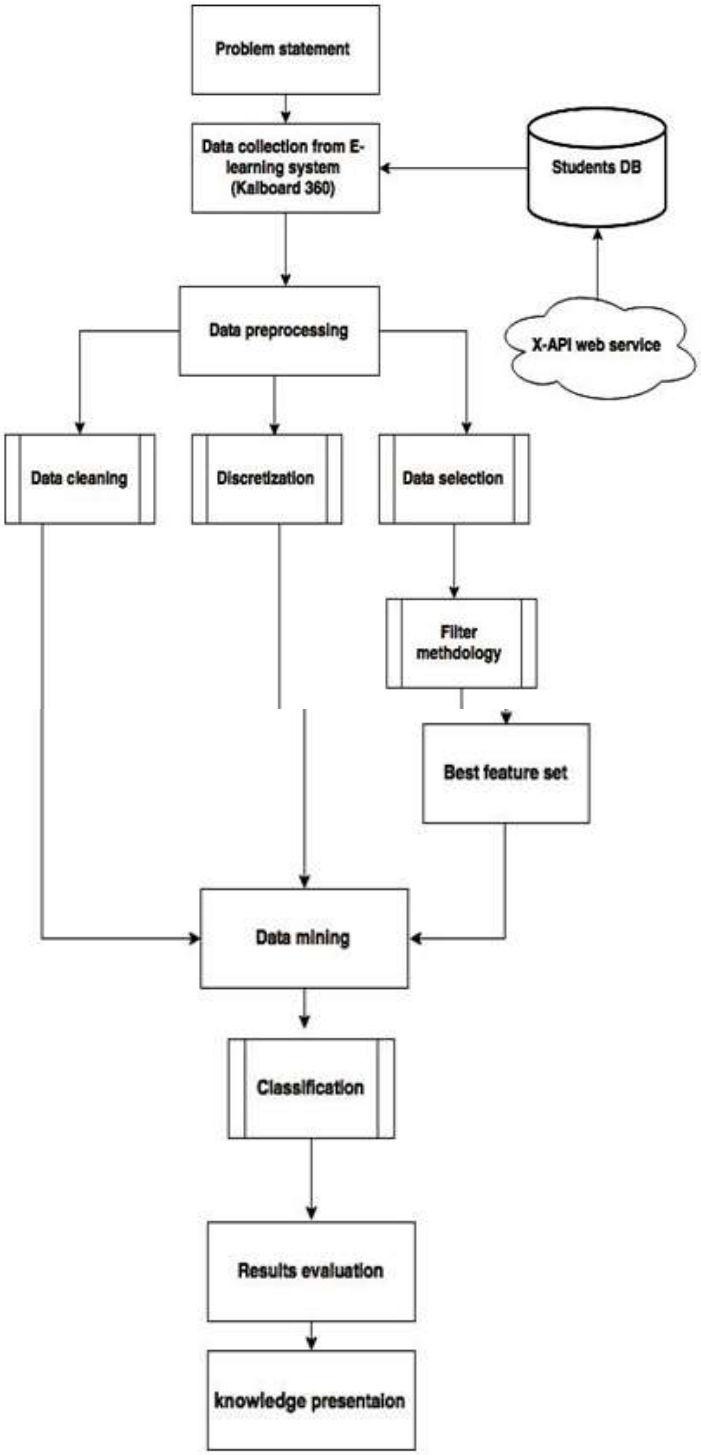


Table 1: The UCI student Dataset variables

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 - < 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour or 4 - > 1 hour).
studytime	weekly study time (numeric: 1 - < 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours or 4 - > 10 hours)
failures	number of past class failures (numeric: n if 1 ≤ n < 3, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

For comparison, Naïve prediction is also tested. For A setup, this model is considered as same as second-period grade G2 or versions of binary/five-level. First-period grade are used when the second grade is not available (i.e., B setup).When the evaluation is not present(C setup) then classification task or regression was returned.

The tested result is shown in Table 4 to 6 [27] with mean and 95% t-student confidence intervals by Flexer [19]. A setup achieves the best result and when the grade of second-
period is not considered (B), and then predictive performance decreases. Therefore results are considered worst when the scores of students are not used (C).

For last evaluation, the naïve predictor is considered first two setups as input which gives the best classification goals for mathematics with binary and 5level, and also regression of Portuguese was considered under input selection A [27]. The inputs with non-evaluation are not used in these cases.

Table 2.The five-level classification system

Country	I excellent/very good	II good	III satisfactory	IV sufficient	V (fail)
Portugal/France	16-20	14-15	12-13	10-11	0-9
Ireland	A	B	C	D	F

Random Forest is considered as the best choice among 8 cases then Decision Trees are considered as best in 4 cases. The nonlinear functions like NN and SVM outperformed due to number of irrelevant inputs. The examples with decision tree are shown in Fig 2.

In binary and 5-level classification good outfits are relieved by considering values that are in majority near the diagonal of matrix. The table 7 shows the importance of relative in percentage is presented for each variable that are considered as input and measured using RF Algorithm [16, 27].

TABLE 3. STUDENT FEATURES AND THEIR DESCRIPTION FROM KAGGLE DATASET

Feature Category	Feature	Description
Demographical Features	Nationality	Student nationality
	Gender	The gender of the student (female or male)
	Place Of Birth	Place of birth for the student (Jordan, Kuwait, Lebanon, Saudi Arabia, Iran, USA)
	Relation	Student's contact parent such as (father or mum)
	Stage ID	Stage student belongs such as (Lower level , Middle level , and high level)
	Grade ID	Grade of students (G-1, G-2, G-3, G-4, G-5, G-6, G-7, G-8, G-9, G-10, G-11, G-12)
	Section ID	Section student belongs such as (A, B, C).

Academic Background Features	Semester	School year semester such as (First or second).
	Topic	Course topic such as (Math, English, IT, Arabic, Science, Quran)
	Teacher ID	Teacher who teach this Particular course.
Parents Participation learning	Parent Answering Survey	parent Answering the surveys that provided from school or not
	Parent school satisfaction	This feature obtains the degree of parent satisfaction from school as follow(Good ,Bad)
Behavioral Features	Raised hand on	interaction with Kalboard 360 e-learning system.
	Visited resources	
	discussion groups	
	Viewing announcements	

Table 4. Results of Binary classification (PCC values are in %, best model represented by underline, bold represents best setup input)

Input	Mathematics					Portuguese				
Setup	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	**91.9**$^{\dagger}_{\pm 0.0}$	88.3$_{\pm 0.7}$	86.3$_{\pm 0.6}$	90.7$_{\pm 0.3}$	91.2$_{\pm 0.2}$	89.7$_{\pm 0.8}$	90.7$_{\pm 0.5}$	91.4$_{\pm 0.2}$	**93.0**$^{\dagger}_{\pm 0.3}$	92.6$_{\pm 0.1}$
B	**83.8**$^{\dagger}_{\pm 0.0}$	81.3$_{\pm 0.5}$	80.5$_{\pm 0.5}$	83.1$_{\pm 0.5}$	83.0$_{\pm 0.4}$	87.5$_{\pm 0.8}$	87.6$_{\pm 0.4}$	88.0$_{\pm 0.3}$	88.4$_{\pm 0.3}$	**90.1**$^{\dagger}_{\pm 0.2}$
C	67.1$_{\pm 0.0}$	66.3$_{\pm 1.0}$	**70.6**$^{*}_{\pm 0.4}$	65.3$_{\pm 0.8}$	70.5$_{\pm 0.5}$	84.6$_{\pm 0.6}$	83.4$_{\pm 0.5}$	84.8$_{\pm 0.3}$	84.4$_{\pm 0.4}$	**85.0**$^{*}_{\pm 0.2}$

† –pair-wise comparisons of statistical significance with other methods
. ∗ – pair-wise comparison of statistical significance with NV.

Table 5. Results of Five-level classifier (values of PCC are in %, best model represented by underline, bold represents best setup input)

Input	Mathematics					Portuguese				
Setup	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	**78.5**$^{\dagger}_{\pm 0.0}$	60.3$_{\pm 1.6}$	59.6$_{\pm 0.9}$	76.7$_{\pm 0.4}$	72.4$_{\pm 0.4}$	72.9$_{\pm 0.0}$	65.1$_{\pm 0.9}$	64.5$_{\pm 0.6}$	**76.1**$^{\dagger}_{\pm 0.1}$	73.5$_{\pm 0.2}$
B	**60.5**$^{\dagger}_{\pm 0.0}$	49.8$_{\pm 1.2}$	47.9$_{\pm 0.7}$	57.5$_{\pm 0.8}$	52.7$_{\pm 0.6}$	58.7$_{\pm 0.0}$	52.0$_{\pm 0.6}$	51.7$_{\pm 0.4}$	**62.9**$_{\pm 0.2}$	55.3$_{\pm 0.4}$
C	32.9$_{\pm 0.0}$	30.4$_{\pm 1.0}$	31.0$_{\pm 0.7}$	31.5$_{\pm 0.6}$	**33.5**$_{\pm 0.6}$	31.0$_{\pm 0.0}$	33.7$_{\pm 0.6}$	34.9$_{\pm 0.5}$	32.8$_{\pm 0.6}$	**36.7**$^{\dagger}_{\pm 0.6}$

† – pair-wise comparisons of statistical significance with some more methods.

Table 6. Results of Regression (values of RMSE, best model represented by underline, bold represents best setup input)

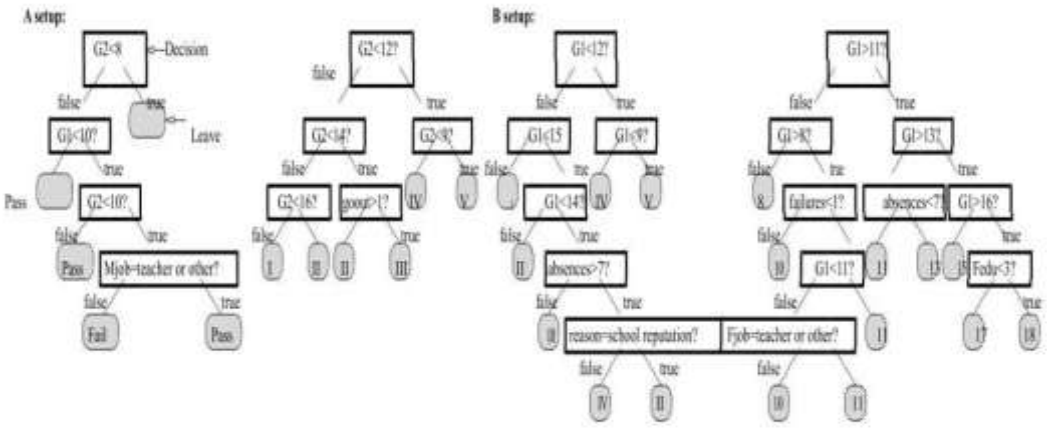
Input	Mathematics					Portuguese				
Setup	NV	NN	SVM	DT	RF	NV	NN	SVM	DT	RF
A	**78.5**†	60.3	59.6	76.7	72.4	72.9	65.1	64.5	**76.1**	73.5
B	**60.5**†	49.8	47.9	57.5	52.7	58.7	52.0	51.7	**62.9**	55.3
C	32.9	30.4	31.0	31.5	**33.5**	31.0	33.7	34.9	32.8	**36.7**†

† – pair-wise comparisons of statistical significance with other methods. ∗ –pair-wise comparison of statistical significance with NV.

Tabl e 7. Shows the importance of relative input variables with RF models

Setup	Relative Importance
C-Mat-Bin	failures: 21.8%, absenses: 9.4%, schoolsup: 7.0%, goout: 6.5%, higher: 6.4%
B-Por-Bin	G1: 22.8%, failures: 14.4%, higher: 11.9%, school: 8.1%, Mjob: 4.1%
C-Por-Bin	failures: 16.8%, school: 13.2%, higher: 13.1%, traveltime: 5.9, famrel: 5.7%
C-Mat-5L	failures: 18.3%, schoolsup: 9.5%, sex: 5.7%, absences: 5.6%, Medu: 4.5%
C-Por-5L	failures: 16.8%, higher: 9.9%, school: 9.3%, schoolsup: 6.9%, Walc: 6.6%
A-Mat-Reg	G2: 30.5%, absenses: 20.6%, G1: 15.4%, failures: 6.7%, age: 4.2%
B-Mat-Reg	G1: 42.2%, absenses: 18.6%, failures: 8.9%, age: 3.3%, schoolsup: 3.2%
C-Mat-Reg	failures: 19.7%, absenses: 18.9%, schoolsup: 8.3%, higher: 5.4%, Mjob:4.2%
C-Por-Reg	failures: 20.7%, higher: 11.4%, schoolsup: 6.9%, school: 6.7%, Medu:5.6%

Figure 2. Examples of Decision Trees



Evaluation on kaggle dataset:

For evaluation on kaggle dataset 4 measures were considered which shows classification confusion matrix in Table 8, based on four equations [29].

Table 8. Confusion Matrix

		Detected	
		Positive	Negative
Actual	Positive	True positive (TP)	False Negative(FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$Fc = 2\frac{Precisionc * Recallc}{Precisionc + Recallc} \quad (4)$$

Results of kaggle dataset

In Evaluation, the results using traditional data mining techniques and the impact of behavior features is evaluated for student academics performance using different classification techniques such as DT, ANN, NB. Each classifier was introduced with two classification results [25]

i. With student behavior features (BF)
ii. Without behavioral features (WBF)

The results are shown in Table 9, where ANN model outperforms with other data mining techniques [29]. From Table 9, ANN model achieves the accuracy of 73.9 with BF and 57.0 without BF that means 380 out of 480 students are classified correctly into right class labels like high, medium and low whereas 100 students are classified incorrectly. The recall measure obtains 79.2 with BF and 57.1 without BF that means 380 students are classified correctly with number of unclassified cases and correctly classified cases. The precision measure obtains 79.1 with BF and 57.2 without BF that means 380 out of 480 students are classified correctly and 100 are misclassified than the F-measure obtains 79.1 with BF and 57.1 without BF.

Therefore the experimental results obtained show the strong effect of student academic performance and learner's behavior [25, 28, 29].

Table 9. Classification Method Results with Behavioral Features (Bf) and Results without Behavioural Features (WBF)

Evaluation Measure	DT (J48)		ANN		NB	
Behavioral features existence	BF	WBF	BF	WBF	BF	WBF
Accuracy	75.8	55.6	79.1	57.0	67.7	46.4
Recall	75.8	55.6	79.2	57.1	67.7	46.5
Precision	76.0	56.0	79.1	57.2	67.5	46.8
F-Measure	75.9	55.7	79.1	57.1	67.1	46.4

Table 10. Classification Method Results Using Ensemble Methods

Evaluation Measure	Traditional classification methods			Bagging			Boosting			Random Forest
Classifiers type	DT	ANN	NB	DT	ANN	NB	DT	ANN	NB	DT
Accuracy	75.8	79.1	67.7	75.6	78.9	67.2	**77.7**	**79.1**	**72.2**	75.6
Recall	75.8	79.2	67.7	75.6	79.0	67.3	**77.7**	**79.2**	**72.3**	75.6
Precision	76.0	79.1	67.5	75.7	78.9	67.1	**77.8**	**79.1**	**72.4**	75.6
F-Measure	75.9	79.1	67.1	75.6	78.9	66.7	**77.7**	**79.1**	**71.8**	75.5

Table 11. Classification Methods Results through Testing and Validation

Evaluation Measure	**Testing results**			**Validation results**		
Classifiers	DT	ANN	NB	DT	ANN	NB
Accuracy	75.8	79.1	67.7	82.2	80.0	80.0
Recall	75.8	79.2	67.7	80.0	80.0	80.0
Precision	76.0	79.1	67.5	84.7	84.7	83.8
F-Measure	75.9	79.1	67.1	79.2	79.2	80.2

In Table 10, ensemble methods [29] are applied for improving the evaluation results of traditional data mining models. The boosting method is outperformed than other ensemble methods in which the accuracy measure of decision tree improves from 75.8 to 77.8 using boosting method and also the recall measures are increased from 75.8 to 77.8 and precision measures in DT using boosting increased from 76.7 to 77.8 and F-measure increases from 75.9 to 77.7 and also boosting method with NB model is also improved with all 4 measures that is observed in Table 10.

Validation process starts once the classification model being trained with ten folds cross-validation. The evaluation results with many classification methods like ANN, NB and DT is shown in Table 10 with testing and validation process therefore 500 students are trained using the models and the model is validated with 25 new comer students. To evaluate the reliability of trained model in validation process unknown labels are considered in the data sets.

Therefore from the Table 11, it is analyzed that evaluation results are increased in validation process for three prediction models. These three models achieve 80% of accuracy that means 20 out of 25 new students are correctly classified into the high, medium and low class labels, and 5 students are classified incorrectly[29].

VI. CONCLUSION

Student academic performance is achieved with various factors of students. The factors considered on UCI focus on demographic attributes and school performance over past years. Mathematics and Portuguese are two classes that are modeled by binary or 5-level classification methods and regression methods. The four DM models like DT, RT, NN and SVM are considered and tested with three selections of input by considering previous grades and not previous grades. High predictive accuracy is obtained by providing first or second school period grades as shown in results.

The factors like demographical features, academic background features, and behavioral features are considered on kaggle dataset where new students performance is predicted by applying data mining techniques like ANN, NB and DT over behavioral features of students. These classifier results are increased by considering ensemble methods.

The advantage of UCI is the results obtained reveals that high predictive accuracy is possible by providing the first or second school period grades.

Used RMiner environment which process on fewer attributes.UCI considered less attributes therefore more attributes are needed for predicting student performance.

In kaggle the reliability of proposed models are increased.

Both the data sets showed that high student performance prediction is obtained more by Decision Trees (DT) which are more reliable and the representation of DT is easily understood. Therefore the comparison is made on two datasets for predicting student performance.

From the analysis of two datasets, the common attributes that influencing student performance is parent responsible for student from kaggle and Pstaus,Medu,Mjob,Fedu,Fjob attributes from UCI are considered which belongs to demographical feature shows the impact on student performances. For future work more student

characteristics need to be analyzed to predict student academic performance accurately.

REFERENCES

[1]. "A case study for Mining Data of students to analyze the Learning Behavior of student "El-Halees A. Arab international Conference of Information Technology. (ACIT2008), in 2008 Dec 15- 18.

[2]. M Mohammed, Abu Tair, M Alaa,and El-Halees, "A Case study of Mining Educational Data to improve student's performance". International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012.

[3]. Romero C and Ventura S. "A review of the state of the art. Systems, Man, and Cybernetics, Part C,Applications and Reviews :Eductional data mining",IEEE Transactions on, 40(6), 601-618.

[4]. Kalboard 360-E-learning system, http://cloud.kalboard360.com/User/Login#home/index/ (accessed July 31, 2015).

[5]. Naser S.A, Zaqout I, Ghosh M.A, Atallah R, and E Alajrami. "Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology". International Journal of Hybrid Information Technology, 2015, 8(2), 221-228.

[6]. "DataMiningClassification",byKevinSwingler,http://www.cs.stir.ac.uk/courses/ITNP60/lectures/1%20Data%20Mining/3%20%20Classification.pdf ,July 10, 2015.

[7]. "Data Mining Concepts and Techniques", Han J,and Kamber M. in 2006, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor.

[8]. Burcu a.m, "A path model for analyzing undergraduate students' achievement", Journal of WEI Business and Economics-December 2013, Volume 2 Number 3.

[9]. Harris, A. in 1999, "Teaching and Learning in the Effective School". Aldershot,Ashgate .

[10]. "Concordia Online Education, Strategies to Improve Classroom Behavior and Academic Outcomes", http://education.cuportland.edu/blog/classroom-resources/strategies-to-improve-classroom-behavior-and-academic-outcomes/. September 19, 2015.

[11]. Fayyad U, Bradley P ,and Renia C, "Scaling EM clustering to large databases. Technical Report", Microsoft Research in1999.

[12]. Shannaq B, Rafael Y ,and AlexandroV. In 2010 "Student Relationship in Higher Education Using Data Mining Techniques", Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59.

[13]. Ayesha S, Sattar A, Mustafa T, and Khan I. in 2010,"Data Mining Model for Higher Education System", European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29.

[14]. Ron Kohavi, and George H. John, "Wrappers for feature subset selection, Artificial Intelligence" 97 (1997) 273-324.

[15]. "Using Information Gain Attribute Evaluation to Classify Sonar Targets", Jasmina Novakovic, 17th Telecommunications forum TELFOR 2009, Serbia, Belgrade, November 24-26, 2009.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[16]. "Random Forest Machine learning", Beriman L, in 2001, 45, no, 1, 5-32.

[17]. "Classification and Regression Trees", Breman L, Friedman J, Ohlesen R, and Stone in C, in 1984, wadsworth, Monterey,CA.

[18]. Cortez P, RMiner, "Data Mining with Neural Networks Support Vector Machines using R. Introduction to Advanced Scientific Software's and Toolboxes", Rajesh(Ed.).

[19]. Flexer A, in 1996,"Statistical Evaluation Of Neural Networks Experiments Minimum Requirements and Current Practice", In proceedings of the 13th European Meeting on Cybemetics and systems Research .Vienna, Australia, Vol.2, 1005-1008.

[20]. "The Elements of Statistical Learning Data Mining, Inference, and Prediction" Hastie T, Tibshirani R, and Friedman J in 2001.Springer Verlag, NY, USA.

[21]. Kotsiantis S, Pierrakeas, and Pinetelas P in2004. "Predicting Students Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence (AAI)", 8, no.5, 411-426.

[22]. Ma Y, Liu B, Wong C, Yu P, and Lee S. In 2000."Targeting the right students using data mining". In Proc. of 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA, 457–464.

[23]. "Predicting student performance: an application of data mining methods with an educational web-based system".Minaei-Bidgoli.B, Kashy D, Kortemeyer G and Punch W. In Proc. of IEEE Frontiers in Education. Colorado, USA, 2003, 13–18.

[24]. Pardos Z, Heffernan N, Anderson B and Heffernan C., in 2006. "Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks". In Proc. of 8th Int. Conf. on Intelligent Tutoring Systems. Taiwan.

[25]. E.A.Amrieh, T. Hamtini and I. Aljarah, "Preprocessing and analyzing educational data set using X-API for improving student's Performance", In Appl ed Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on. IEEE, **(2015)**, pp. 1-5.

[26]. "Predicting School Failure a Dropout by Using Data Mining Techniques", Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto IEEE Journal of Latin American Learning Technologies,vol. 8, no. 1,February 2013.

[27]. Paulo Cortez and Alice Silva, "Using Data Mining to Predict Secondary School Student performance", Dept.Information Systems, Algoritmi, R&D Centre, University of Minho. http://www3.dsi.uminho.pt/pcortez .

[28]. Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah, The University of Jordan, Amman, Jordan, http://www.Ibrahimaljarah.com www.ju.edu.jo.

[29]. Mining Educational data to predict students academic performance using Ensembel Methods. Elaf AbuAmrieh,Thair Hamtini,Rresearch Gate September 2016.

Mrs. K. Deepika "Comparison Of Student Academic Performance On Different Educational Datasets Using Different Data Mining Techniques "International Journal of Computational Engineering Research (IJCER), vol. 08, no. 09, 2018, pp 28-38