

Student Academic Performance Prediction using Supervised Learning Techniques

<https://doi.org/10.3991/ijet.v14i14.10310>

Muhammad Imran ^(✉), Shahzad Latif, Danish Mehmood, Muhammad Saqlain Shah
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad, Pakistan
dr.imran@szabist-isb.edu.pk

Abstract—Automatic Student performance prediction is a crucial job due to the large volume of data in educational databases. This job is being addressed by educational data mining (EDM). EDM develop methods for discovering data that is derived from educational environment. These methods are used for understanding student and their learning environment. The educational institutions are often curious that how many students will be pass/fail for necessary arrangements. In previous studies, it has been observed that many researchers have intension on the selection of appropriate algorithm for just classification and ignores the solutions of the problems which comes during data mining phases such as data high dimensionality, class imbalance and classification error etc. Such types of problems reduced the accuracy of the model. Several well-known classification algorithms are applied in this domain but this paper proposed a student performance prediction model based on supervised learning decision tree classifier. In addition, an ensemble method is applied to improve the performance of the classifier. Ensemble methods approach is designed to solve classification, prediction problems.

This study proves the importance of data preprocessing and algorithms fine-tuning tasks to resolve the data quality issues. The experimental dataset used in this work belongs to Alentejo region of Portugal which is obtained from UCI Machine Learning Repository. Three supervised learning algorithms (J48, NNge and MLP) are employed in this study for experimental purposes. The results showed that J48 achieved highest accuracy 95.78% among others.

Keywords—Educational Data mining, Educational Data Mining, Predicting Student Performance, Decision Tree, Ensemble

1 Introduction

Educational quality is compulsory in the development of each country. The data amount in education domain is getting increase day by day with the help of admission system, academic information system, learning management system, e-learning etc. The data collected from students are usually used for making simple quires for decision making. But most of the data remain unused due to complexity and large volume data sets. Therefore, to analyze this huge amount of educational data is the great interest to predict student performance. Data mining is the practice of find out useful information

from huge sets of data, also known as knowledge discovery in databases (KDD). It has been applied successfully in multiple domains including banking, medical, business and now has been used for educational purposes called Educational Data Mining.

The prediction of student performance is a crucial task which is being researched by using EDM. This task foresees the value of an unidentified variable which describes the students regarding outcome (Pass/Fail), grades, marks etc. Predicting student Attrition, failures, success are the main areas which are discussed in the literature review of this study. Each stakeholder belongs to this domain wants an early warning system to predict learning on early stages. This early warning system not only reduced the learning costs but also time and space requirements.

One of the biggest challenges is to improve the quality of the educational processes so as to enhance student's performance. Instructors can update their teaching methodology to fulfill the requirement of poor performance students and can provide additional guidance to deserving students. The prediction results might help students develop a good understanding of how well or bad they would perform in a course and then can take steps accordingly. Increasing the student retention is a long-term target of any educational institutions around the globe. There are many positive impacts of increased retention such as increased college reputation, ranking and better job opportunities for alumni etc.

To analyze data using classification technique, well known classification algorithms such as Decision tree (DT), Artificial neural networks (ANN), K-nearest neighbor (KNN) and Rule Induction (RI) are being used for prediction purposes. Quality of a predictive classification model is measured by its ability to find out the unknown patterns accurately. This study employed three classification algorithms J48 from DT, NNge from IR and MLP from ANN for experimental purposes. The major objective of the proposed methodology is to build the ensemble classification model that classifies a students' performance as Pass or Fail.

2 Previous Work

Dorina et al. [1] proposed a predictive model for student's performance by classifying students into binary class (successful / unsuccessful). The proposed model was constructed under the CRISP-DM (Cross Industry Standard Process for Data Mining) research approach. The classification algorithms (OneR, J48, MLP and IBK) were applied on the given dataset. The results show that the highest accuracy was achieved by the MPL model (73.59%) for identification of successful while other three models perform better for the identification of unsuccessful students. The model was unable to work out for data high dimensionality and class balancing problems.

Edin Osmanbegovic et al. [2] builds a model to predict student academic success in a course by reducing data dimensionality problem. Various machine learning classifiers such as NB, MLP and j48 were evaluated in this study. The result shows that the Naïve Bayes gained the highest accuracy 76.65%. The proposed model not handles the class imbalance problem.

Carlos et al. [3] addressed a student failure prediction model based on machine learning techniques to resolve the class imbalance and data dimensionality problems. Ten classifiers were applied on dataset. The ICRM classifier achieved the highest accuracy 92.7% among others. Due to varying student's characteristics at each educational level, the performance of proposed model was not tested for other levels of education.

Another EDM Challenge is to predict the drop-outs of the students from their courses [4]. Four data mining methods with six combinations of attributes were participated in this study. The result shows that the support vector machine model with the combination of the predictor variables was more accurate while classifying the data. The inclusion of an attribute, earned grades of pre-requisite courses in the data set was the limitation of this study because it might be possible that during study of any course the student might have improved his knowledge of pre-requisite of this course.

Ajay et al. [5] conducted study on the prediction of student performance. The main contribution of the study was to introduce a new social factor called "CAT" which describes that in early times Indians were divided into four types of groups on the basis of their social status etc, which have a direct effect on the student education. Four classifiers oneR, MLP, J48, and IB1 were applied on the data set. The results indicated that the IB1 model was the highest accuracy (82%) achieved.

Build an improved version of the ID3 model, which predicts the student academic performance [6]. The weakness of the ID3 model was its intension to select those attributes as a node which had more values. In a result generated tree was not efficient. The proposed model overcomes such problem. Two output classes were produced by this model (Pass and Fail). The classifiers including J48, wID3 and Naïve Bayes were applied and results compared. The wID3 achieved high accuracy 93%.

Alaa Khalaf et al. [7] proposed a model to predict student success performance in courses. Three Decision Tree classifiers such as (J48, Hoeding tree, Reptree) were employed by this study. The highest accuracy 91.47 % was achieved by Reptree. The model was unable to work out for data high dimensionality and class balancing problems.

Dech Thammasiri et al. [8] proposed a model to provide early classification of poor academic performance of freshmen. Four classification methods with three balancing methods were applied to resolve class imbalance problem. In results the combination of support vector machine and SMOTE achieved the 90.24% highest overall accuracy.

An early warning system was proposed to predict the student learning performances during an online course based on their learning portfolios data [9]. The results showed the approaches accompanied by time dependent variables had high accuracy than other approaches which were not included it. The model was not tested on offline mode. The performance might be decreased in offline mode using time dependent attributes.

Mostly previous studies were assumed that the data mining algorithms performed well with only large data sets but this study supported that the data mining is also suitable for small datasets as well [10]. This research proposed a student success prediction model. A small dataset including student academic data was used by using three decision tree approaches (Reptree, J48, M5P). The result claims that the Reptree obtained the highest accuracy above 90% among them. The proposed model not supported to data high dimensionality and class balancing problems.

Camilo et al. [11] proposed a model to predict student academic attrition by overcoming class imbalance problem. Two algorithms Naïve Bays and Decision tree were used by this study. A cost-sensitive approach. Metacost was used to manage this problem. After that highest accuracy was got by naviey bays upto 85%. The data collection at the end of academic period is not feasible because no one can get benefit at that time.

A student academic performance prediction model was proposed in this study [12]. The classifiers namely J48, Decision Stump, Reptree, NB and ANN with three kinds of attribute setups were evaluated in this study. The J48 classifier achieved the high accuracy 90.51%.

Proposed approached was contributed by evaluated three number of classes (drop-out, persisting, and completed) while predicting student dropout [13]. Ten classification models were assessed. The results of experiments depict that the Naïve Bayes algorithm had the highest predicting levels for the three classes of students.

Bilal et al. [14] presented a student failure prediction model which identified the students that might be at-risk. Four output classes (Average, Risk, below Average and Above Average) were generated by the proposed model based on the CGPA of the students. Six classifiers including were applied on the given dataset. The ID3 got the highest accuracy 79.23%. The model was unable to work out for class imbalance problem.

An ensemble model including classifiers (NB, SVM, KNN) was proposed for the identification of weak students [15]. The dataset included a most effective attribute as standard based grading assessment in addition to typical score-based grading. The results of proposed model with six other individual classifiers were compared and conclude that the accuracy of ensemble model was 85% which is higher than others.

A multilevel classification model was proposed to resolve the multiclass classification problem in the prediction of student performance [17]. The goal of study was not only to increase the model accuracy but also increase the accuracy of the individual classifier. The model contains two levels. Initially a re-sampling technique was performed on the dataset to overcome the class distribution problem in the preprocessing phase. In the first level, four classification models were applied on the dataset namely IBK, MLP, NB, J48. Results were evaluated and compared. The results show that the decision classifier (j48) was highly accurate and selected for use in the next level. In the level two, outliers were identified by comparing the previously predicted results with actual results and removed accordingly. Once again re-sampling technique with high accurate classifier which was selected previous (J48) was applied onto the filtered dataset and results were compared with the results of applying remaining classifiers also on the filtered dataset. The results depict that the J48 classifier got the above 90% accuracy for overall model as well as for individual classes prediction.

An early student failure identification model was proposed in this study by evaluating data mining techniques as well as preprocessing approaches. Several techniques and models were applied (ANNs, decision trees, support vector machines, naïve bayes) in this study and conclude that the support vector machines is outperformed from the others ones [18]. The data was collected from two different types of data sources. Model not supported for reducing the classification errors.

The rest of the report is organized as Section-III Methodology Section –IV Results and Discussion and Section -V Conclusion

3 Methodology

To address the common issues of above literature review such as class imbalance, data hi-dimensionality and classification errors, this study has proposed a model which have following phases. Figure 1 shows the main steps of proposed methodology.

3.1 Data collection

A student performance data set used in this study has collected from UCI Machine Learning Repository [16]. The data was collected for academic session 2005-2006 of two schools of Alentejo region Portugal. It includes 1044 instances with 33 attributes including student grades, demographic, social and school related features.

3.2 Data preprocessing

Pre-processing plays an important in data mining. Its purpose is to convert raw data into a suitable form which can be used by mining algorithms. Following tasks are performed in this phase.

Data integration: Data Integration means to gather the data from the multiple sources into single repository. Redundancy is the common problem occurred when integrating data. The dataset consists of two comma separated values files which were taken from UCI Machine learning repository. These files contained the performance data of two courses (Portugal Language holds 395 instances and Mathematics holds 649 instances) which were studied by Portuguese Students. In this step, multiple files are integrated into one file. In order to perform consolidation, an attribute (Course) is added to describe the course such as (P for Portugal or M for Mathematics).

Data cleaning: In this phase, missing and noisy data is handled to achieve data consistency. The dataset occupied by this study not have any missing and outliers etc.

Discretization: The discretization mechanism is used to transform the desire data from numerical values into nominal values. Some classifiers are not applicable on continues data. That's why target attribute G3Grade has converted into nominal. Such as other countries, Portuguese education system follows the 20 point grading scale. In which 0 shows the lowest and 20 is the perfect score. The student availed grade points of three sessions have converted into binary target nominal intervals by applying following rule. Declared Pass as P, if points are greater than ten and declared Fail as F, if points are less than or equal to ten. The target variable (Class Label) is G3Grade which describes whether student is pass or fail.

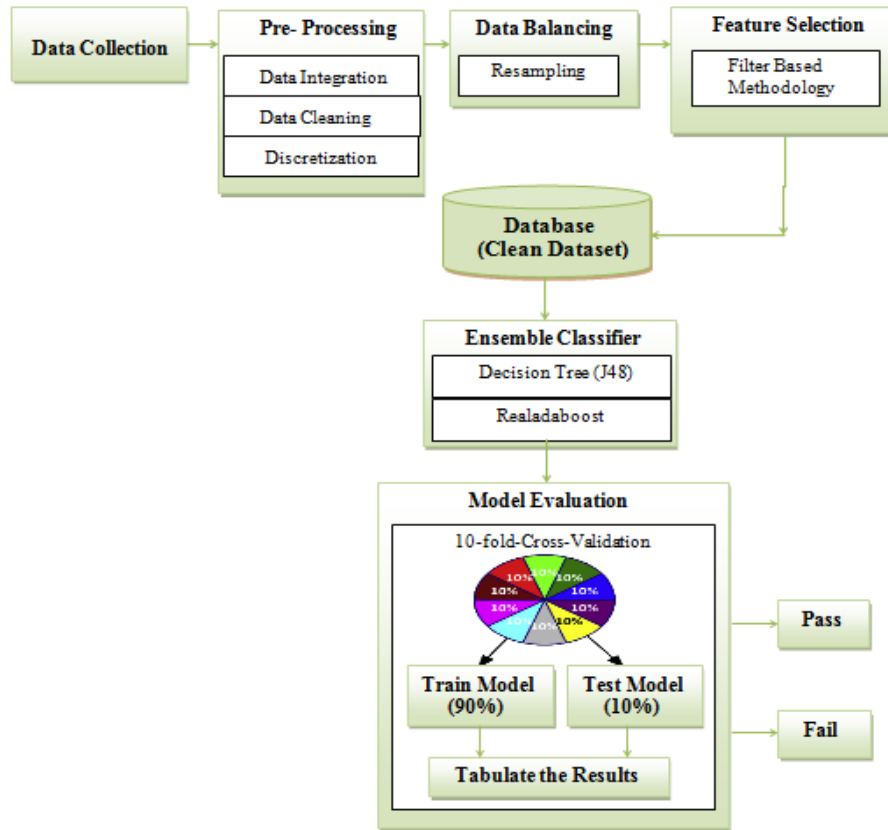


Fig. 1. Proposed System for Student Academic Performance Prediction

3.3 Class balancing

In this phase, data balancing approach is applied after data pre-processing for solving the class imbalance problem. The class imbalanced problem arises when the number of instances in one class is much smaller than the number of instances in another class or other classes. Traditional classification algorithms provide high accuracy for majority classes when data is un-balance because during classification, they have much intension towards majority class instances and have less intension for minority class instances. In the collected dataset, 22.03% of students failed the course, resulting in a serious class imbalance problem. The adjustment of the ratio of two class samples can improve the machine's learning performance. Therefore, we employed a class balancing method known as re-sampling in this phase.

Figure 2 shows the class distribution.

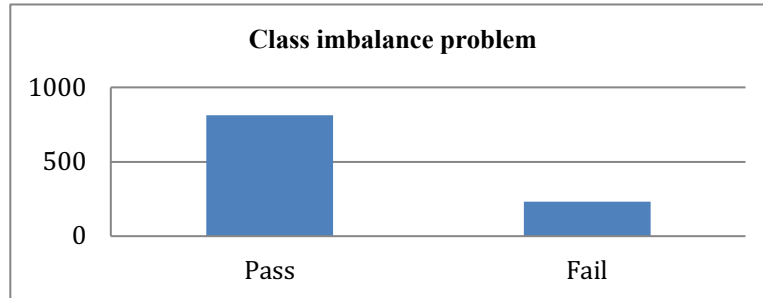


Fig. 2. Class Distribution on Scale

After re-sampling on the training set, 50% PASS and 50% FAIL students are obtained.

3.4 Feature selection

The student performance dataset may contain many attributes, which may be inappropriate for classification purposes. The problem of data high dimensionality arises when included large amounts of student’s characteristics which can influence student performance such as educational background, social, demographics, family, socioeconomic status etc. This issue can be resolve by selecting important features from the dataset.

The purpose of feature selection is to select an appropriate subset of features which can efficiently describe the input data, which reduces the dimensionality of feature space and removes irrelevant data. Feature selection methods are mainly categorized into wrapper-based and filter-based methods. Filter method is searching for the minimum set of relevant features while ignoring the rest. It uses variable ranking techniques to rank the features where the highly ranked features are selected and applied to the learning algorithm.

This study applied filter method using information gain-based selection algorithm to evaluate the feature ranks. It’s checking which features are most important to build students’ performance model. During feature selection, a rank value is assigned to each feature according to their influence on data classification. The highly ranked 12 out of 33 features have selected while others are excluded.

Table 1, shows the list of some selected features as sample features after filter-based evaluation.

Table 1. Samples of Selected Features with Rank values

Sr. #	Selected Attribute	Description	Ranked Values
1	G2Grade	Pass / Fail in 2 nd Term	0.5936
2	G1Grade	Pass / Fail in First Term	0.4921
3	Failures	No of failures in past class	0.4747
4	AbsRate	Absence rate in class	0.4038
6	Fedu	Father qualification	0.3331

3.5 Model construction

The literature review recommends that in general there is no single classifier that works best in all contexts to provide good prediction.

In this paper, student's performance prediction model is build using ensemble method. Ensemble method is a learning approach that combines multiple models to reduce their classification errors as well as to enhance the accuracy of weak classifiers. The predictions made by ensembles are usually more accurate than predictions made by a single model.

Generally, there are following two types of Ensemble approaches:

- **Homogeneous ensembles:** A combination of one ensemble learning (Meta model) such as Bagging, Boosting and one base model.
- **Heterogeneous ensembles:** An ensemble that combines at least two different base methods.

This study has employed homogeneous ensemble approaches.

In model building, first J48 classifier is applied on the clean dataset. Then a Meta classifier Realadaboost has applied to enhance the accuracy of J48 classifier by reducing its classification errors. After studied literature survey it has been observed that the educational data sets are belongs to category of small kind of datasets and Decision Tree (DT) classifiers has ability to work accurately on small kind of datasets. Therefore we have selected DT classifier (J48) as base classifier for our proposed study. DT is a well-known and powerful supervised learning technique. It comprises a hierarchical structure comprising nodes and branches; an internal node represents an input variable, the branch of an internal node represents a subset of the values of the corresponding input variable, and a leaf node is associated with a value (or a class label) of the output variable.

Realadaboost is a version of boosting algorithm. The main idea behind this algorithm is to pay more attention to patterns that are hard to classify truly. It increased the predictive accuracy and reduced the misclassification instances ration.

4 Results and Discussion

4.1 Model evaluation

For our experiments, three classifiers J48, NNge and MLP have evaluated using 10-folds cross validation technique. This technique divides the data set into 10 subsets of equal size; nine of the subsets are used for training, while one is left out and used for testing. The process is iterated for ten times, the final result is estimated as the average error rate on test examples.

4.2 Evaluation measures

In our experiments, we use five common different measures for the evaluation of the classification quality. Details are under as:

- **CCI** (Correctly Classified Instances): represents the number of correctly classified instances divided by the total instances. It is also known as accuracy.

The CCI formula is= $(TP+TN)/(TP+FP+TN+FN)$

- **ICI** (Incorrectly Classified Instances): represents the number of incorrectly classified instances divided by the total instances

The ICI formula is= $(FP+FN)/(TP+FP+TN+FN)$

- **Precision**: of algorithm represents the percentage of accurate classified instances from all truly classified instances.

Precision = $TP/TP+FP$

- **Recall**: reflects the division number of correctly classified instances by the total number of all instances (almost recall value be same as CCI).

TPrate= $TP/TP+FN$

- **F-Measure**: measured from recall and precision values (double value of precision multiplied by recall divided by the value of summation of recall and precision).

4.3 Result analysis

In the first experiment, three classification algorithms (J48, NNge and MLP) are simply executed individually on dataset without using proposed model steps. In the light of results mentioned in Table 2 and graphical representation of figure 3, it has been observed that the highest accuracy achieved by J48 which is not enough as compared to previous studies and lowest accuracy has achieved by MLP.

Table 2. Classification Method Results with original dataset

Algorithm	CCI	ICI	Precision	Recall	F-Measure
J48	90.13 %	9.86 %	0.902	0.901	0.902
NNge	89.55 %	10.46%	0.897	0.896	0.896
MLP	88.98 %	11.01%	0.895	0.890	0.892

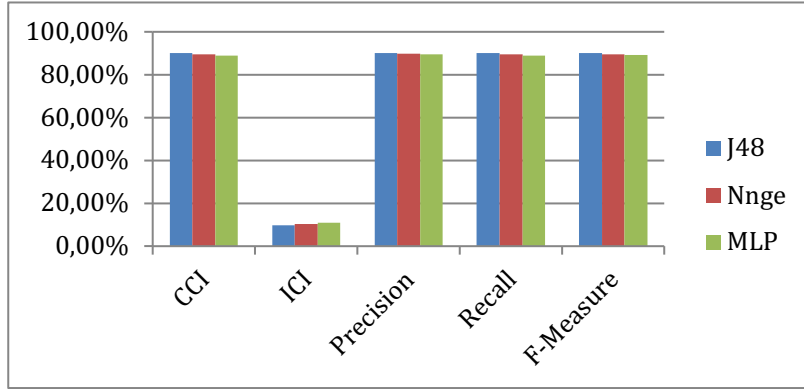


Fig. 3. Single Classifier Based Performance Measures

In the second experiment, proposed methodology has performed step by step. Results can be seen in table 3 with graphical representation in figure 4. It has been observed that the highest accuracy achieved is 95.78% by J48 classifier and the lowest accuracy achieved 92.81% by NNge. It has observed that after reducing class imbalance, data high dimensionality as well as by using ensemble method the proposed model accuracy has improved significantly for all classifiers.

Table 3. Results of proposed Model

Algorithm	CCI	ICI	Precision	Recall	F-Measure
J48	95.78 %	4.21 %	0.958	0.958	0.958
NNge	92.81 %	7.18 %	0.929	0.928	0.928
MLP	93.39 %	6.60 %	0.934	0.934	0.934

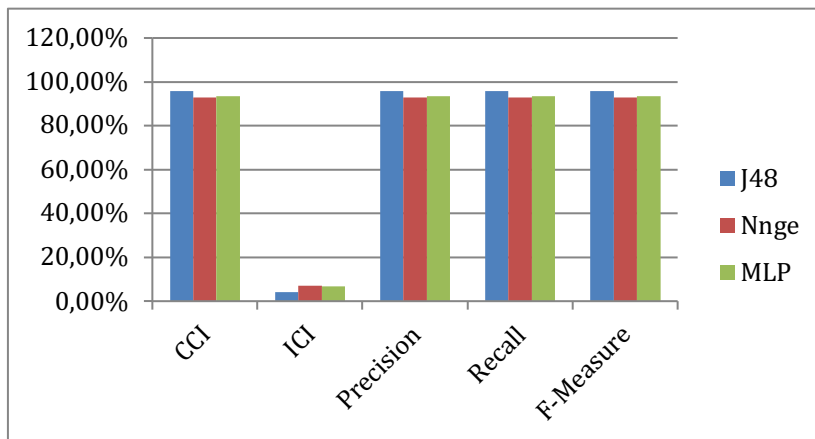


Fig. 4. Proposed Model Performance Measures

During this experiment, we have also measured the classification errors in terms of Root Mean Squared Error (RMSE). Figure 5 shows the graphical representation of RMSE. This clearly shows the results with and without ensemble classification errors rates.

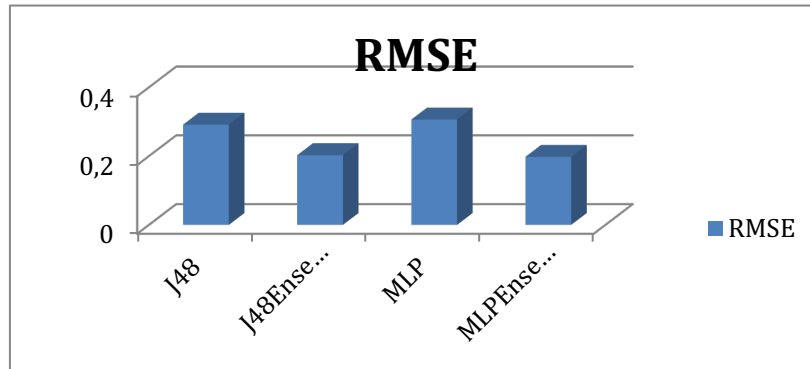


Fig. 5. RMSE of single and Proposed Model Comparison

4.4 Model comparison

In this section we compared the results of other student performance prediction model with our proposed system. Table 4 shows the detail of participated approaches such as SVM, Decision Stump, Reptree and Our proposed ensemble method with their performed actions. Our proposed model performed all the actions such as class balancing, feature selection and use of ensemble methods. The results show that our proposed study has an improvement, being higher than others in terms of accuracy.

Table 4. Comparison of Proposed Model with other Techniques

Ref	Technique	Data source	Data set	Class Balancing	Feature Selection	Ensemble Method	Results of Reference study	Results Using Proposed study Dataset
[8]	SVM	Tulsa, USA	2165	Handled	Handled	Not Used	90.24%	92.22 %
[12]	Decision Stump	UCI Student Performance	649	Not Handled	Handled	Not Used	90.51%	90.81 %
[7]	Reptree	UCI Student Performance	1044	Not Handled	Not Handled	Not Used	91.47%	91.28 %
Proposed Model	Ensemble (J48, Realada-boost)	UCI Student Performance	1044	Handled	Handled	Used		95.78 %

5 Conclusion

The accurate student academic performance prediction model is demand of every educational institute nowadays. But to resolve the data quality issues in student performance prediction model is often biggest challenge. This research work, presented a student performance prediction model based on supervised learning technique Decision Tree. The performance of Student's predictive model is assessed on dataset by set of classifiers namely; J48, NNge, and MLP. In addition, an ensemble method is applied to improve the performance of these classifiers. The result shows that the proposed ensemble model including Decision tree (J48) classifier achieved the high accuracy which is 95.78 %.

In future the proposed model will be tested on large dataset with more number of attributes.

6 References

- [1] Dorina Kababchieva. (2012). Student Performance Prediction using Data Mining Classification Algorithms. *International Journal of Computer Science and Management Research*, vol. 1.
- [2] Edin Osmanbegovic and Mirza Suljic. (2012). Data mining approach for predicting student performance. *Journal of Economics and Business*, vol. X, Issue 1.
- [3] Carlos Marques-Vera and Alberto Cano. (2013). Predicting student failure at school using genetic programming. *ApplIntell*, vol. 38, pp.315–330.
- [4] Shaobo Huang and Ning Fang. (2013). predicting student academic performance in an engineering dynamic course: A comparison of four types of predictive mathematical models. *Computers & Education*, vol. 61, pp. 133–145. <https://doi.org/10.1016/j.compedu.2012.08.015>
- [5] Ajay Kumar Pal and Saurabh Pal. (2013). Data Mining Techniques in EDM for Predicting the Performance of Students. *International Journal of Computer and Information Technology*, vol. 02, Issue 06.
- [6] Ramanathan, Saksham Dhanda and Suresh Kumar D. (2013). Predicting Student Performance using Modified ID3 Algorithm. *International Journal of Engineering and Technology*, vol. 5 No 3.
- [7] Alaa Khalaf Hamoud. (2016). Selection of Best Decision Tree algorithm for prediction and classification of student Action. *American International Journal of Research in Science, Technology, Engineering & Mathematics*, vol 1, pp. 26-32.
- [8] Dech Thammasiri, DursunDelen, PhayungMeesad andNihatKasap. (2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41, pp.321–330. <https://doi.org/10.1016/j.eswa.2013.07.046>
- [9] Ya-Han Hu, Chia-Ling L and Sheng-Pao Shih. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, 36, pp. 469–478. <https://doi.org/10.1016/j.chb.2014.04.002>
- [10] SreckoNatek and Moti Zwilling. (2014). Student data mining solution–knowledge management system related to higher education institutions. *Expert Systems with Applications*, 41, pp.6400–6407. <https://doi.org/10.1016/j.eswa.2014.04.024>

- [11] Camilo Ernesto Lopez Guarín, Elizabeth León Guzmán, and Fabio A. González. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining IEEE Transactions, vol. 10(3).
- [12] Sana Akhai and Ruchi Karia. (2017). Automated Performance Evaluation System. IJARIT, 3(2).
- [13] Laci Mary Barbosa Manhaes, Sergio Manuel Serra da Cruz and Geraldo Zimbrão. (2016). Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs. ACM. <https://doi.org/10.1145/2695664.2695918>
- [14] Bilal Mehboob, Rao Muzamal Liaqat and Nazar Abbas Saqib. (2016). Predicting Student Performance and Risk Analysis by Using Data Mining Approach. International Journal of Computer Science and Information Security (IJCSIS), vol. 14, No 7.
- [15] Farshid Marbouti and Krishna Madhavan. (2016). Models for early prediction at risk students in a course using standard based grading. Computers & Education, 103, pp. 1-15. <https://doi.org/10.1016/j.compedu.2016.09.005>
- [16] Cortez, Paulo, Alice Maria and Gonçalves Silva. (2008). Using data mining to predict secondary school student performance. [Online]. Available: <http://archive.ics.uci.edu/ml/machine-learning-databases/00356/student.zip>. [Accessed: 23- Jun- 2016].
- [17] Mrinal Pandey and S.Tarun. (2014). A Multi-Level Classification Model Pertaining to the Students' Academic Performance Prediction. International Journal of Advances in Engineering & Technology, 4: 1329-1341.
- [18] Santana, M.A., Costa, E.B., Fonseca, B., Rego, J., and de Araújo, F.F (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computer. Human. Behavior, 73: 247–256. <https://doi.org/10.1016/j.chb.2017.01.047>

7 Authors

Muhammad Imran is Assistant Professor of computer science in the department of computer science at Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan. He received his PhD degree in Information Technology from University Tun Hussin onn Malaysia. His main research interests include data mining, Evolutionary computing, Computer Vision and Medical image processing. (dr.imran@szabist-isb.edu.pk)

Shahzad Latif received PhD degree in Electronics Engineering from ISRA University, Islamabad Pakistan. Currently he is Assistant Professor in the department of computer science at Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan. His main research interests include Computational Intelligence, Digital Signal Processing and Fuzzy systems (dr.shahzad@szabist-isb.edu.pk)

Danish Mehmood received PhD degree in Smart Grid from Comsats University, Islamabad Pakistan. Currently he is Assistant Professor in the department of computer science at Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan. (dr.danish@szabist-isb.edu.pk)

Muhammad Saqlain Shah got his degree of Master of Science in the Faculty of Computing and engineering sciences from Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan. His main research interests include data mining, machine learning (msaqlain80@hotmail.com)

Article submitted 2019-02-13. Resubmitted 2019-03-22. Final acceptance 2019-03-25. Final version published as submitted by the authors.