Prediction of Academic Performance During Adolescence Based on Socioeconomic, Psychological and Academic Factors

A.T.M. Shakil Ahamed, Navid Tanzeem Mahmood and Rashedur M. Rahman

Abstract Educational effectivity is paramount towards enhancing modernization. In our study we have taken into account various socioeconomic, psychological and academic factors to properly understand a student's life during adolescence and their effect on academic performance. In order to build a predictive model, we have pre-processed the data using dimensionality reduction, data balancing, discretization, and normalization and then classified the data using different machine learning techniques like Artificial Neural Net, K-Nearest Neighbors and Support Vector Machine. Lastly we have discovered patterns throughout the dataset in relation with academic performance.

Keywords Data mining · Predictive model · Machine learning

1 Introduction

Our study falls into the field of Educational Data Mining: a rather new area where academic performance is investigated and steps are figured out to enhance it. The aim of this study is set towards Higher Secondary Certificate (HSC) (12th grade equivalent) candidates; all data has been collected from recent HSC examinees: university freshers.

The goal of our study is to provide a prediction of HSC result from socioeconomic, psychological and academic factors and to discover pattern in attributes which allow students to visualize how the attributes influencing their educational

A.T.M. Shakil Ahamed · N.T. Mahmood · R.M. Rahman (✉)
Department of Electrical and Computer Engineering, North South University, Plot-15,
Block-B, Bashundhara, Dhaka, Bangladesh
e-mail: rashedur@northsouth.edu; rashedur.rahman@northsouth.edu

A.T.M. Shakil Ahamed
e-mail: ahnafshakil@gmail.com

N.T. Mahmood
e-mail: navid.mahmood@hotmail.com

performance which help students to enhance the outcome of HSC exam. Guardians of students and responsible authorities could use such information generated from our model and help the students to enhance the HSC performance.

2 Related Work

To group the student according to their skill level in [1] the authors have used model based clustering, other clustering and K-mean technique. In paper [2] the authors have used quiz score, midterm marks, final marks, lab marks and CGPA as attribute to predict semester final grade of individual courses. They have used ANN, C4.5 and Naïve Bayes for the classification and the have managed to reach accuracy around 82%. In order to predict academic performance [3] class attendance, seminar, class test score and assignment marks have been used as attributes. Authors have used C4.5, CART and ID3 algorithms to design their predictive model.

3 Designing Survey Questionnaire and Data Set

Designing a predictive model greatly depends on the quality of the data set. We are very careful while designing the survey questionnaire, as the survey is the source of attributes for the predictive model. In order to get the relevant attributes, we have interviewed experts and explored different factors related to academic performance. We have made sure that every attribute is proven to be related to academic performance through published research. We have surveyed 423 students from different location and economic status for maintaining the versatility in the data set. Our data set consists of 33 attributes grouped into 3 main categories namely socioeconomic, psychological and academic as shown in Table 1.

4 Overview of Designing Predictive Model

The design of the predictive model is shown in Fig. 1. After getting raw data we have realized that data set needs to be preprocessed before applying any learning algorithm to augment the predictive accuracy. We have used data balancing technique to balance the data set and also used Principal Component Analysis (PCA) for dimensionality reduction then Optimal equal width binning for discretization and finally used normalization technique as pre-processing. Then we have applied 3 different classifiers namely Artificial Neural Network (ANN), K-Nearest Neighbors (K-NN) and Support Vector Machine (SVM) and after comparative study of the classifiers we have derived the best model for the prediction.

Table 1 All the attributes used in designing the predictive model

Categories	Factors
Socioeconomic and demographic factors	• Sex • Age • Mother's education • Father's education • Mother's job • Father's job • Average monthly family income • Place of education • Private tuition
Psychological factors	• Family size • Age difference with most immediate sibling • Parents' involvement in education • Parent's marital status • Absence/Death of parent(s) • Involvement in physical activity • Motivating teacher • Motivating parents • Academic ambitions • Time spent on social media • Romantic relation • Hanging out with friends • Alcohol/Drugs consumption • Smoking • Health status of participant • Health status of family members
Academic factors	• Early childhood education • Weekly study time • Previous failures • Number of absence(s) • Extracurricular activities • Accessibility of the internet • SSC (Secondary School Certificate) result (10th grade equivalent) • HSC result

5 Data Pre-processing

5.1 Data Balancing

Our data set is not balanced as shown in Fig. 2. The problem with of the imbalanced data can be visualized while mammographic pictures are used as data set. The picture contains 98% of normal pixels and 2% of abnormal pixels. So simply by guessing the pixel as normal one can be right for the 98% of the time [4]. We have used SMOTEBoost technique [5] which consists of Synthetic Minority oversampling technique and a boosting procedure to provide equal entropy to the both majority and minority class to enrich the quality of the accuracy of the classifiers.
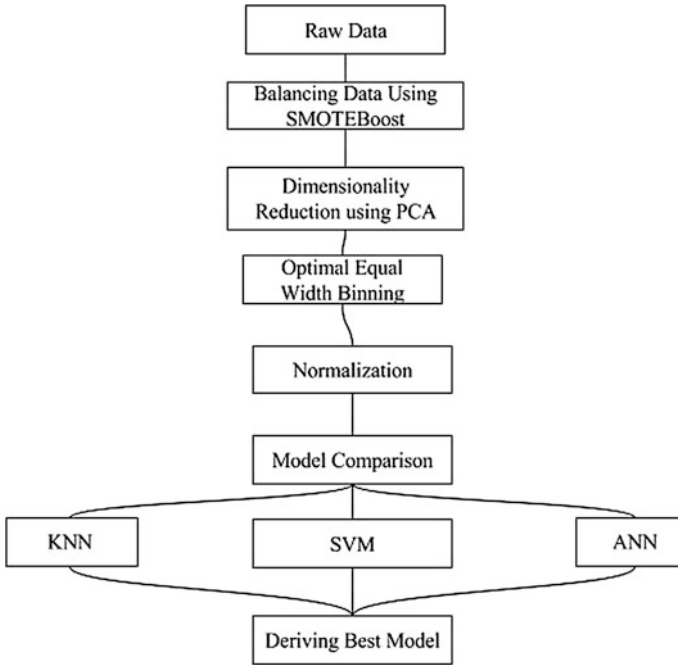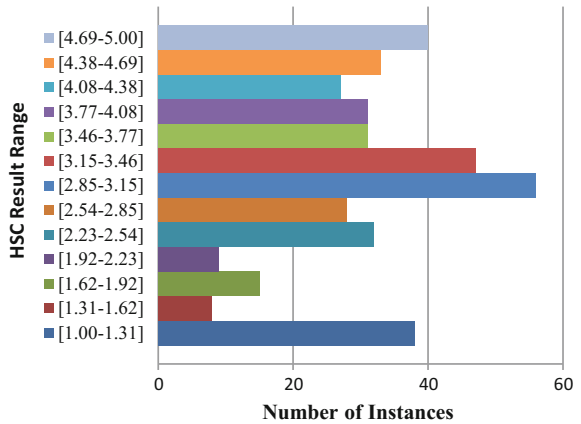
Fig. 1 An overview of our system

Fig. 2 Instances of target class (HSC result)

5.2 Dimensionality Reduction

The data set consists of 33 attributes; most of them are dispersed from each other and some of them are redundantly related to the target class. Removing redundancy bolsters classifiers by excluding the curse of the dimensionality reduction [6]. We

have used principal component analysis (PCA) to reduce the dimensionality of our data set. PCA applies an orthogonal transformation; so all resultant attributes are correlated to only the target class and not each other [7].

5.3 Discretization

Our data set consists of attribute such as SSC and HSC result which are continuous valued attributes. To discretize the data set we have implemented Optimal Equal Width binning [8] that searches dynamically for the optimal width and number of bins for the target class. In our data set we got 8 bins with the width of 0.5. As the HSC result is in the range of 1–5 the bins we got from the binning are: $[-\infty–1.5]$, $[1.5–2]$, $[2–2.5]$, $[2.5–3]$, $[3–3.5]$, $[3.5–4]$, $[4–4.5]$ and $[4.5–\infty]$.

5.4 Normalization

Our data set consists of attributes with different scales and units, for instance, "Weekly study time" has the unit of hour, on the other hand, average family income has the unit of Taka and "Family involvement" is measured in a scale of 5, conversely "Health status" is measured in the scale of 10. To ensure the equal contribution of the attribute to predict the target class we have used min-max [9] normalization technique to rescale the data set in order to produce better training set for the classifiers using the Eq. (1).

$$x^{'} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{1}$$

This technique is able to preserve all the relationship among attributes and rescale them for equal contribution of the attribute to predict target class.

6 Learning Algorithms

After applying the pre-processing techniques we have used learning algorithms to predict the performance of HSC examination. We have spilt the data set in 70:30 for training and testing. In order to ensure consistency in class distribution throughout the data set we have approached stratified sampling to maintain the equal distribution in the training and testing set in order to discard the possibility of having adverse impact on predictive accuracy due to exhibiting inconsistency in the split parts of the data set.

6.1 K-Nearest Neighbors (K-NN)

K-NN learns from each instance from other known outputs of the data set. All attributes are transformed into vector space where each of the N dimensions of an attribute is represented by vectors [10]. K-NN uses Euclidian distance to measure the similarity metric amongst neighbors using the Eq. (2).

$$d_E(x, y) = \sum_{i=0}^{N} \sqrt{(x_i^2 - y_i^2)} \tag{2}$$

It works the following way in our system:

(a) Finding appropriate k-value to get the similar neighbors for classifying.
(b) Calculating distance between training sample and testing sample.
(c) Sorting the samples based on the distance.
(d) Applying majority vote across the k-nearest neighbors produced the predicted HSC result.

6.2 Support Vector Machine (SVM)

By constructing hyperplanes in a multidimensional feature space it breaks down class levels into different cases. Iteratively, SVM applies a training algorithm to produce an optimal hyperplane and also minimizes the error function in each iteration [11]. Error function is described in the Eq. (3).

$$\frac{1}{2} w^t w + c \sum_{i=1}^{N} \xi_i \tag{3}$$

Subject to: $y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i \ and \ \xi_i \geq 0, i = 1, \ldots\ldots, N$

where,

- C = capacity function
- w = Vectors of coefficients
- b = Constant
- ξ_i = Parameter that handles non-separable data
- i = Number of training cases (N)
- $y \in \pm 1$ = Class labels
- x_i = Independent variable
- Kernel ϕ transforms data to the feature space from the input.

As C increases, the error gets penalized more; the value of C can be set to avoid the problem of overfitting. Dot kernel has been applied for our system as shown below in Eq. (4).

$$K(X_i, X_j,) = \begin{cases} X_i \cdot X_j & \text{Linear} \\ (\gamma X_i \cdot X_j + C)^4 & \text{Polynomial} \\ \exp(-\gamma|X_i - X_j|^2) & \text{RBF} \\ \tanh(\gamma|X_i \cdot X_j + C) & \text{Sigmoid} \end{cases} \tag{4}$$

6.3 Artificial Neural Network

Artificial neurons perform its actions almost like their biological counterparts: they process their mechanism by transmitting information among neurons that helps each other to 'fire' based on given inputs. To 'fire' the activation function shown in Eq. (5) must be satisfied [12].

$$a_i = \sum_{J=1}^{N} W_{ji} x_j + \theta_i \tag{5}$$

where, x_j is the output from a neuron or an external input, W_{ji} is the weight and θ_i is the threshold.

We have used 'feed forward' architecture for designing our system; the outputs of a layer of neurons get fed into the inputs of the neurons from the next layer. The 'hidden' layer(s) exist in between input and output layers. Number of hidden layers of our system can be calculated using the Eq. (6).

$$(Number\ of\ Attributes + Number\ of\ Classes)/2 + 1$$
$$= (32 + 1)/2 + 1 = 17(approx.) hidden\ layers. \tag{6}$$

7 Result Analysis

7.1 K-Nearest Neighbors

While no pre-processing is used the KNN provides the accuracy of 30% as shown in Table 2 but after introducing pre-processing technique KNN it starts achieving better accuracy. After using PCA and SMOTEBoost it increases its accuracy to 50.35%. In addition when we have used Optimal Equal Width binning the accuracy has reached around 63% and after the normalization the accuracy has notably augmented to 70%. It is because KNN uses Euclidian algorithm to measure the

Table 2 Performance comparison of the different predictive models

Models	Accuracy (%)	Precision (%)	Recall (%)	F-measure (%)
ANN	70.05	55.96	59.21	57.54
ANN (PCA + SMOTEBoost)	71.43	79.87	65.68	72.08
ANN (PCA + SMOTEBoost + Binning)	77.01	81.77	87.50	84.54
ANN (PCA + SMOTEBoost + Binning + Normalized)	77.78	75.68	68.95	72.16
K-NN	30.35	16.32	19.25	17.66
K-NN (PCA + SMOTEBoost)	50.35	17.02	22.25	54.29
K-NN (PCA+ SMOTEBoost + Binning)	63.74	63.92	54.77	58.99
K-NN (PCA + SMOTEBoost + Binning + Normalized)	70.00	82.37	61.48	70.41
SVM	70.88	71.99	55.52	62.69
SVM (PCA + SMOTEBoost)	78.50	69.22	56.02	61.92
SVM (PCA + SMOTEBoost + Binning)	76.17	86.85	62.75	72.86
SVM (PCA + SMOTEBoost + Binning + Normalized)	75.93	85.43	57.36	68.64

distance between neighbors; by rescaling the attribute and normalizing the range bolster K-NN to achieve such high score.

7.2 Artificial Neural Network (ANN)

ANN has provided 70% of the accuracy without using any pre-processing technique, after SMOTEBoost and PCA it does not provide any promising change in accuracy. Using discretization with Optimal Equal Width Binning it enhances its accuracy by 7% and normalization does not have any significant impact on accuracy of ANN.

7.3 Support Vector Machine (SVM)

SVM provides accuracy of 70.88% without using any pre-processing technique. After balancing the data set using SMOTEBoost the accuracy has increased by 8%. As SMOTEBoost provides equal entropy to the majority and minority class, it bolsters SVM to provide better accuracy. Binning and Normalization lowers the accuracy of the SVM.

7.4 Key Findings

We have generated key findings using decision trees as shown in below. One of the observations is "higher family involvement notably increases the performance producing better result with same time spent on studying weekly". Family involvement has measured in the scale of 1–5 where 1 is least involvement and 5 is the highest involvement. As we can see in the rule below that with increasing family involvement performance gets better for students with having lower study time.

```
SSC result = 4.250
| Family Involvement = 1: (HSC result) 2.750
| Family Involvement = 3
| | Weekly Study Time = 1 hour: (HSC result) 3.750
| | Weekly Study Time = 2 hours: (HSC result) 4.000
| Family Involvement = 4: (HSC result) 4.250
| Family Involvement = 5
| | Weekly Study Time = 1 hour: (HSC result) 4.000
| | Weekly Study Time = 2 hours: (HSC result) 4.500
| | Weekly Study Time = 3 hours: (HSC result) 4.750
```

Students with the previous failure get better results with the increasing family involvement in their study as shown in the tree below.

```
Previous failures = 3 (Number of previous failures)
| Family Involvement = 2: (HSC result) 2.750
| Family Involvement = 3: (HSC result) 3.250
| Family Involvement = 4: (HSC result) 3.500
```

Another observation is "while parents get separated then having a romantic relation among students hampers the academic attainments". Conversely, another observation exhibits romantic relationship can cause enhancement in academic performance while parents are together.

```
Parent Status = Living Apart
| Romantic Relation = no: (HSC result) 2.750
| Romantic Relation = yes: (HSC result) 1
Parent Status = Living Together
| Romantic Relation = no: (HSC result) 2.250
| Romantic Relation = yes: (HSC result) 3.250
```

8 Conclusion

Highest accuracy of the system is 78.5% while SMOTEBoost is used along with PCA and the second highest accuracy is the 77.78% by ANN along with the PCA, SMOTEBoost, binning and normalization. Pre-processing has significant impact on classifiers in most of the time. Predictive model and the key finding through the visualization of the data set provide students and their parents an important instrument to get better academic performance in HSC.

References

1. Ayers, E., Nugent, R., Dean, N.: A comparison of student skill knowledge estimates. In International Conference on Educational Data Mining, Cordoba, Spain, pp. 1–10 (2009)
2. Jishan, S.T., Rashu, R.I., Mahmood, A., Billah, F., Rahman, R.M.: Application of optimum binning technique in data mining approaches to predict students' final grade in a course. Adv. Intell. Syst. Comput. 159–170 (2015). doi:10.1007/978-3-319-13153-5_16
3. Yadav, S.K., Bharadwaj, B., Pal, S.: Data mining applications: a comparative study for predicting student's performance (2012). arXiv preprint arXiv:1202.4815
4. Woods, K., Doss, C., Bowyer, K., Solka, J., Priebe, C., Kegelmeyer, P.: Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. Int. J. Pattern Recognit. Artif. Intell. 7(6), 1417–1436 (1993)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357 (2002)
6. Jimenez, L.O., Landgrebe, D.A.: Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. IEEE Trans. Syst. Man Cybern. 28(1), 39–45 (1997)
7. Jolliffe I.T.: Principal component analysis. Series: Springer Series in Statistics, XXIX, 487 p. 28 illus, 2nd edn. Springer, NY (2002). ISBN: 978-0-387-95442-4
8. Kayah, F.: Discretizing Continuous Features for Naive Bayes and C4. 5 Classifiers. University of Maryland Publications, College Park, MD, USA (2008)
9. Jayalakshmi, T., Santhakumaran, A.: Statistical normalization and back propagation for classification. IJCTE 3(1), 89–93 (2011). doi:10.7763/ijcte.2011.v3.288
10. Mullin, M., Sukthankar, R.: Complete cross-validation for nearest neighbor classifiers. Accessed from http://www.cs.cmu.edu/~rahuls/pub/icml2000-rahuls.pdf (2002)
11. Min, J., Lee, Y.: Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. Expert Syst. Appl. 28(4), 603–614 (2005). doi:10.1016/j.eswa.2004.12.008
12. Kar, A.: Stock prediction using artificial neural networks. Department of Computer Science and Engineering, IIT, Kanpur (n.d.)